

SEMANTIC DISTANCES BETWEEN MEDICAL ENTITIES

FIB



Universitat Politècnica de Catalunya
Facultat d'Informàtica de Barcelona
Master in Artificial Intelligence

BARCELONA

APRIL 2018

Author: Alberto Olivares Alarcos
Director: Horacio Rodríguez Hontoria

ABSTRACT

Processing medical data is an active research task in both industry and academia. Medical data can be found in different formats: textual, taxonomic, chemical structures, etc. However, most of that data is in textual format or contain semantic information (medical records and reports, articles, etc.), so that the processing of those data commonly includes the use of Natural Language Processing (NLP) techniques. There exists a large list of applications within the medical domain in which NLP becomes essential. Just to cite some of them: Clinical Decision Support, Medical Question Answering, Semantic tagging of medical categories or Metrics in Ontologies in the Medical Domain. The computation of semantic similarity or distance measures between medical entities becomes essential for many of those tasks.

In this thesis, three different similarity measures between medical entities (drugs) have been implemented. Each of those measures have been computed over one or more dimensions of the drugs: textual, taxonomic and molecular information. All the information has been extracted from the same resource, the DrugBank database.

Text similarity is the task of determining the degree of similarity between two texts. Texts length can vary from single words to paragraphs to complete novels or even books. In this work, drugs were represented in a vector space model, which is an algebraic model for representing text documents, where similarities can be computed. In particular, three data fields from the DrugBank database: description, indication and pharmacodynamics –all expressed in natural language– were concatenated and, after removing stop words and transforming to lowercase, their term frequency-inverse document frequency (tf-idf) representation was computed. In this case, each document used to compute the tf-idf is the concatenation of the textual fields of each drug, while the corpus is formed by all those documents as a whole. The obtained result is a sparse matrix in which the Euclidean distance is meaningless. A dimension reduction based on LSA is performed. The Euclidean distance is then computed over the reduced data, then, the similarity is obtained from the distance.

Topological similarity is the task of determining the degree of similarity between two taxonomic or ontological concepts/entities. Essentially, there are two sorts of approaches: Edge-based (which use the edges and their types as the data source) and Node-based (in which the main data sources are the nodes and their properties). The DrugBank database contains two kinds of different taxonomic structures: a set of fields named 'Classification' and the ATC Codes¹. In this project, the tag 'Classification' is used to build a graph, which is used to compute the distance between drugs as the shortest path. The similarity measure is obtained from the distance.

¹The Anatomical Therapeutic Chemical (ATC) Classification System is used for the classification of active ingredients of drugs according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties.

Measures of structural similarity play an important role in chemoinformatics for applications such as similarity searching, database clustering and molecular diversity analysis. The importance of structural similarity derives in large part from the Similar Property Principle, which states that molecules that are structurally similar are likely to have similar properties. In our approach, a molecular structure similarity measure is computed using the Tanimoto (Jaccard) coefficient over the representation of the molecules, specifically, the 2D fingerprints.

In order to study how good is each of the similarity measures, two different evaluations have been performed: indirect and direct. The indirect evaluation is based on clustering the drugs and evaluating how good the obtained clusters are. The direct evaluation is done over the similarities, comparing them with a ground truth annotated by hand by experts in the domain. A whole analysis of the obtained results is performed and written in this document.

The thesis is closed with our conclusions and the statement of all contributions of this work. Some future lines of work are also included in order to show the path which could follow our efforts.

ACKNOWLEDGEMENTS

José Luis Sampedro nos dijo a todos, escribiendo, pero nos dijo, que uno no escribe lo que ha vivido, vive lo que ha escrito. Cada letra de este documento cuenta una historia, la historia del segundo en que fue escrita. Segundo a segundo, escribí esta tesis de master de manera que llegué a sentirla, a respirarla, a vivirla. Segundo a segundo, escribí esta tesis de master sabiendo que no era yo solo quien escribía, pues sois muchos los que hacéis que mis manos se muevan, que mi cerebro funcione, que mi corazón no se muera. Ahora, que miro hacia atrás sabiendo que debo caminar hacia delante. Ahora, sólo puedo sentirme agradecido, realmente agradecido...

Agradecido..por haber podido conocer a Horacio, mi supervisor, quien me ha guiado de manera magistral y calmada, haciendo uso de toda su experiencia y conocimiento. Tras unos minutos conversando en su despacho sentí el palpito de compartir esta aventura con él. Ahora sé con hechos empíricos que ese pellizco, ese palpito, era cierto. Gracias por todo, Horacio, es mucho lo que me llevo de ti.

Agradecido..por las raíces sobre las que me alcé..unas raíces ancladas a un lugar donde la tierra es roja..los gigantes siguen alzándose desafiantes..y el viento sopla tan fuerte que hasta mueve a los gigantes..una tierra roja que fluye en mi roja sangre, que late en cada una de mis venas..

Agradecido..por las alas que un día me arrancaron de mis raíces y me llevaron tan lejos que hasta sentí olvidarlas..las alas que me hablaron de vientos..de sueños..haciéndome de mi destino, al mismo tiempo, servidor y dueño..

Agradecido..a todos vosotros..por ser mis raíces..por formar parte de mis alas..

CONTENTS

1	Scope and contextualization	1
1.1	Introduction	1
1.2	Motivation	3
1.2.1	Why similarity measurements?	3
1.2.2	Difficulty of the problem	4
1.2.3	Conclusion	6
1.3	Proposed Methodology	6
1.3.1	Text Based Similarity	7
1.3.2	Taxonomy Based Similarity	8
1.3.3	Molecular Structure Based Similarity	9
1.3.4	Evaluation	10
1.4	Structure of the document	12
2	Background and Theory	13
2.1	Natural Language Processing in the Medical Domain	13
2.1.1	Tasks	13
2.1.2	Issues on Processing Medical Texts	17
2.1.3	Genres	18
2.1.4	Resources	18
2.2	Similarity Measurements In Natural Language Processing	22
2.2.1	Distance vs Similarity	22
2.2.2	Applications	23
2.2.3	Relevant Information	24
2.2.4	A suit of methods and similarities	24
2.3	Clustering	25
2.3.1	Cluster Models	26
2.3.2	Cluster Similarity	27
2.3.3	Clustering Algorithms	27
2.3.4	Clustering Evaluation	29

2.4	Programming Tools	30
2.4.1	Jupyter Notebook	30
2.4.2	Libraries	30

3 Measuring similarity between drugs 33

3.1	DrugBank	33
3.1.1	Database Fields	34
3.1.2	Available Files	35
3.1.3	Classification Field	37
3.1.4	ATC Code	38
3.1.5	Discussion	38
3.2	Text Based Similarity	39
3.2.1	Data representation	40
3.2.2	Sparseness as a problem	40
3.2.3	Dimensional reduction as a solution	40
3.2.4	Similarity measure	42
3.3	Taxonomy Based Similarity	42
3.3.1	Data representation	42
3.3.2	Similarity measure	45
3.4	Molecular Structure Based Similarity	49
3.4.1	Data Format	49
3.4.2	Data representation	51
3.4.3	Similarity measure/coefficient	52
3.5	Evaluation Setup	53
3.5.1	Clustering	53
3.5.2	Ground Truth	56

4 Experiments and Analysis 61

4.1	General Experimental Setup	61
4.2	Text Based Similarity	62
4.2.1	Experimental Setup	63
4.2.2	Similarity Matrix	63
4.2.3	Indirect Evaluation: Clustering	63
4.2.4	Direct Evaluation: Ground Truth	69
4.3	Taxonomy Based Similarity	70
4.3.1	Experimental Setup	71
4.3.2	Similarity Matrix	71
4.3.3	Indirect Evaluation: Clustering	71
4.3.4	Direct Evaluation: Ground Truth	75

4.4	Molecular Structure Based Similarity	79
4.4.1	Experimental Setup	79
4.4.2	Similarity Matrix	79
4.4.3	Indirect Evaluation: Clustering	81
4.4.4	Direct Evaluation: Ground Truth	89
5	Conclusion	91
5.1	Statement and Contributions	91
5.2	Conclusions	92
5.2.1	Clustering Evaluation: Conclusions	92
5.2.2	Ground Truth Evaluation: Conclusions	93
5.3	Future Work	94
	Bibliography	97

LIST OF FIGURES

1.1	Triangular Inequality: visual intuition (distance).	3
3.1	The SVD decomposition of an $n \times d$ matrix.	41
3.2	Example of a subgraph of the total graph built in our project. The drugs used for this example are: Acetaminophen and Acetylsalicylic acid . . .	44
3.3	Example of a subgraph of the total graph built in our project in which an unweighted distance path is computed between two drugs. The drugs used for this example are: Acetaminophen and Acetylsalicylic acid . . .	47
3.4	Example of a subgraph of the total graph built in our project in which a weighted distance path is computed between two drugs. The drugs used for this example are: Acetaminophen and Acetylsalicylic acid	48
3.5	Molecular Structure in 2D of the Acetaminophen drug.	50
3.6	Molecular Structure in 3D of the Acetaminophen drug.	50
3.7	A comparison of the clustering algorithms in scikit-learn library (Python)	54
3.8	Three molecule-pairs with the corresponding fractions of YesNo responses to the question: 'Are these molecules similar?' The similarity values in the right-hand column were computed by the authors using the Tanimoto coefficient and ECFP4 fingerprints.	57
3.9	Extract of the CSV file used to evaluate our similarities against the proposed ground truth.	58
4.1	Distribution of the first level of all ATC Codes contained in DrugBank.	62
4.2	Similarity matrix based in text mining. The textual data has been reduced from original number of features to 100 using LSA.	64
4.3	Similarity matrix based on text mining ordered using the clusters. The textual data has been reduced from original number of features to 100 using LSA.	65
4.4	Distribution of the first level of all ATC Codes for the 1,661 drugs used in the textual mining experiment.	65
4.5	Distribution of the first level of all ATC Codes for the drugs contained within the Clusters 0-5 for the text experiment.	67
4.6	Distribution of the first level of all ATC Codes for the drugs contained within the Clusters 6-13 for the text experiment.	68
4.7	Similarity matrix based on taxonomy for the case: unweighted graph. .	72
4.8	Similarity matrix based on taxonomy for the case: weighted graph. . .	72
4.9	Similarity matrix based on taxonomy ordered using the clusters for the case: unweighted graph.	73

4.10	Similarity matrix based on taxonomy ordered using the clusters for the case: weighted graph.	74
4.11	Distribution of the first level of all ATC Codes for the 1,661 drugs used in the taxonomic experiment.	74
4.12	Distribution of the first level of all ATC Codes for the drugs contained within the Clusters 0-5 for the taxonomy experiment.	76
4.13	Distribution of the first level of all ATC Codes for the drugs contained within the Clusters 6-13 for the taxonomy experiment.	77
4.14	Similarity matrix based on molecular structure similarity. The molecular structure has been represented using ECFP fingerprints (1,024 bits).	80
4.15	Similarity matrix based on molecular structure similarity. The molecular structure has been represented using MACCS fingerprints (167 bits).	81
4.16	Similarity matrix based on molecular structure similarity ordered using the clusters. The molecular structure has been represented using ECFP fingerprints (1,024 bits).	82
4.17	Similarity matrix based on molecular structure similarity ordered using the clusters. The molecular structure has been represented using MACCS fingerprints (167 bits).	83
4.18	Distribution of the first level of all ATC Codes for the 8,738 drugs used in the molecular structure experiment.	83
4.19	Distribution of the first level of all ATC Codes for the drugs contained within the Clusters 0-5 for the molecular structure experiment. The clustering was done using the similarity matrix computed with the ECFP fingerprints.	85
4.20	Distribution of the first level of all ATC Codes for the drugs contained within the Clusters 6-13 for the molecular structure experiment. The clustering was done using the similarity matrix computed with the ECFP fingerprints.	86
4.21	Distribution of the first level of all ATC Codes for the drugs contained within the Clusters 0-5 for the molecular structure experiment. The clustering was done using the similarity matrix computed with the MACCS fingerprints.	87
4.22	Distribution of the first level of all ATC Codes for the drugs contained within the Clusters 6-13 for the molecular structure experiment. The clustering was done using the similarity matrix computed with the MACCS fingerprints.	88

LIST OF TABLES

1.1	SNOMED CT top categories	2
3.1	First Level ATC-code Meaning	39
3.2	Graph information for both cases: unweighted and weighted.	45
3.3	First Level ATC-code Meaning	56
4.1	Direct Evaluation against a ground truth of the Text Based Similarity .	70
4.2	Direct Evaluation against a ground truth of the Taxonomy Based Similarity	78
4.3	Direct Evaluation against a ground truth of the Molecular Based Similarity	90

SCOPE AND CONTEXTUALIZATION

This chapter introduces our project to the reader. Here we provide some background about the involved topics and the motivation behind the project. The project's objectives, as well as the proposed approach are stated. Finally, we show the structure of the document.

1 1

Introduction

We address the task of obtaining similarity measurements among medical entities, specifically, drugs. The term 'medical entities' includes several relevant concepts one can find within the medical domain. There is not one but several possible classifications for those entities. However, a good and well-known classification of medical terms, accepted by the experts in the domain, is the one provided by SNOMED CT¹ [Donnelly, 2006]. SNOMED Clinical Terms are a systematically organized (as a taxonomy) computer processable collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting. Of course, there exist other classifications: Wikipedia (although Wikipedia is not domain specific, it provides a rich coverage of the medical domain), MeSH² [Lipscomb, 2000], UMLS³ [Bodenreider, 2004] or other medical resources. Nevertheless, SNOMED CT is considered to be the most comprehensive, multilingual clinical health care terminology in the world. The top SNOMED's categories are shown in the Table 1.1. In our project, we are interested in Drugs, which in the cited classification is named as 'Pharmaceutical/biologic products'.

Processing medical data is an active research task in both industry and academia. Medical data can be found in different formats: textual, taxonomic, chemical structures, etc. However, most of that data is in textual format or contain semantic information (medical records and reports, articles, etc.), so that the processing of those data commonly includes the use of Natural Language Processing (NLP) techniques. Again, we can find plenty of applications within the medical domain in which NLP becomes essential. Just to cite some of them:

¹<https://www.snomed.org>. Last visit: April 2018.

²<https://www.nlm.nih.gov/mesh/>. Last visit: April 2018.

³<https://www.nlm.nih.gov/research/umls/>. Last visit: April 2018.

Pharmaceutical / biologic product	Physical force
Special atomic mapping values	Clinical finding
Substance	Special concept
Observable entity	Procedure
Qualifier value	Linkage concept
Environment or geographical location	Physical object
Situation with explicit context	Organism
Body structure	Event
Staging and scales	Social context

Table 1.1: SNOMED CT top categories

Clinical Decision Support [Demner-Fushman et al., 2009], Medical Question Answering [Goodwin and Harabagiu, 2016], Semantic tagging of medical categories [Goeuriot et al., 2015, Vivaldi and Rodríguez, 2015] or Metrics in Ontologies in the Medical Domain [Melnikov and Vorobkalov, 2014].

In this project, our aim is to make use of medical data in different formats in order to extract information which let us compute distinct sorts of similarity measurements among drugs. Those similarity measures can play an important role in NLP tasks.

Although there is no definitive consensus of axioms defining a similarity measure, we will adopt the definition stated in [Gower, 1971].

Definition 1.1.1. *Similarity Measure* Let D be a set of items represented in an euclidean space and let $S : D \times D \Rightarrow \mathbb{R}$. S is a similarity measure if it satisfies the following properties:

- Boundary conditions : there are two x numbers a and b such that $\forall x,y : 0 \leq S(x,y) \leq 1$
- Symmetry $\forall x, y : S(x, y) = S(y, x)$
- Identity and indiscernibility: $\forall x,y : x = y \Leftrightarrow , S(x,y) = 1$
- Metric : S is positive semi-definite (PSD)

Sometimes, it is also considered the constrain named as *triangular inequality*, which is not considered within this project since it is not needed. The inequality states:

$$\forall x, y, z : S(x, y) + S(y, z) \leq S(x, z) \quad (1.1)$$

A visual intuition of the the cited inequality is provided in the Figure 1.1. Please, note that the image shows the inequality from the perspective of distances, that is the reason why the inequality's symbol is the opposite. However, the principle is the same. We use distances for better visualization, in fact, simple mappings can be used between distances and similarities (see Section 2.2.1).

The problem of computing the similarity can be faced at *type level* (the concept) or at a *token level* (mentions). The distinction is relevant when the word form of the mentions are polysemic (as is the frequent case of acronyms). In this project, however, we focus on the computation of similarity following the first interpretation (the concept, not the mention).

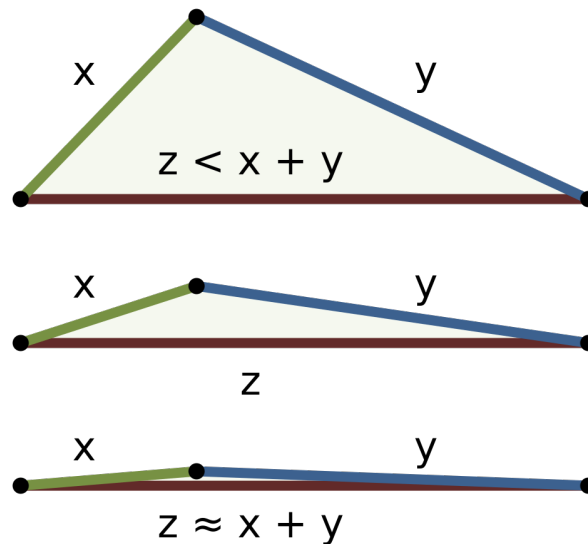


Figure 1.1: Triangular Inequality: visual intuition (distance).

1 2

Motivation

This section is devoted to make clear which is the interest and motivation behind this project. In the following sections we analyze why computing similarity measurements is important and which are the problems one faces when addressing the approach. Finally, we show a conclusion which puts all the pieces together.

1 2 1 Why similarity measurements?

Similarity is all around us, we can see it in several events of our daily life: searching for new music we could like, making decisions similar to previous ones, etc.

Every time we Humans make a decision taking into account previous events, we use any similarity measure to infer which actions we should perform to achieve any desired result. Specifically, we compare previous actions which led us to successful situations with the set of possible actions we could perform now. Let's consider a use case related to the task 'Clinical Decision Support'. A doctor has a patient with a specific illness which in other patients was treated using a concrete drug. If now the illness persists, the doctor will look for a similar drug to treat the patient. Of course, knowing the similarity among too many drugs can be unfeasible for a human and here is where computing similarities can help.

Previous case is just an example of possible application in which computing similarities among medical entities can be useful. However, there exist several tasks on the inside of the medical domain in which similarity measurements are used. For instance, Finding Patterns in Annotation Graphs [Benik et al., 2012a, Saha et al., 2010], which is based on a complementary methodology of graph summarization and dense subgraphs. For a graph G , dense subgraphs are highly connected subgraphs of G that are almost cliques. The elements of a graph summary correspond to a pattern. Another example is the task of Semantic tagging of medical categories [Yeganova et al., 2012, Goeriot et al., 2015, Vivaldi and Rodríguez, 2015], in which

similarity measurements are used to tag and group medical entities automatically.

1 2 2 Difficulty of the problem

The difficulty of the problem resides in the heterogeneity of the domain. On the one hand, there exist a large list of different representations of the information: textual (formal and informal), chemical, etc. On the other hand, all medical entities have several properties which can be used to extract information: name, description, etc. Of course, coming back to our task, the high dimensionality of the domain, makes difficult to compute similarity between drugs. In the upcoming paragraphs, we explain in detail the two main difficulties which are present in our task.

Variety of Genres

As said before, the heterogeneity of the genres (sorts of resources) we find within the medical domain is overwhelming. A list of some possible genres is explained below:

- **Electronic Medical/Health Records.** [Vasiljeva and Arandelovic, 2016, Vasiljeva and Arandjelović, 2017] Electronic Health Record (EHR), or electronic medical record (EMR), is the systematized collection of patient and population electronically-stored health information in a digital format. EHRs may include a range of data, including demographics, medical history, medication and allergies, immunization status, laboratory test results, radiology images, vital signs, personal statistics like age and weight, and billing information.
- **Medical books and articles.** This sort of documents are usually formal and contain a lot of concrete vocabulary, which might be a difficulty for a NLP system.
- **Social media.** [Beykikhoshk et al., 2015, Nikfarjam et al., 2015, Pierce et al., 2017] Medical domain is a hot topic on the Internet, one can find plenty of forums, blogs and unofficial sources of information related to this domain.
- **Wikipedia pages.** Wikipedia is a huge source of information for any domain, including the medical domain.
- **Taxonomies.** Some resources organize drugs into a taxonomy which can be easily translated into similarity measurement by knowing the paths (relationships) among the the drugs. DrugBank, the resource used within this project, has two different taxonomies: one based on the relation *IS-A* and another one based on the ATC Codes⁴.
- **Prospects.** Again, a textual document containing information potentially useful to compute the similarity.
- **Clinical trials.** [Arandjelović, 2015, Arandjelović, 2017] Clinical trials are experiments or observations done in clinical research. Such prospective biomedical or behavioral research studies on human participants are designed to answer specific questions about biomedical or behavioral interventions, including new

⁴The Anatomical Therapeutic Chemical (ATC) Classification System is used for the classification of active ingredients of drugs according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties.

treatments (such as novel vaccines, drugs, dietary choices, dietary supplements, and medical devices) and known interventions that warrant further study and comparison.

Apart from the problem of heterogeneity, some of the previous resources present difficulties for being used in a task related to NLP, in general.

For instance, textual resources might be written by doctors who can have their own jargon (depending on the specialty) or can write with a lot of abbreviations and/or acronyms. Another difficulty can be found in textual radiological reports, in which the lack of the image can be relevant.

Another example is the context or the frame in which the text is found. On the web, the expressions and vocabulary are normally less formal than in prospectuses, books or articles, what is something to consider when using NLP techniques.

The main resource used in this project is the database DrugBank. Of course, there were other options but, as we discuss during the Section 3.1.5, DrugBank is the more suitable to us.

Several properties to choose from

Every time a similarity between entities of any sort is computed, it is necessary to do so with respect a specific property or group of properties. In our case, we look for computing the similarity between drugs, where several dimensions are found. A fact which makes that computation a difficult process, since a pair of drugs can be similar when considering one of their properties but completely different if the chosen property is another one. Some of this properties are included into the main resource we use in this project, DrugBank. Just to list some of the dimensions or properties we can use to compute similarity between drugs:

- **Name.** There are different kind of names: Chemical, Generic (nonproprietary) or drug brands. In some cases, the name of a drug gives information about the family of the drug. For instance, the generic names usually indicate via their stems what drug class the drug belongs to. For example, one can tell that *aciclovir* is an antiviral drug because its name ends in the -vir suffix.
- **Description.** Description of the drug describing general facts, composition and/or preparation.
- **Pharmacodynamics.** Description of how the drug works at a clinical or physiological level.
- **Indication.** Description or common names of diseases that the drug is used to treat.
- **Chemical (Molecular) Structure.** Similar chemical compounds are meant to show similar effects. Nevertheless, this statement is not always true.
- **Classification.** This is a relevant field for us, so we explain it in detail in the subsection 3.1.3. Used in order to compute the taxonomy based similarity measure (see Section 3.3).
- **ATC Code.** This is a relevant field for us, so we explain it in detail in the subsection 3.1.4. Used to evaluate the three computed similarity measures.

The items above show just a short list of possible drug properties to be used to compute similarity between drugs, but there exist much more.

Apart from the problem of high dimensionality of properties, some of them, while useful for the desired task, present some drawbacks. For instance, in the case of using the name of the drugs, we cannot ensure that all the drugs follow the rules for the prefixes and suffixes, so we cannot perfectly group all known drugs by just knowing their names. Another problem is that in the medical domain is rather common the use of acronyms. Sometimes, one acronym could be referred to different drugs (polysemy) what makes more difficult the task of identifying similarities.

In each of the proposed cases we find specific problems which would complicate the computation of the similarity measure, a fact to take into account.

1 2 3 Conclusion

We can claim that the problem addressed within this project arouses enough interest to be a research topic and this can be proved by the following statements:

1. The applicability of the approach in the medical domain.
2. The scope of the approach goes further than our proposal. Computing similarity measurements among other medical entities (body parts, illnesses, etc.) might be equally useful. Thus, our approach not only is interesting because its possible applications but also because it could be extended.
3. The problem presents several difficulties which make it challenging enough to be a research branch.

1 3

Proposed Methodology

We propose the implementation of various similarity measurements applied to drugs. Specifically, a total of three similarities have been implemented within the development of this project, each of them based on one dimension or property of the drugs. In particular, we have used: textual information, the molecular structure and the taxonomic structure of the drugs. The implementation of these similarities, can be found on a free access repository on GitHub created by the author of this thesis⁵.

In order to evaluate how good the similarities are, two different evaluations are performed, one of them indirect, the another one direct. In the case of the indirect evaluation, we have used the similarities to cluster the drugs and then, we have evaluated the clustering. For the direct evaluation, we have used an external ground truth. The evaluation is performed for each of the three measurement.

The data used for the experiments come from the same resource, the database DrugBank. It is a unique bioinformatics/cheminformatics resource that combines detailed drug (i.e. chemical) data with comprehensive drug target (i.e. protein) information [Wishart et al., 2006]. Specifically, we have used the latest version, DrugBank 5.0 [Wishart et al., 2017]. In the section 3.1 there is a complete section devoted to DrugBank, as well as a brief discussion why we have chosen it instead

⁵<https://github.com/albertoOA/Medical-Entities-Similarity-Measurements>

of others. We can advance now that the main reason is that DrugBank is the most complete database about drugs which there exists nowadays.

The latest release of DrugBank before April 2018 (version 5.0.11, released 2017-12-20) contains 11,002 drug entries including 2,503 approved small molecule drugs, 943 approved biotech (protein/peptide) drugs, 109 nutraceuticals and over 5,110 experimental drugs. Additionally, 4,910 non-redundant protein (i.e. drug target/enzyme/transporter/carrier) sequences are linked to these drug entries. Each DrugCard entry contains more than 200 data fields with half of the information being devoted to drug/chemical data and the other half devoted to drug target or protein data. Some of those fields are textual, like the ones used in the text based similarity explained in the sub-section 1.3.1. Some other fields are related to the chemical structure of a drug and have been used in the experiment explained in the sub-section 1.3.3. Some of the fields we can find within the database are listed below:

- Description (textual)
- Indication (textual)
- Pharmacodynamics (textual)
- Name
- Kingdom
- Synonyms
- Brand names

Details of these fields are presented in Section 3.1.1.

1 3 1 Text Based Similarity

Text similarity is the task of determining the degree of similarity between two texts. Texts length can vary from single words to paragraphs to complete novels or even books. In our case, the texts are a concatenation of different textual fields extracted from the DrugBank database. Since the way of computing the text-based similarity lies on the bag of words (BoW)⁶ paradigm, simple concatenation of textual fields seems to be a good choice (we do not care about the order of the words, just if they appear or not). Single words constitute a special case of text similarity which is commonly referred to as the task of computing word similarity [Zesch and Gurevych, 2010] and is not the focus of this project.

The computation of text similarity is a very difficult task for machines. This is mainly due to the enormous variability in natural language, in which texts can be constructed using different lexical and syntactic constructions. Even so, computing text similarity has been for several years a fundamental means for many NLP tasks and applications. Nowadays, still a lot of works are devoted to this topic [Kenter and De Rijke, 2015, Kashyap et al., 2016, Ho et al., 2018].

⁶The bag-of-words model is a simplifying representation used in NLP and information retrieval (IR). Also known as the vector space model. In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity.

Our aim is to find a measure of similarity (or dissimilarity) among the drugs found in DrugBank by means of text similarity. To this purpose, the drugs were represented in a vector space model, which is an algebraic model for representing text documents and, thus, similarities can be computed in this space. To obtain the vector space model representation of the drugs, the data fields: description, indication and pharmacodynamics –all expressed in natural language– were concatenated and, after removing stop words and transforming to lowercase, their term frequency-inverse document frequency (tf-idf) representation was computed. In this case, each document used to compute the tf-idf is the concatenation of the textual fields of each drug, while the corpus is formed by all those documents as a whole. Thus, the data is represented as the matrix $M \in \mathbb{R}^{n \times d}$, where n is the number of drugs and d the number of words in the whole corpus.

Usually, the number of terms within a corpus is large, this together with the fact that only few terms appear in a specific document give room to a sparse matrix. The high dimensionality and sparseness of the matrix M entail to a well-known phenomenon called 'curse of dimensionality'. In a nutshell, we lose statistical significance and the Euclidean distance becomes meaningless.

Reducing the dimension of the vector space model we have computed is the solution proposed in this work. Specifically, we use the technique that in Information Retrieval is known as Latent Semantic Indexing (LSI) [Dumais et al., 1995], for us, Latent Semantic Analysis (LSA) [Deerwester et al., 1990]. LSA uses Singular Value Decomposition (SVD) to find the most discriminative features of our data vectors. As a result, we obtain a representation of our data in a reduced dimensional space in return for losing part of the information. The similarity matrix is computed using the Euclidean distance over the dimensionally reduced data.

1.3.2 Taxonomy Based Similarity

The DrugBank database contains two kinds of different taxonomic structures: a set of fields named 'Classification' and the ATC Codes⁷. The taxonomy contains implicit information about the similarity of the drugs we can use for our purpose. For this project, we have chosen to use the first one (Classification) to build a graph which is used to compute the similarity among the drugs. The second graph (ATC Codes) is used to evaluate the result. The classification field of DrugBank has 5 levels in total, enumerated from the highest to the lowest:

- Kingdom - Organic or Inorganic
- Classes - drug classes form the major component of the classification system. Drugs with the same class are considered structurally similar.

The Classes are divided into:

- SuperClass, for example - "Organic Acids"
- Class, for example - "Carboxylic Acids and Derivatives"
- SubClass, for example - "Amino Acids, Peptides, and Analogues"

⁷The Anatomical Therapeutic Chemical (ATC) Classification System is used for the classification of active ingredients of drugs according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties.

- DirectParent, for example - "Peptides" (can coincide with SubClass)

In our approach, similarity between drugs is computed using the graph structure in which they are organized. Thus, it is logically inevitable for us to build a graph in which the nodes are the drugs and the edges are the relationship between them. The semantics of our taxonomy has only one sort of relationship: 'is-a' relationship, (e.g. Acetaminophen is-a SubClass of Benzenoids, or which is the same, Benzenoids is-a SuperClass of Acetaminophen).

We have used the classification tag in the Drugbank database to construct 2 trees of 6 levels (depth equals to 5) which would connect the drugs in the database through undirected edges. Two different cases were contemplated: unweighted and weighted graphs. On the one hand, in unweighted graphs all the edges have the same meaning and value. On the other hand, in weighted graphs the cost of moving from one node to another is different depending on the level of the taxonomy in which the nodes are. This is to say, the edges between levels of the taxonomy imply a higher cost than edges between the same level. The distance between drugs is calculated as a shortest path distance. For the case of the weighted graph, the higher the level of the closest common ancestor in the tree, the higher the weight for the distance.

The motivation behind having two trees instead of one is because the drugs belong to either Organic or Inorganic kingdom, so we have not contemplated the most general class 'Drug'. Thus, we have decided that the path between those kingdoms should not exist, because of the very nature of the taxonomy (no or very little information gain). Additionally, introducing full connectivity (any drug can be reached from any drug in the database), by adding a common root, drastically increases computation time.

There are three main approaches to compute distances among concepts organized in a taxonomic structure: path-based (weighted and unweighted), density-based and information/content-based. In this project, we use the first one in its two forms (when the graph's edges are weighted and not).

Specifically, compute the distance between every pair of drugs as the length of the path between them. In the cases in which there is no path, we set the distance to -1. There exist several ways of turning those distances into similarities though, we have chosen the method proposed by Leacock and Chodorow [Leacock and Chodorow, 1998]. The Leacock and Chodorow Similarity between two nodes of a graph (drugs, in this case, $d1$ and $d2$) is as follows:

$$Sim(d1, d2) = -\log\left(\frac{length}{2D}\right)$$

Where *length* is the length of the shortest path between the two concepts (using node-counting) and *D* is the maximum depth of the taxonomy. Based on this measure, the shortest path between two concepts of the ontology restricted to taxonomic links is normalized by introducing a division by the double of the maximum hierarchy depth.

1.3.3 Molecular Structure Based Similarity

Measures of structural similarity play an important role in chemoinformatics for applications such as similarity searching, database clustering and molecular diversity analysis.

The importance of structural similarity derives in large part from the Similar Property Principle, which states that molecules that are structurally similar are likely

to have similar properties [Johnson and Maggiora, 1990].

The main three elements of any similarity measure based on Molecular Structure are:

- **Representation or Descriptor.** It is used to characterize the two molecules that are being compared. Among all the possible descriptors we use the fingerprints⁸.
- **Weighting Scheme.** It is used to reflect the relative importance of different parts of the representation. No weights are used in this project.
- **Similarity Coefficient.** It is used to quantify the degree of resemblance between two appropriately weighted structural representations. In our case, we use the Tanimoto (Jaccard) Coefficient.

In our approach, we first calculate the fingerprints of each drug, using the information about the Molecular Structure which DrugBank contains. Although molecular description can be obtained in two or three dimensions, we used 2D fingerprints since the number of drugs with 3D information is limited in DrugBank and actually, even though it does make a difference, there is not any instance of 3D representation as well established as the fingerprints in the case of 2D representations [Willett, 2014].

Using the fingerprints, we compute the similarity among all of them using the Tanimoto Coefficient. The computation of the Tanimoto Coefficient for two binary vectors (a and b) of length k is defined as:

$$\frac{\sum_{j=1}^k a_j \times b_j}{(\sum_{j=1}^k a_j^2 + \sum_{j=1}^k b_j^2 - \sum_{j=1}^k a_j \times b_j)} \quad (1.2)$$

Another important issue to address, is how to actually represent the Molecular Structure of a chemical compound so that a computer can process it efficiently. Normally, the Molecular Structure is represented by well-known methods like: InChi Key or SMILES. However, we cannot use those sorts of representation to compute similarity between drugs. A more efficient representation is provided by fingerprints, a list of binary values (0 or 1) which characterize a molecule. Obviously, the more bits we use, the more precise the representation is. In this project, we have explored two of the most well-known types: MACCS [Keys, 2011] and ECFPs [Rogers and Hahn, 2010].

1.3.4 Evaluation

There are two main sorts of evaluation: direct and indirect. On the one hand, a direct evaluation is the one performed directly over the result you want to study. On the other hand, an indirect evaluation is the one in which you use the obtained result to solve a task and then you evaluate the performance of it over the task. Normally, the ideal evaluation is a direct one, in which the result is compared with a '*golden standard*'. However, it is difficult to evaluate our work since there is not any clear '*golden standard*' to compare our results with.

⁸A fingerprint is a vector, each element of which describes the presence of one or more substructures in a molecule, with typical fingerprints containing a few hundred or a few thousand elements, and with two molecules being considered to be similar if their fingerprints share common values for many of the constituent elements.

In this project, we have performed two different evaluations over the computed similarities:

- **Clustering.** This is an example of indirect evaluation. We have used the similarities to cluster the drugs into groups. Then, we study the ATC Code distribution of those clusters in order to check if our similarity measurements are good.
- **Ground Truth.** This evaluation is a small direct evaluation we have done with a ground truth annotated by experts in the domain. The similarity of a list of 100 pairs of drugs were annotated by 143 experts. We have taken it from [Franco et al., 2014] and modified and adapted to our convenience. We compare the similarity computed for us with the similarity following the experts's opinion.

Clustering

As said before within the present section, the similarity measurements we have implemented are used to cluster the used drugs. This is meant to have an evaluation method to measure the quality of the computed similarities.

The type of clustering we are using is Spectral Clustering. Spectral clustering techniques make use of the spectrum (eigenvalues) of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions. The similarity matrix is provided as an input and consists of a quantitative assessment of the relative similarity of each pair of points in the dataset.

Spectral Clustering needs as input argument the number of clusters. Therefore, we need to choose that number. Drugs in DrugBank has one unique identifier which is named: 'Anatomical Therapeutic Chemical (ATC) Classification System'.

The Anatomical Therapeutic Chemical (ATC) Classification System is used for the classification of active ingredients of drugs according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties. It is controlled by the World Health Organization Collaborating Center for Drug Statistics Methodology (WHOC), and was published in 1976.

The system has a total of 5 levels, and the code consists of 7 alphanumerical characters, which can be read in the following way:

- First level: character 0 - for example 'A'
- Second level: characters 1-2, for example - 02
- Third level: character 3 - for example 'C'
- Fourth level: character 4 - for example 'A'
- Fifth level: character 5-6 - for example 04

Although each level has it's significance, we have decided to focus on the first one of the system, which determines the anatomical main group and consists of 14 categories, as shown in the Table 3.3:

Ground Truth

The external (direct) evaluation consisted of comparing the computed similarities values with the degree of similarity between 100 pairs of drugs which were annotated by experts. That annotated data has been taken from [Franco et al., 2014] and modified and adapted to our convenience.

Specifically, the ground truth was built using the opinion of 143 experts, who provided Yes/No decisions on a set of 100 DrugBank 3.0 [Knox et al., 2010] molecule-pairs. Basically, all those experts were asked to answer with Yes/No to the question: “Are these molecules similar?”. The answers were collected and a distribution of Yes/No answers was computed. In this project, we use the proportion (percentage) of ‘Yes’ answers as degree of similarity. Of course, the reader should note that the experts were not asked about the degree of similarity.

In order to evaluate how our similarity measures are related to the ground truth values, we have studied three different aspects:

- **Order.** We order the pairs by the value of their similarity in both cases, the list annotated by the experts and the one with our similarity measures. The correlation between both ordered lists is studied using Kendall’s Tau Correlation.
- **Value.** The correlation between the value of the two lists (ground truth and computed in this project) is studied using Pearson’s Correlation.
- **Threshold.** We have selected a threshold to classify the pairs of drugs into two different categories: similar and non-similar. If their similarity value is greater than the threshold, then, the drugs are similar. The threshold we have chosen is 0.85. The reason is because one of our similarity measures, the Tanimoto Coefficient, is considered relevant from that value. Then, we compute the precision and the recall of the classification process.

1 4

Structure of the document

In this document, we have structured all information we have considered relevant into five chapters. First chapter introduces the main topic of the thesis: semantic distances/similarity measures between medical entities (drugs, in this case). The motivation of doing this project and an introduction to the proposed approach are explained in that chapter. Second chapter is devoted to explain the state of the art and all the theoretical background we have researched for the development of the project. Third chapter explains in detail how we have faced the exposed problem and which is the proposed methodology for each experiment. Fourth chapter shows the obtained results and their interpretation. Finally, the fifth chapter contains the final conclusion, contributions and the future lines of work.

BACKGROUND AND THEORY

In this chapter we talk about the theoretical background related to the thesis. In order to compute similarity/distance measures between medical entities and evaluate the results, three main aspects have been researched:

- NLP in the medical domain
- Similarity measurements in NLP
- Clustering (since it is used as evaluation method)

Apart from those topics, we have also included a short section devoted to the programming tools used within the project.

2 1

Natural Language Processing in the Medical Domain

This section aims to provide useful information on the topic of Medical Data Processing. Since the topic is huge, the section is, obviously, incomplete, however, it gives a flavor of the principle aspects of the domain and it serves as a good introduction to the general framework in which this thesis belongs to.

The section is divided into several subsections. The first one accounts for the tasks involved. The second subsection presents the issues associated to the tasks. In the next subsection, we can find the different genres of medical information. Finally, we have a subsection devoted to the presentation of the resources (both data and processors).

2 1 1 Tasks

In this section we list and explain some relevant NLP tasks which can be found within the medical domain. Specifically, we have focused on tasks in which distance/similarity measurements between drugs play a relevant role. Actually, the first explained task is '*Metrics in Ontologies in the Medical Domain*'. The concept of 'tasks' is referred here to specific applications in which NLP techniques are used to solve a problem. Note that we do not talk about final user applications (products), that would be for us named as 'systems'. In some of the references we provide along this section, the authors implement complete systems which are based on solving one or several tasks we expose.

Metrics in Ontologies in the Medical Domain

With a wide range of research topics, such as Drug Discovery [Hauser et al., 2017, Moffat et al., 2017], Drug Targets [Overington et al., 2006, Santos et al., 2017] and Drug Interaction [Melnikov and Vorobkalov, 2014, Yi et al., 2017], it is not difficult to imagine that detecting the underlying patterns of the functioning of the human body with different drugs can have a wide specter of applications. An example of such as application could be detecting the area of the body where one drug has an impact and explore the possible correlations, as well as detecting different drugs that affect the same organs or organ systems. For performing such tasks an initial pre-requirement is disposing of appropriate metrics over the underground ontologies. A paradigmatic case is DrugBank.

This is totally aligned to the task we try to address in this thesis. Several similarity measurements are implemented to be used as metrics over the DrugBank database.

DrugBank is not the only ontology for which distance or similarity measures should/could be defined. Metrics over disease databases (ICD-9, ICD-10 coding), anatomical terms (Gray's coding, ATC), generic Medical Terms (SNOMED-CT, MeSH, UMLS, etc.). Some of those resources are explained in detail in Section 2.1.4.

Clinical Decision Support (CDS)

Computerized clinical decision support (CDS) aims to aid decision making of health care providers and the public by providing easily accessible health-related information at the point and time it is needed [Demner-Fushman et al., 2009]. NLP is instrumental in using free-text information to drive CDS, representing clinical knowledge and CDS interventions in standardized formats, and leveraging clinical narrative. The goal of clinical decision support (CDS) is to 'help' health professionals make clinical decisions, deal with medical data about patients or with the knowledge of medicine necessary to interpret such data.

The benefits of this topic are obvious, thus, it is a quite active research line [Roberts et al., 2015, Roberts et al., 2016, Goodwin and Harabagiu, 2016, Ran et al., 2017].

Medical Question Answering (MQA)

MQA is a concrete instance of Question Answering (QA), which is a computer science discipline within the fields of information retrieval and NLP. The task here is to automatically answer questions proposed by humans in natural language.

A QA implementation, usually a computer program, may construct its answers by querying a structured database of knowledge or information, usually a knowledge base. More commonly, QA systems can pull answers from an unstructured collection of natural language documents.

Questions occurring in clinical situations could pertain to:

- Information on particular patients
- Data on health and sickness within the local population
- Medical knowledge
- Local information on doctors available for referral
- Information on local social influences and expectation

- Information on scientific, political, legal, social, management, and ethical changes affecting both how medicine is practiced and how doctors interact with individual patients

Some questions do not need NLP and can be answered directly by a known resource. Questions about particular patients are currently answered by manually browsing or searching the Electronic Health Record (EHR). Answering these questions can be facilitated by summarization (which requires NLP if information is extracted from free-text fields) and visualization tools. Facilitating access to medical knowledge by providing answers to clinical questions is an area of active NLP research [Goodwin and Harabagiu, 2016, Goodwin and Harabagiu, 2017, Zhang et al., 2017]. This sort of task can be embedded inside QA systems, whose goal is to satisfy medical knowledge questions providing answers in the form of short action items supported by strong evidence.

Finding Patterns in Annotation graphs

Biological knowledge is increasingly being represented using graphs, e.g., protein interactions, metabolic pathways, gene regulation, gene annotation, etc. This graph representation of medical entities and their relations is used for knowledge discovering.

One way of using this information is the location, extraction, normalization and generalization of patterns of entities, relations, and events occurring in clinical narratives. [Benik et al., 2012b], for instance, exploit the NCI Thesaurus¹ for extracting meaningful patterns. [Benik et al., 2012a] use Annotation graphs (see next use case, 12.1.18) with a tool, PAnG (Patterns in Annotation Graphs), that is based on a complementary methodology of graph summarization and dense subgraphs. The elements of a graph summary correspond to a pattern and its visualization can provide an explanation of the underlying knowledge. The paper presents and analyzes two distance metrics to identify related concepts in ontologies.

This task is currently a branch of research which actually has captured the attention of many researchers [Palma et al., 2014, Grover and Leskovec, 2016, Galkin et al., 2017, Singh et al., 2017] since it has plenty of applications embedded in different systems.

Relation Extraction in the Medical Domain

Once semantically tagged, the obvious extension of processing over clinical documents consists of detecting meaningful relations between the tagged entities. A wide variety of relations exist and are relevant to be detected and extracted. For instance, in the genre of radiology reports, we could be interested on detecting a relation between a clinical finding and a body part or between an affected body part and an anchoring body part.

In Electronic Health Records (EHR) reports, we could be interested on a relation between a disease and body part, between a drug and a disease, between a drug and a dose, between a disease and a procedure, between a drug and a commercial brand denomination, and many other relations.

In the Semeval-2013 task 9 [Segura-Bedmar et al., 2014], focusing on drug-drug interaction (DDI), one of the challenges consisted of extraction of Drug-Drug inter-

¹NCI Thesaurus covers vocabulary for cancer-related clinical care, translational and basic research, and public information and administrative activities. Webpage: <https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NCI/>. Last visit: April 2018.

actions. Several types of such interactions can be considered: reinforcement, adverse effect, combination, etc.

Many approaches can be used for the task:

- Rule based
- Supervised Machine Learning
- Semi-supervised Machine Learning
- Unsupervised Machine Learning
- Clustering

Processing radiology reports

Radiology reports are probably the most studied type of clinical narrative, thus, it is an active branch of research in NLP, [Pons et al., 2016, Campos et al., 2017]. As a matter of fact, even experts from the medical domain (doctors) are exploring the potential of the use of Machine Learning techniques (including NLP) in tasks related to radiology [Lakhani et al., 2018]. This extremely important source of clinical data provides information not otherwise available in the coded data and allows performing tasks from coding of the findings and impressions, to detection of imaging technique suggested for follow up or repeated examinations, to bio-surveillance.

Radiology reports contains not only terminology from the Medical domain but also from the domain of imaging and graphical software. As narrative is related to images many special references can occur as well as challenging forms of anaphora.

Types of Radiology Reports include:

- Computed Tomography (CT)
 - CT Angiography (CTA)
 - CT Venography (CTV)

Automatic extraction of clinical trial characteristics from medical literature

Clinical trials are one of the most important sources of evidence for guiding evidence-based practice and the design of new trials. However, most of this information is available only in free text - e.g., in journal publications - which is labor intensive to process for systematic reviews, meta-analyses, and other evidence synthesis studies.

[Milian et al., 2013] face the problem of extracting eligibility criteria. Since eligibility criteria of clinical trials are represented as free text, their automatic interpretation and the evaluation of patient eligibility is challenging.

[Dunn et al., 2018] evaluated the use of document similarity methods to identify unreported links between ClinicalTrials.gov and PubMed. They extracted terms and concepts from a dataset of 72,469 ClinicalTrials.gov registrations and 276,307 PubMed articles, and tested methods for ranking articles across 16,005 reported links and 90 manually-identified unreported links. Distance measures used were Euclidean distance, cosine , and Jaccard.

Others

Of course, there are plenty of tasks within this domain, so we cannot cover all of them. We have explained in detail those we have considered more relevant nowadays or more related to our work. Some other examples of tasks are listed below:

- Representing clinical knowledge and CDS interventions in standardized formats
- Developing specific NLP processors for the Medical domain
- Clinical events monitoring
- Timeline extraction from clinical reports
- Clinical data and evidence summarization for clinicians and/or patients
- Management of patients' narratives for diagnostic and prognostic purposes

2.1.2 Issues on Processing Medical Texts

Processing Medical Texts presents several issues which are not easy to tackle. In this section, we cite and analyze just some instances of those issues among all possible which could exist:

- Difficulties in finding medical (chemical) named entities [Krallinger et al., 2015]:
 - The official IUPAC nomenclature guidelines are only partially followed in practice in the literature.
 - Chemical compounds/drugs often have many synonyms or aliases (e.g. systematic names, trivial names and abbreviations referring to the same entity).
 - Existence of hybrid chemical mentions (e.g. mentions that are partially systematic and trivial).
 - Chemical compounds are ambiguous with respect to other entities or terms (in particular abbreviations and short formula).
 - Existence of naming variation: typographical variants (alternating uses of hyphens, brackets, spacing, etc.) and alternative word order.
 - New chemical compound are discovered and described in papers every day (novel chemical names).
 - Definition of both chemical entity mention boundaries and word tokenization is complicated.
- Drug-Drug similarity. Consider the following example relevant to a group of monoclonal antibodies (mab) drugs. Ranibizumab and Bevacizumab belong to this group as their suffix “mab” points out.
- Incidental findings, i.e. asymptomatic lesions that are discovered through routine radiography.

2.1.3 Genres

- **Electronic Medical/Health Records.** [Vasiljeva and Arandelovic, 2016, Vasiljeva and Arandjelović, 2017] Electronic Health Record (EHR), or electronic medical record (EMR), is the systematized collection of patient and population electronically-stored health information in a digital format. EHRs may include a range of data, including demographics, medical history, medication and allergies, immunization status, laboratory test results, radiology images, vital signs, personal statistics like age and weight, and billing information.
- **Medical books and articles.** This sort of documents are usually formal and contain a lot of concrete vocabulary, which might be a difficulty for a NLP system.
- **Social media.** [Beykikhoshk et al., 2015, Nikfarjam et al., 2015, Pierce et al., 2017] Medical domain is a hot topic on the Internet, one can find plenty of forums, blogs and unofficial sources of information related to this domain.
- **Wikipedia pages.** Wikipedia is a huge source of information for any domain, including the medical domain.
- **Taxonomies.** Some resources organize drugs into a taxonomy which can be easily translated into similarity measurement by knowing the paths (relationships) among the the drugs. DrugBank, the resource used within this project, has two different taxonomies: one based on the relation *IS-A* and another one based on the ATC Codes².
- **Prospects.** Again, a textual document containing information potentially useful to compute the similarity.
- **Clinical trials.** [Arandjelović, 2015, Arandjelović, 2017] Clinical trials are experiments or observations done in clinical research. Such prospective biomedical or behavioral research studies on human participants are designed to answer specific questions about biomedical or behavioral interventions, including new treatments (such as novel vaccines, drugs, dietary choices, dietary supplements, and medical devices) and known interventions that warrant further study and comparison.

2.1.4 Resources

There exists a colossal amount of resources (data and processors) within the medical domain. We devote this section to the explanation of some of them we have considered relevant enough or somehow related to the work done in this thesis.

DBpedia

DBpedia³ (from 'DB' for 'database') is a project aiming to extract structured content from the information created in the Wikipedia project. This structured information is made available on the World Wide Web. DBpedia allows users to semantically query

²The Anatomical Therapeutic Chemical (ATC) Classification System is used for the classification of active ingredients of drugs according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties.

³<http://wiki.dbpedia.org>. Last visit: April 2018.

relationships and properties of Wikipedia resources (e.g. medical resources), including links to other related datasets. Tim Berners-Lee described DBpedia as one of the most famous parts of the decentralized Linked Data effort.

BioPortal

BioPortal⁴ [Whetzel et al., 2011] the world’s most comprehensive repository of biomedical ontologies. It provides access to commonly used biomedical ontologies and to tools for working with them. BioPortal allows you to:

- Browse the library of ontologies
- Search for a term across multiple ontologies
- Browse mappings between terms in different ontologies
- Receive recommendations on which ontologies are most relevant for a corpus
- Annotate text with terms from ontologies
- Search biomedical resources for a term
- Browse a selection of projects that use BioPortal resources

More than 300 ontologies are currently included allowing a federated access to their content through a *sparql* endpoint [Salvadores et al., 2012]. All information available through the BioPortal Web site is also available through the NCBO Web service REST API⁵.

WordNet

WordNet⁶ is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser (link is external). WordNet is also freely and publicly available for download. WordNet’s structure makes it a useful tool for computational linguistics and NLP.

Even though WordNet is a general purpose resource, we have included it here because we consider it one of the most used resources in NLP. Furthermore, there are several Medical related words included within WordNet.

PubMed

PubMed⁷ is a free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. The United States National Library of Medicine (NLM) at the National Institutes of Health maintains the database as part of the Entrez system of information retrieval.

From 1971 to 1997, MEDLINE online access to the MEDLARS Online computerized database primarily had been through institutional facilities, such as university libraries.

⁴<https://bioportal.bioontology.org>. Last visit: April 2018.

⁵<http://data.bioontology.org/documentation>. Last visit: April 2018.

⁶<https://wordnet.princeton.edu>. Last visit: April 2018.

⁷<https://www.ncbi.nlm.nih.gov/pubmed/>. Last visit: April 2018.

PubMed, first released in January 1996, ushered in the era of private, free, home- and office-based MEDLINE searching. The PubMed system was offered free to the public in June 1997, when MEDLINE searches via the Web were demonstrated, in a ceremony, by US Vice President Al Gore.

Medline

MEDLINE (Medical Literature Analysis and Retrieval System Online, or MEDLARS Online) is a bibliographic database of life sciences and biomedical information. It includes bibliographic information for articles from academic journals covering medicine, nursing, pharmacy, dentistry, veterinary medicine, and health care. MEDLINE also covers much of the literature in biology and biochemistry, as well as fields such as molecular evolution.

Compiled by the United States National Library of Medicine (NLM), MEDLINE is freely available on the Internet and searchable via PubMed and NLM's National Center for Biotechnology Information's Entrez system.

DrugBank

DrugBank is a unique bioinformatics/cheminformatics resource that combines detailed drug (i.e. chemical) data with comprehensive drug target (i.e. protein) information [Wishart et al., 2006]. Specifically, we have used the latest release, DrugBank 5.0 [Wishart et al., 2017].

The latest release of DrugBank (version 5.0.11, released 2017-12-20) contains 11,002 drug entries including 2,503 approved small molecule drugs, 943 approved biotech (protein/peptide) drugs, 109 nutraceuticals and over 5,110 experimental drugs. Additionally, 4,910 non-redundant protein (i.e. drug target/enzyme/transporter/carrier) sequences are linked to these drug entries. Each DrugCard entry contains more than 200 data fields with half of the information being devoted to drug/chemical data and the other half devoted to drug target or protein data.

It is the main resource utilized in our work. In the section 3.1 there is a complete section devoted to DrugBank, as well as a brief discussion why we have chosen it instead of others. We can advance now that the main reason is that DrugBank is the most complete database about drugs which there exists nowadays.

SnoMed CT

SNOMED CT⁸ or SNOMED Clinical Terms is a systematically organized computer processable collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting. SNOMED CT is considered to be the most comprehensive, multilingual clinical healthcare terminology in the world [Donnelly, 2006]. The primary purpose of SNOMED CT is to encode the meanings that are used in health information and to support the effective clinical recording of data with the aim of improving patient care. SNOMED CT provides the core general terminology for electronic health records. SNOMED CT comprehensive coverage includes: clinical findings, symptoms, diagnoses, procedures, body structures, organisms and other etiologies, substances, pharmaceuticals, devices and specimens.

⁸<https://www.nlm.nih.gov/healthit/snomedct/index.html>. Last visit: April 2018

UMLS

The UMLS⁹ [Bodenreider, 2004], or Unified Medical Language System, is a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems.

You can use the UMLS to enhance or develop applications, such as electronic health records, classification tools, dictionaries and language translators.

One powerful use of the UMLS is linking health information, medical terms, drug names, and billing codes across different computer systems. Some examples of this are:

- Linking terms and codes between your doctor, your pharmacy, and your insurance company
- Patient care coordination among several departments within a hospital

The UMLS has many other uses, including search engine retrieval, data mining, public health statistics reporting, and terminology research. UMLS contains a unique identifier, Concept Unique Identifier (CUI), which is frequently used as the facto standard identifier.

MeSH

Medical Subject Headings (MeSH) [Head-ingB, 1965] is a comprehensive controlled vocabulary for the purpose of indexing journal articles and books in the life sciences; it serves as a thesaurus that facilitates searching. Created and updated by the United States National Library of Medicine (NLM), it is used by the MEDLINE/PubMed article database and by NLM's catalog of book holdings. MeSH is also used by ClinicalTrials.gov registry to classify which diseases are studied by trials registered in ClinicalTrials.gov.

MeSH was introduced in 1960, with the NLM's own index catalogue and the subject headings of the Quarterly Cumulative Index Medicus (1940 edition) as precursors. The yearly printed version of MeSH was discontinued in 2007 and MeSH is now available online only. It can be browsed and downloaded free of charge through PubMed. Originally in English, MeSH has been translated into numerous other languages and allows retrieval of documents from different languages.

NCI Thesaurus

The NCI Metathesaurus is product of the US National Cancer Institute's Enterprise Vocabulary Service¹⁰, a collaborative effort of the NCI Center for Bioinformatics and the NCI Office of Communications. The NCI Metathesaurus is based on NLM's Unified Medical Language System Metathesaurus supplemented with additional cancer-centric vocabulary.

The public version of the NCI Metathesaurus currently contains all public domain vocabularies from the National Library of Medicine's UMLS Metathesaurus, as well as a growing number of NCI-specific vocabularies developed by the National Cancer Institute.

⁹<https://www.nlm.nih.gov/research/umls/>. Last visit: April 2018.

¹⁰<https://ncit.nci.nih.gov/ncitbrowser/>. Last visit: April 2018.

BioScope corpus

BioScope¹¹ contains biomedical texts annotated for uncertainty, negation and their scopes. The corpus [Vincze et al., 2008] consists of three parts, namely medical free texts, biological full papers and biological scientific abstracts. The dataset contains annotations at the token level for negative and speculative keywords and at the sentence level for their linguistic scope. The annotation process was carried out by two independent linguist annotators and a chief linguist – also responsible for setting up the annotation guidelines – who resolved cases where the annotators disagreed. The resulting corpus consists of more than 20.000 sentences that were considered for annotation and over 10% of them actually contain one (or more) linguistic annotation suggesting negation or uncertainty.

Foundational Model of Anatomy Ontology

The Foundational Model of Anatomy Ontology (FMA¹²) [Rosse and Mejino, 2008] is an evolving computer-based knowledge source for biomedical informatics; it is concerned with the representation of classes or types and relationships necessary for the symbolic representation of the phenotypic structure of the human body in a form that is understandable to humans and is also navigable, parseable and interpretable by machine-based systems. Specifically, the FMA is a domain ontology that represents a coherent body of explicit declarative knowledge about human anatomy. Its ontological framework can be applied and extended to all other species.

The Foundational Model of Anatomy (FMA) ontology is one of the information resources integrated in the distributed framework of the Anatomy Information System developed and maintained by the Structural Informatics Group at the University of Washington. The FMA is open source.

2 2

Similarity Measurements In Natural Language Processing

In statistics and related fields, a similarity measure or similarity function is a real-valued function that quantifies the similarity between two objects. Although no single definition of a similarity measure exists, usually such measures are in some sense the inverse of distance metrics: they take on large values for similar objects and either zero or a negative value for very dissimilar objects.

The concept of similarity is also related to other concepts like: proximity, affinity, distance, difference or divergence. Whenever we see any of those concepts, we will be talking about the same: how close (and far) are two or more entities.

2 2 1 Distance vs Similarity

As stated above, there is not a clear definition of similarity, so normally, it is computed as the inverse of distance metrics. Some simple methods to compute similarity from a distance are shown in 2.2.1 and 2.2.1. The expression 2.2.1 is really simple but it makes sense only if the Distance is normalized from 0 to 1. The formula 2.2.1 is more general and could be used with other distances (e.g. simple Euclidean).

¹¹Website: <http://rgai.inf.u-szeged.hu/index.php?lang=en&page=bioscope>. Last visit: April 2018.

¹²Visit <http://si.washington.edu/projects/fma>. Last visit: April 2018.

$$S(x, y) = 1 - D(x, y) \quad (2.1)$$

$$S(x, y) = \frac{1}{1 + D(x, y)} \quad (2.2)$$

We could think about the question: *'why would we use similarity if there is not a proper definition?'* There are some benefits of using Similarity instead of Distance since the Distance can only be used when some metric properties hold (see 2.2.1).

Definition 2.2.1. *Metric Properties* Constrains to be hold in the case of using a Distance measure (D).

- $\forall x : D(x, x) \neq 0$
- $\forall x, y : D(x, y) \geq 0$ when $x \neq y$
- $\forall x, y : D(x, y) = D(y, x)$ (Symmetry)
- $\forall x, y, z : S(x, y) + S(y, z) \leq S(x, z)$ (Triangular Inequality)

On the other hand, similarity can be used in more general cases:

- Function: $\text{sim} : A \times B \rightarrow S$ (where S is often $[0, 1]$)
- Homogeneous: $\text{sim} : A \times A \rightarrow S$ (e.g. word-to-word)
- Heterogeneous: $\text{sim} : A \times B \rightarrow S$ (e.g. word-to-document)
- Not necessarily symmetric, or holding triangular inequality.

2.2.2 Applications

The range of possible applications of similarity in the NLP domain is wide. In this section we just list some of them in order to give a flavor of it.

- Clustering, case-based reasoning, Information Retrieval, etc.
- Discovering related words - Distributional similarity
- Resolving syntactic ambiguity - Taxonomic similarity
- Resolving semantic ambiguity - Ontological similarity
- Acquiring selectional restrictions/preferences
- Others

2.2.3 Relevant Information

In order to compute a similarity measure, it is necessary to know the problem we try to tackle. Three main aspects have been identified:

- Content or information about compared units:
 - Words: form, morphology, Part of Speech (PoS), ...
 - Senses: synset, topic, domain, ...
 - Syntax: parse trees, syntactic roles, ...
 - Documents: words, collocations, Name Entities (NEs), ...
- Context or information about the situation in which the similarity is computed. For instance, we could have Window-based vs. Syntactic-based.
- External Knowledge: Monolingual/bilingual dictionaries, ontologies, corpora, etc.

2.2.4 A suit of methods and similarities

In this section, we analyze several methods and similarities which can be used in NLP. Note that some of them are just general purpose measurements and the use of them in the NLP domain rely upon a proper representation of the units. For instance, sometimes, we need to represent our units (which can be words) as vectors.

Vectorial Methods

This sort of methods can be used when our data has been represented in a vectorial space (units are vectors), where the distances exposed below make sense. Representing linguistic units as vectors (both, large dimensional or sparse vectors and low dimensional or dense vectors) is an active area of research [Bowman et al., 2017]. All methods explained in this section are distances, in order to compute similarities from them, it is possible to use the formula 2.2.1. In the following definitions, x_i and y_i are the two vectors representing our units (e.g. vectors representing words, texts, etc.) and N is the length of those vectors.

- Manhattan Distance (L1 Norm). $D(\vec{x}, \vec{y}) = \sum_{i=1}^N |x_i - y_i|$
- Euclidean Distance (L2 Norm). $D(\vec{x}, \vec{y}) = |\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^N |x_i - y_i|^2}$
- Cosine Distance. $D(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| \cdot |\vec{y}|} = \frac{\sum_{i=1}^N x_i y_i}{\sqrt{\sum_{i=1}^N x_i^2} \cdot \sqrt{\sum_{i=1}^N y_i^2}}$
- Camberra Distance. $D(\vec{x}, \vec{y}) = \sum_{i=1}^N \frac{x_i - y_i}{x_i + y_i}$
- Chebychev Distance. $D(\vec{x}, \vec{y}) = \max |x_i - y_i|, i = [1, n]$

Set-oriented Methods

In this case, we also have a specific representation for our units, binary-valued vectors. The similarities are computed considering the values which agree in those boolean vectors. All methods showed in this section have values in $[0,1]$, so we could compute the Distance as: $D = 1 - S$, where D is the distance and S is the similarity. In the following expressions, X and Y are binary vectors which represent our data (e.g. words, text, etc.).

- Dice. $S(X, Y) = \frac{2 \cdot |X \cap Y|}{|X| + |Y|}$
- Jaccard (Tanimoto). $S(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}$
- Overlap. $S(X, Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)}$
- Cosine. $S(X, Y) = \frac{|X \cap Y|}{\sqrt{|X| \cdot |Y|}}$

Distributional Similarity

It is a particular case of vectorial method, where the attributes of the vectors are probability distributions computed over the context of the linguistic unit. Some examples of this sort of similarity are: Relative Entropy and Mutual Information.

Semantic Similarity

Consist of the projection of words to a semantic space (concepts) where the similarities/distances are computed. It is not a straightforward task because of several reasons. On the one hand, it is not straightforward to project words, since semantic space is composed of concepts, and a word may map to more than one concept. On the other hand, it is not obvious how to compute distance in the semantic space.

Instances of semantic spaces are Ontologies (WordNet, SUMO, etc.) or Graph-like Knowledge Bases (Wikipedia).

Ready to go similarities: WordNet

WordNet provides several similarity measures already computed among all the words contained inside the database. Some of them are available through the WordNet's accessor implemented in the Python library *'nltk'*, a complete library with NLP tools.

2 3

Clustering

Clustering is the basis of one part of our evaluation framework. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense, we need a similarity measure) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

The assignment of objects to clusters can be: hard or soft. On the one hand, an assignment is hard when we assign one cluster per object. On the other hand, we say

that the assignment is soft when there is a degree of membership, thus, one same object can belong to several clusters with a specific probability.

Each cluster has a representative, which is named as Centroid, its definition is as follows:

$$\vec{\mu} = \frac{1}{|c|} \sum_{\vec{x}_{ec}} \vec{x} \quad (2.3)$$

where c is the number of clusters and \vec{x} is the vectorial representation of each object. Note that the Centroid is artificially computed from the members of the cluster (it is not a member). The Medoid is the analogue concept but with the restriction of actually being a member of the cluster. Meaning, the Medoid is the closest member to the rest of members of a cluster.

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances between cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including parameters such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It is often necessary to modify data preprocessing and model parameters until the result achieves the desired properties.

In Section 2.4, we present a list of possible clustering algorithms which are already implemented within one Python library which is used in this thesis.

2.3.1 Cluster Models

The notion of a 'cluster' cannot be precisely defined, which is one of the reasons why there are so many clustering algorithms. There is a common denominator: a group of data objects. However, different researchers employ different cluster models, and for each of these cluster models again different algorithms can be given. The notion of a cluster, as found by different algorithms, varies significantly in its properties. Understanding these 'cluster models' is key to understanding the differences between the various algorithms. Typical cluster models include:

- Connectivity models: for example, hierarchical clustering builds models based on distance connectivity.
- Centroid models: for example, the k-means algorithm represents each cluster by a single mean vector.
- Distribution models: clusters are modeled using statistical distributions, such as multivariate normal distributions used by the expectation-maximization algorithm.
- Density models: for example, DBSCAN and OPTICS defines clusters as connected dense regions in the data space.

- Subspace models: in biclustering (also known as co-clustering or two-mode-clustering), clusters are modeled with both cluster members and relevant attributes.
- Group models: some algorithms do not provide a refined model for their results and just provide the grouping information.
- Graph-based models: a clique, that is, a subset of nodes in a graph such that every two nodes in the subset are connected by an edge can be considered as a prototypical form of cluster. Relaxations of the complete connectivity requirement (a fraction of the edges can be missing) are known as quasi-cliques, as in the HCS clustering algorithm.
- Neural models: the most well known unsupervised neural network is the self-organizing map and these models can usually be characterized as similar to one or more of the above models, and including subspace models when neural networks implement a form of Principal Component Analysis or Independent Component Analysis.

2.3.2 Cluster Similarity

As it is stated before, in order to group objects into clusters, it is necessary to use a measure of closeness among those objects. Therefore, we need to compute a similarity/distance measure. Any of the similarity and/or distances metrics we have explained within the Section 2.2 could be used.

2.3.3 Clustering Algorithms

Clustering algorithms can be categorized based on their cluster model, as listed above, or based on the structure they produce. In this section, we show two main groups of algorithms based on the structure among the objects of the produced clusters.

Hierarchical Clustering

Connectivity-based clustering, also known as hierarchical clustering, is based on the core idea of objects being more related to nearby objects than to objects farther away. These algorithms connect 'objects' to form 'clusters' based on their distance. A cluster can be described largely by the maximum distance needed to connect parts of the cluster. At different distances, different clusters will form, which can be represented using a dendrogram, which explains where the common name 'hierarchical clustering' comes from: these algorithms do not provide a single partitioning of the data set, but instead provide an extensive hierarchy of clusters that merge with each other at certain distances. In a dendrogram, the y-axis marks the distance at which the clusters merge, while the objects are placed along the x-axis such that the clusters don't mix.

There are two different types of hierarchical clustering:

- **Bottom-up.** Also named 'Agglomerative Clustering', it is based on merge objects to form the clusters. At the initial state, every object is a cluster, then we group the most similar iteratively until a degree of satisfaction is reached.

- **Top-Down.** In this case we address the problem in the contrary direction, we start with a unique cluster. Then, an iterative process of division form new clusters until a degree of satisfaction is reached. This technique receives also the name of 'Divisive Clustering'.

Apart from the usual choice of distance functions, the user of this algorithm also needs to decide on the linkage criterion (since a cluster consists of multiple objects, there are multiple candidates to compute the distance) to use. Popular choices are known as single-linkage clustering (the minimum of object distances), complete linkage clustering (the maximum of object distances) or UPGMA ('Unweighted Pair Group Method with Arithmetic Mean', also known as average linkage clustering).

- **Single link.** In single-link clustering or single-linkage clustering, the similarity of two clusters is the similarity of their most similar members. This single-link merge criterion is local. We pay attention solely to the area where the two clusters come closest to each other. Other, more distant parts of the cluster and the clusters' overall structure are not taken into account. Thus, we obtain local coherence, since close objects are clustered in the same group). However, we also obtain elongated clusters because of the local nature of the criterion. Since the merge criterion is strictly local, a chain of points can be extended for long distances without regard to the overall shape of the emerging cluster. This is know as *chaining effect*.
- **Complete link.** In complete-link clustering or complete-linkage clustering, the similarity of two clusters is the similarity of their least similar members. This is equivalent to choosing the cluster pair whose merge has the smallest diameter. This complete-link merge criterion is non-local; the entire structure of the clustering can influence merge decisions. This results in a preference for compact clusters with small diameters over long, straggly clusters, but also causes sensitivity to outliers. A single object far from the center can increase diameters of candidate merge clusters dramatically and completely change the final clustering.
- **UPGMA.** Unweighted Pair Group Method with Arithmetic Mean. The UPGMA algorithm constructs a rooted tree (dendrogram¹³) that reflects the structure present in a pairwise similarity matrix (or a dissimilarity matrix). At each step, the nearest two clusters are combined into a higher-level cluster. The distance between any two clusters A and B is taken to be the average of all distances $d(x, y)$ between pairs of objects x in A and y in B , that is, the mean distance between elements of each cluster:

$$\frac{1}{|A| \cdot |B|} \sum_{x \in A} \sum_{y \in B} d(x, y) \quad (2.4)$$

In this method there is a trade-off between global coherence and efficiency.

¹³Dendogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering.

Non-hierarchical Clustering

In this type of clustering there is not a hierarchy between the objects. Within this group, we can find a large list of algorithms, depending on the foundation in which they are based to cluster the objects. Two relevant groups are listed below, however, the reader should notice that there exist much more sorts of algorithms.

All these sorts of clusterings have the following three steps:

1. Initial partition of the set in clusters based on random seeds.
 2. Iteratively refine the partition by means of reallocating objects.
 3. Stop when the cluster quality does not improve further. The cluster quality can be: group average similarity, mutual information between adjacent clusters or likelihood of data given a cluster model, among others.
- **Centroid-based Clustering.** In centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. One of the most well known methods is K-means [MacQueen et al., 1967]. When the number of clusters is fixed to k , k-means clustering gives a formal definition as an optimization problem: find the k cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized.
 - **Distribution-based Clustering.** The clustering model most closely related to statistics is based on distribution models. Clusters can then easily be defined as objects belonging most likely to the same distribution. A convenient property of this approach is that this closely resembles the way artificial data sets are generated: by sampling random objects from a distribution. A good example is the Clustering based on the Expectation Maximization Algorithm [Moon, 1996].

2 3 4 Clustering Evaluation

This section is devoted to show some approaches commonly used to evaluate the performance of a clustering and indirectly, of the similarity measure used for the clustering process. There are several techniques but all of them can be included into two groups: internal and external evaluations.

Internal Evaluation

Evaluations which can be done without the need of other external resources. The performance of the clustering is evaluated based on the data that was clustered itself.

Internal evaluation methods usually assign the best score to the algorithm that produces high intra-cluster similarity and low inter-cluster similarity, that is to say, clusters with high similarity between their own objects and low similarity between all clusters. One drawback of using internal criteria in clustering evaluation is that high scores on an internal measure do not necessarily result in effective information retrieval applications. Additionally, this evaluation is biased towards algorithms that use the same cluster model. For example, k-means clustering naturally optimizes object distances, and a distance-based internal criterion will likely overrate the resulting clustering.

Somme cluster quality measures: coherence, average internal distance, average external distance, etc.

External Evaluation

This is the case when a gold standard is available. Meaning, we have an external resource which can be compared to the clustering result. An example of this is to have a set of clusters manually annotated by experts or stemming from a well established classification.

In those cases, we can compute several cluster quality measurements. Maybe the most well-known is the metrics of *Purity* 2.5 and *Inverse Purity* 2.6.

$$P = \frac{1}{|D|} \sum_c \max_x |c \cap x| \quad (2.5)$$

$$IP = \frac{1}{|D|} \sum_x \max_c |c \cap x| \quad (2.6)$$

2 4

Programming Tools

The code implementation of this work has been done in Python¹⁴, specifically, Python 3.6. Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales.

Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library.

Python provides several tools for Artificial Intelligence tasks. In this section, we analyze some of the specific purpose libraries and tools we have used within this project. Please, note that we have used several libraries, but some of them are the common libraries included in Python, thus, we do not explain them.

2 4 1 Jupyter Notebook

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

The code implemented in this project, includes three notebooks, apart from the classes and other support codes. The notebooks present the implementation and use of each of the similarity measures we have developed. In there, we not only program but also explain the steps of our implementation, thus, the notebooks are a good resource for anyone who would be interested in the work performed in this thesis.

2 4 2 Libraries

As stated before, Python includes several libraries and tools which help you to develop more powerful codes. In this section, we analyze some of them which have been used

¹⁴<https://python.org>. Last visit: April 2018.

in the thesis.

NumPy

NumPy¹⁵ is the fundamental package for scientific computing with Python. It contains among other things:

- a powerful N-dimensional array object
- sophisticated (broadcasting) functions
- tools for integrating C/C++ and Fortran code
- useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

SciPy

SciPy¹⁶ is an open-source Python library used for scientific computing and technical computing. It contains modules for optimization, linear algebra, integration, interpolation, special functions, FFT, signal and image processing, ODE solvers and other tasks common in science and engineering.

SciPy builds on the NumPy array object and is part of the NumPy stack which includes tools like Matplotlib, pandas and SymPy, and an expanding set of scientific computing libraries.

Scikit-learn

Scikit-learn¹⁷ (formerly scikits.learn) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means, spectral clustering and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

NLTK

NLTK¹⁸ is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

¹⁵<http://www.numpy.org>. Last visit: April 2018.

¹⁶<https://www.scipy.org>. Last visit: April 2018.

¹⁷<http://scikit-learn.org/stable/>. Last visit: April 2018.

¹⁸<http://www.nltk.org>. Last visit: April 2018.

RDKit

RDKit¹⁹ is a collection of cheminformatics and machine-learning software written in C++ and Python. Among its features, we can find:

- BSD license - a business friendly license for open source
- Core data structures and algorithms in C++
- Python (2.x and 3.x) wrapper generated using Boost.Python
- Java and C# wrappers generated with SWIG
- 2D and 3D molecular operations
- Descriptor and Fingerprint generation for machine learning
- Molecular database cartridge for PostgreSQL supporting substructure and similarity searches as well as many descriptor calculators
- Cheminformatics nodes for KNIME²⁰
- Contrib folder with useful community-contributed software harnessing the power of the RDKit

¹⁹<http://www.rdkit.org>. Last visit: April 2018.

²⁰<https://www.knime.com>. Last visit: April 2018.

MEASURING SIMILARITY BETWEEN DRUGS

In this chapter, we explain in depth the core of the work developed in this thesis. First, we explain the main resource we have used: DrugBank. Then, we explain each of the similarity measures between drugs we have implemented. Finally, we tackle the explanation of the evaluation process done over our similarities.

The implementation of this work can be found on a free access repository on GitHub created by the author of this thesis¹.

3 1

DrugBank

The DrugBank² database is a comprehensive, freely accessible, online database containing information on drugs and drug targets. As both a bioinformatics and a cheminformatics resource, DrugBank combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information. Because of its broad scope, comprehensive referencing and unusually detailed data descriptions, DrugBank is more akin to a drug encyclopedia than a drug database. As a result, links to DrugBank are maintained for nearly all drugs listed in Wikipedia. DrugBank is widely used by the drug industry, medicinal chemists, pharmacists, physicians, students and the general public. Its extensive drug and drug-target data has enabled the discovery and repurposing of a number of existing drugs to treat rare and newly identified illnesses.

The latest release of DrugBank before April 2018 [Wishart et al., 2006] (version 5.0.11, released 2017-12-20) contains 11,002 drug entries including 2,503 approved small molecule drugs, 943 approved biotech (protein/peptide) drugs, 109 nutraceuticals and over 5,110 experimental drugs. Additionally, 4,910 non-redundant protein (i.e. drug target/enzyme/transporter/carrier) sequences are linked to these drug entries. Each DrugCard entry contains more than 200 data fields with half of the information being devoted to drug/chemical data and the other half devoted to drug target or protein data.

DrugBank is used in this project as the main resource. All the information about drugs which is utilized within our work, comes from this database. Please, note that

¹<https://github.com/albertoOA/Medical-Entities-Similarity-Measurements>

²See <https://www.drugbank.ca>. Last visit: April 2018.

the used version is not the latest one. There exists a new version which was published at the beginning of April, obviously, we had not time to use it and change all our analysis of the results. However, the changes are minimal, instead of 11,002 drugs now there are 11,037.

3.1.1 Database Fields

DrugBank is a detailed database on small molecule and biotech drugs. Each drug entry (“DrugCard”) includes extensive information on properties, structure, and biology (what the drug does in the body). Below you can find some definitions, and a detailed reference of the sources used for each field.

In this subsection, we analyze part of the documentation or fields which are contained in the DrugBank. We have included not only the fields we are using but also some of the ones we consider important.

Below we list part of the possible fields³ a drug can have within the data contained in DrugBank. Please, note that not all of the possible fields are available for all the drugs.

- **Drug Type.** Drugs are categorized by type, which determines their origin. Here is the list of possible types:
 - **Small Molecule.** Low molecular weight drugs (900 daltons) which are produced by chemical synthesis. These drugs have well defined structures and chemical properties. In DrugBank, some drugs larger than 900 daltons are considered small molecule drugs (such as monomers: ribo- or deoxyribonucleotides, amino acids, and monosaccharides), as long as they are chemically synthesized.
 - **Biotech.** Drugs with a biological origin (manufactured in, extracted from, or semisynthesized from biological sources). These include vaccines, blood, blood components, allergenics, somatic cells, gene therapies, tissues, recombinant therapeutic protein, and living cells used in cell therapy. Biotech drugs are also known as biopharmaceuticals or biologics.
- **Drug Group(s).** Drugs are categorized by group, which determines their drug development status. Here is the list of possible groups:
 - **Approved.** A drug that has been approved in at least one jurisdiction, at some point in time.
 - **Vet Approved.** A drug that has been approved in at least one jurisdiction, at some point in time for the treatment of animals.
 - **Nutraceutical.** A drug that is a pharmaceutical-grade and standardized nutrient (with confirmed or unconfirmed health benefits)
 - **Illicit.** A drug that is scheduled in at least one jurisdiction, at some point in time.

³For a complete list visit: <https://www.drugbank.ca/documentation#drug-cards>. *Last visit April 2018.*

- **Withdrawn.** A previously approved drug that has been withdrawn from the market in at least one jurisdiction, at some point in time. Note that because a drug can be approved in one jurisdiction, and withdrawn in another, it's possible for a drug to be in both groups.
 - **Investigational.** A drug that is in some phase of the drug approval process in at least one jurisdiction.
 - **Experimental.** A compound that has been shown experimentally to bind specific proteins in mammals, bacteria, viruses, fungi, or parasites. This includes compounds that are Pre-Investigational New Drug Applications (Pre-IND, or Discovery Phase compounds).
- **DrugBank ID.** It is the Primary Accession Number and unique identifier for a drug.
 - **Name.** Standard name of drug as provided by drug manufacturer. Note that there are other fields including synonyms and other names like 'brand names'. Used in order to compute the text based similarity measure (see Section 3.2).
 - **Description.** Description of the drug describing general facts, composition and/or preparation. Used in order to compute the text based similarity measure (see Section 3.2).
 - **Pharmacodynamics.** Description of how the drug works at a clinical or physiological level. Used in order to compute the text based similarity measure (see Section 3.2).
 - **Indication.** Description or common names of diseases that the drug is used to treat. Used in order to compute the text based similarity measure (see Section 3.2).
 - **Classification.** This is a relevant field for us, so we explain it in detail in the subsection 3.1.3. Used in order to compute the taxonomy based similarity measure (see Section 3.3).
 - **ATC Code.** This is a relevant field for us, so we explain it in detail in the subsection 3.1.4. Used to evaluate the three computed similarity measures.
 - **Chemical Formula.** Describing atomic or elemental composition.
 - **Structure.** The 2D/3D chemical structure including links to download and view the structure in various formats. In our case we are not using this field but a specific file in which we find all information related to the molecular structure of the drugs. See Section 3.4 for more detail.

3.1.2 Available Files

On account of there are a lot of information on inside of the DrugBank database, there exist several sorts of files⁴ which offer different content. In this section we talk briefly about all possible those files and we explain which ones are used in our project.

⁴See <https://www.drugbank.ca/releases/latest#full>. *Last visit, April 2018.*

- **Complete Database.** It is a XML file which contains all fields for all available drugs. An example of the DrugBank fields used in this work, extracted from the complete XML, can be seen in the Listing 3.1. This file is utilized in two of the similarities we compute (Sections 3.2 and 3.3) as well as to evaluate using the ATC Codes.
- **Structures.** It consists of a file in SDF format. The format is one of a family of chemical-data file formats developed by MDL; it is intended especially for structural information. "SDF" stands for structure-data file, and SDF files actually wrap the molfile (MDL Molfile) format. A feature of the SDF format is its ability to include associated data (not only molecular, but also general purpose data like 'drug id'). We use it for the Molecular Structure Based Similarity (see Section 3.4). The reason why we decided to use this document was because we discovered there was a specific library for Python to deal with this sort of document. That tool (RDKit) was perfectly aligned to the work we wanted to do.
- **External Links.** CSV with links to other databases. Not used in this project.
- **Protein Identifiers.** Protein identifiers include external IDs to resources such as UniProt and PDB. These downloads are divided first by protein/compound type (target, transporter, etc.). Secondly they are divided by drug group (approved, illicit, etc.). Not used in this project.
- **Target Sequences.** Not used in this project.
- **Drug Sequences.** Not used in this project.

Listing 3.1: Extract of just used fields from the full XML DrugBank database.

```
<?xml version="1.0" encoding="UTF-8"?>
<drugbank xmlns="http://www.drugbank.ca"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.drugbank.ca
http://www.drugbank.ca/docs/drugbank.xsd"
version="5.0" exported-on="2017-12-20">
<drug type="biotech" created="2005-06-13" updated="2017-11-06">
  <drugbank-id primary="true">DB00001</drugbank-id>
  <drugbank-id>BTD00024</drugbank-id>
  <drugbank-id>BIOD00024</drugbank-id>
  <name>Lepirudin</name>
  <description>Lepirudin is identical to natural hirudin except for
substitution of leucine for isoleucine at the N-terminal end of
the molecule and the absence of a sulfate group on the tyrosine at
position 63. It is produced via yeast cells. Bayer ceased the
production of lepirudin (Refludan) effective May 31, 2012.
</description>
<indication>For the treatment of heparin-induced thrombocytopenia
</indication>
<pharmacodynamics>Lepirudin is used to break up clots and to reduce
```

thrombocytopenia. It binds to thrombin and prevents thrombus or clot formation. It is a highly potent, selective, and essentially irreversible inhibitor of thrombin and clot-bound thrombin. Lepirudin requires no cofactor for its anticoagulant action. Lepirudin is a recombinant form of hirudin, an endogenous anticoagulant found in medicinal leeches.

```

</pharmacodynamics>
<atc-codes>
  <atc-code code="B01AE02">
    <level code="B01AE">Direct thrombin inhibitors</level>
    <level code="B01A">ANTITHROMBOTIC AGENTS</level>
    <level code="B01">ANTITHROMBOTIC AGENTS</level>
    <level code="B">BLOOD AND BLOOD FORMING ORGANS</level>
  </atc-code>
</atc-codes>
<classification>
  <description/>
  <direct-parent>Peptides</direct-parent>
  <kingdom>Organic Compounds</kingdom>
  <superclass>Organic Acids</superclass>
  <class>Carboxylic Acids and Derivatives</class>
  <subclass>Amino Acids, Peptides, and Analogues</subclass>
</classification>

```

3.1.3 Classification Field

The DrugBank database contains some kinds of different taxonomic structures. One of them is the field named 'Classification'. A taxonomy contains implicit information about the similarity of the drugs we can use for our purpose. For this project, we have chosen to use the Classification to build a graph which is used to compute the similarity among the drugs. The classification field of DrugBank has 5 levels in total, enumerated from the highest to the lowest:

- Kingdom - Organic or Inorganic
- Classes - drug classes form the major component of the classification system. Drugs with the same class are considered structurally similar.

The Classes are divided into:

- SuperClass, for example - "Organic Acids"
- Class, for example - "Carboxylic Acids and Derivatives"
- SubClass, for example - "Amino Acids, Peptides, and Analogues"
- DirectParent, for example - "Peptides" (can coincide with SubClass)

In our approach, similarity between drugs is computed using the graph structure in which they are organized. Thus, it is logically inevitable for us to build a graph in which the nodes are the drugs and the edges are the relationship between them.

The semantics of our taxonomy has only one sort of relationship: 'is-a' relationship, (e.g. Acetaminophen is-a SubClass of Benzenoids, or which is the same, Benzenoids is-a SuperClass of Acetaminophen).

3.1.4 ATC Code

The Anatomical Therapeutic Chemical (ATC) Classification System is used for the classification of active ingredients of drugs according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties. It is controlled by the World Health Organization Collaborating Center for Drug Statistics Methodology (WHOCC), and was published in 1976.

ATC Codes are one of the fields of the DrugBank database and it can be used to build a taxonomy. We use this information in order to perform one of the evaluations our approach.

The system has a total of 5 levels, and the code consists of 7 alphanumerical characters, which can be read in the following way:

- First level: character 0 - for example 'A'
- Second level: characters 1-2, numbers - 02
- Third level: character 3 - for example 'C'
- Fourth level: character 4 - for example 'A'
- Fifth level: character 5-6 - for example 04

Although each level has its significance, we have decided to focus on the first one of the system, which determines the anatomical main group and consists of 14 categories (the same number of clusters we use), as shown in the Table 3.1:

3.1.5 Discussion

Our aim is to compute different and heterogeneous similarity measurements: text, taxonomy and molecular structure based similarities. This is the principal constrain when found when looking for data resources. For us, it was important to try to find one unique resource which could be used for all the experiments developed within the project. This mainly the reason why we chose DrugBank.

Of course, there are plenty of resources related to the medical domain we considered potentially useful: MeSH [Lipscomb, 2000], SnoMed [Donnelly, 2006], UMLS [Bodenreider, 2004] or ChEMBL [Gaulton et al., 2011]. However, none of them are so complete as DrugBank, actually, some of them are just vocabularies or just contain chemical information. DrugBank has enough information to perform the three experiments proposed in this project and it is accepted in the domain as the most complete resource in terms of medical drugs. For all those reasons, it is the database we use during the development of this project.

Code	Contents
A	Alimentary tract and metabolism
B	Blood and blood forming organs
C	Cardiovascular system
D	Dermatologicals
G	Genito-urinary system and sex hormones
H	Systemic hormonal preparations, excluding sex hormones and insulins
J	Antiinfectives for systemic use
L	Antineoplastic and immunomodulating agents
M	Musculo-skeletal system
N	Nervous system
P	Antiparasitic products, insecticides and repellents
R	Respiratory system
S	Sensory organs
V	Various

Table 3.1: First Level ATC-code Meaning

3 2

Text Based Similarity

Text similarity is the task of determining the degree of similarity between two texts. Texts length can vary from single words to paragraphs to complete novels or even books. In our case, the texts are a combination of different textual fields extracted from the DrugBank database. Single words constitute a special case of text similarity which is commonly referred to as the task of computing word similarity [Zesch and Gurevych, 2010] and is not the focus of this project.

The computation of text similarity is a very difficult task for machines. This is mainly due to the enormous variability in natural language, in which texts can be constructed using different lexical and syntactic constructions. Even so, computing text similarity has been for several years a fundamental means for many NLP tasks and applications. Nowadays, still a lot of works are devoted to this topic [Kenter and De Rijke, 2015, Kashyap et al., 2016, Ho et al., 2018].

Our aim is to find a measure of similarity (or dissimilarity) among the drugs found in DrugBank by means of text similarity. To this purpose, we propose to use several textual fields extracted from DrugBank and compute the similarity of them from one drug to another.

The number of drugs used in this experiment were 1,661. DrugBank has much more drugs, however, we have selected just the ones which contain some information in the textual fields we are interested in. In the cases in which the drug missed one of the following fields, it was discarded: description, indication, pharmacodynamics or ATC Code.

In the upcoming subsections, we explain the main parts of this approach as well as

the decision making we faced.

3.2.1 Data representation

As said before, we are going to use several textual fields to compute similarity between the drugs. In order to do so, the drugs were represented in a vector space model, which is an algebraic model for representing text documents and, thus, similarities can be computed in this space.

To obtain the vector space model representation of the drugs, the data fields: description, indication and pharmacodynamics –all expressed in natural language– were concatenated and, after removing stop words and transforming to lowercase, their term frequency-inverse document frequency (tf-idf) representation was computed. We have chosen those three textual fields because we think they contain information which can be relevant to discriminate the drugs from each other. We thought also about using the field *'name'* because, as said at some point in the introduction, the name of a drug can contain useful information about the drug in its prefixes and suffixes. However, we considered that adding just one more word to the concatenation will be meaningless, since the high dimension of our vectorial space. In future approaches, the use of the cited field could be interesting but just alone, so that the weight of the prefixes and suffixes is noted.

In information retrieval, tf-idf or TFIDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. In this case, each document used to compute the tf-idf is the concatenation of the textual fields of each drug, while the corpus is formed by all those documents as a whole.

Thus, the data is represented as the matrix $M \in \mathbb{R}^{n \times d}$, where n is the number of drugs and d the number of words in the whole corpus. In other words, the rows of the matrix are the samples while the columns correspond to features of each sample.

3.2.2 Sparseness as a problem

Usually, the number of terms within a corpus is large, this together with the fact that only few terms appear in a specific document give room to a sparse matrix. The high dimensionality and sparseness of the matrix M entail to a well-known phenomenon called *'curse of dimensionality'*. The curse of dimensionality refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces (often with hundreds or thousands of dimensions) that do not occur in low-dimensional settings such as the three-dimensional physical space of everyday experience. The expression was coined by Richard E. Bellman when considering problems in dynamic optimization [Bellman, 2015]. The issue with this phenomena is that the Euclidean distance becomes meaningless. For us, this fact is obviously a problem to be solved, since we want to use the data represented in the matrix M to compute similarity (which is basically the unit minus the distance).

3.2.3 Dimensional reduction as a solution

Reducing the dimension of the vector space model we have computed, is the solution proposed in this work. Specifically, we use the technique that in Information Retrieval is known as Latent Semantic Indexing (LSI) [Dumais et al., 1995], for us, Latent Semantic

Analysis (LSA) [Deerwester et al., 1990]. Latent semantic analysis (LSA) is a technique in NLP, in particular, distributional semantics, of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. LSA assumes that words that are close in meaning will occur in similar pieces of text (the distributional hypothesis). Singular Value Decomposition (SVD) [Golub and Reinsch, 1970] is used to rank the features (unique words in this case) from the more relevant (named 'singular values') to the less relevant.

LSA uses Singular Value Decomposition (SVD) to find the most discriminative components of our data vectors. The SVD of $M \in \mathbb{R}^{n \times d}$ is the factorization given by:

$$M = U \Sigma V^T \quad (3.1)$$

where U is an orthonormal $n \times r$ matrix, V^T is a $r \times d$ orthonormal matrix and Σ is a diagonal $r \times r$ matrix which elements are the ordered singular values σ_i , $i \in [1, \text{rank}(M)]$. Thus, r is the rank of the space. In the Figure 3.1 we see a visual intuition of this in which the Matrices M and Σ are represented as A and D respectively.

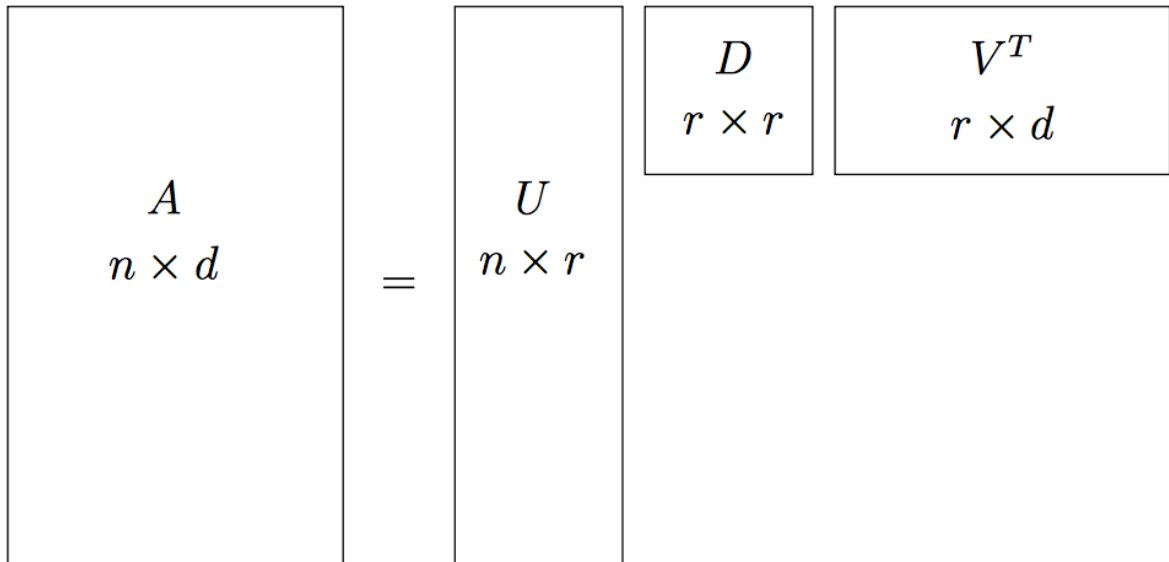


Figure 3.1: The SVD decomposition of an $n \times d$ matrix.

By taking the first k singular values σ_k , the expression 3.1 can be rewritten as 3.2:

$$M_{(k)} = U_{(k)} \Sigma_{(k)} V_{(k)}^T \quad (3.2)$$

where $U_{(k)}$ and $V_{(k)}^T$ are, respectively, the $n \times k$ and $k \times d$ matrices with orthonormal columns and $\Sigma_{(k)}$ is the $k \times k$ diagonal matrix with elements $\sigma_1; \dots; \sigma_k$. The matrix $M_{(k)}$ is then an approximation of M in a space of dimension k . With regard to LSA, $V_{(k)}^T$ can be thought as a matrix which maps terms to concepts and $U_{(k)}$ a matrix that maps concepts to documents. In this sense, the LSI is said to capture the semantic content of a text corpus. For the case in which we actually do not approximate (e.g. $k = r$), LSA reduces the dimensionality with no loss using a new basis for the semantic space. For $k < r$, some information is lost, so the result is an approximation.

In this project, the LSA was done with k equals to 500, 200 and 100. We have used all sets of data (reduced and original) in most of the upcoming steps, however, for the

clustering we used the similarity matrix obtained with the reduced data with k equals to 100. The reason is that for that value we obtained the most interesting results in the evaluation performed in Section 4.2.4.

As a result, we obtain a representation of our data in a three reduced dimensional spaces in return for losing part of the information.

3 2 4 Similarity measure

For this approach, the measurement we are using to compute the similarity is quite simple. The similarity matrix is computed based on the Euclidean distance over the dimensionally reduced data. Specifically, we first compute the Euclidean distance then we normalize it (to have a distance between zero and one) and finally, we calculate the similarity as the unit minus the normalized distance.

3 3

Taxonomy Based Similarity

Semantic similarity is a metric defined over a set of documents or terms, where the idea of distance between them is based on the likeness of their meaning or semantic content as opposed to similarity which can be estimated regarding their syntactical representation (e.g. their string format). Semantic relatedness includes any relation between two terms, while semantic similarity only includes 'IS-A' relations. The semantics of our taxonomy has only one sort of relationship: 'is-a' relationship, (e.g. Acetaminophen is-a SubClass of Benzenoids, or which is the same, Benzenoids is-a SuperClass of Acetaminophen).

Computationally, semantic similarity can be estimated by defining a topological similarity, by using ontologies to define the distance between terms/concepts (as we propose in the present section). For example, a naive metric for the comparison of concepts ordered in a partially ordered set and represented as nodes of a directed acyclic graph (e.g., a taxonomy), would be the shortest-path linking the two concept nodes. Based on text analyses, semantic relatedness between units of language (e.g., words, sentences) can also be estimated using statistical means such as a vector space model to correlate words and textual contexts from a suitable text corpus (as we do in Section 3.2).

In our approach, similarity between drugs is computed using the graph structure in which they are organized, that is, topological similarity. There are essentially two types of approaches that calculate topological similarity between ontological concepts:

- **Edge-based** [Cheng et al., 2004, Pekar and Staab, 2002, Del Pozo et al., 2008], also named path-based, which uses the edges and their types as the data source. This is the type we are using.
- **Node-based** [Resnik, 1995, Lin et al., 1998] in which the main data sources are the nodes and their properties.

3 3 1 Data representation

DrugBank organizes the data in some taxonomic structures, but we have used the classification tag to construct **2 trees of 6 levels** which would connect the drugs in

the database through undirected edges. So we do not have a graph but two trees. Two different cases are contemplated: unweighted and weighted graphs. On the one hand, in unweighted graphs all the edges have the same meaning and value. On the other hand, in weighted graphs the cost of moving from one node to another is different depending on the level of the taxonomy in which the nodes are. This is to say, the edges between levels of the taxonomy imply a higher cost than edges between the same level. The distance between drugs is calculated as a shortest path distance. For the case of the weighted graph, the higher the level of the closest common ancestor in the tree, the higher the weight for the distance.

The motivation behind having two trees instead of a unique graph is because the drugs belong to either Organic or Inorganic kingdom, so we have not contemplated the most general class 'Drug'. Thus, we have decided that the path between those kingdoms should not exist, because of the very nature of the taxonomy (no or very little information gain). Additionally, introducing full connectivity (any drug can be reached from any drug in the database), by adding a common root, drastically increases computation time. From now on, even though our graph is actually formed by two trees, we are going to refer to them as 'graph' just to simplify the writing.

Shape of our graph

We have built the graphs with six levels: the five fields from the Classification tag of DrugBank (see Section 3.1.3) and the DrugBank ID of the drugs. In the Figure 3.2 we can see an extract of the final graph. As we see, there are five levels which come from the Classification tag of DrugBank, finally, at the bottom, the DrugBank ID. In the Figure we can also see in a different tone of gray the Direct Parent of the drug 'DB00316'. The reason is because it is completely equal to the Sub-class so in our graph, for cases like that one, we omitted the Direct Parent (in order to have less nodes). Meaning, that node actually does not exist in our graph, the edge goes directly from the subclass to the drug.

Data of the graph

In this section, we analyze some properties that our graph has. First, we introduce the properties we have considered and finally, we show a table with the value of those properties.

- **Number of drugs.** Number of drugs used to build the graph.
- **Number of nodes.** Number of nodes of a graph.
- **Number of edges.** Number of edges (connectors between nodes) of a graph.
- **Average degree.** The node degree is the number of edges adjacent to that node, the average is the mean for all nodes.
- **Directed.** If the edges of a graph have direction, then the graph is directed.
- **Radius.** The radius is the minimum eccentricity⁵.
- **Center.** The center is the set of nodes with eccentricity equal to radius.

⁵The eccentricity of a node v is the maximum distance from v to all other nodes in the graph.

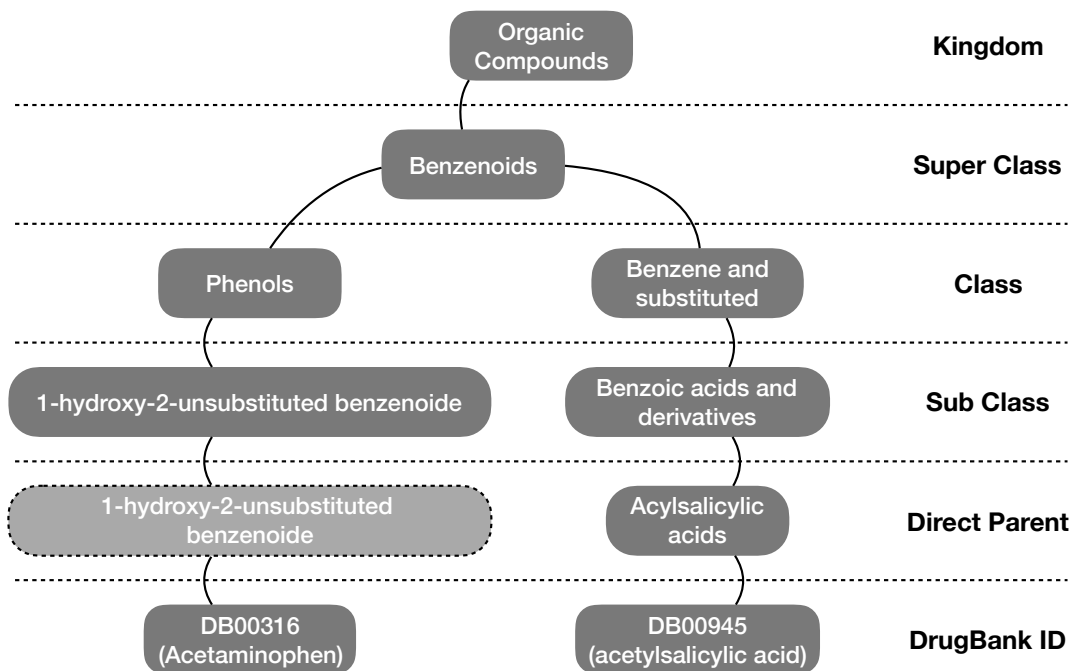


Figure 3.2: Example of a subgraph of the total graph built in our project. The drugs used for this example are: Acetaminophen and Acetylsalicylic acid

Code	Contents
Number of drugs	1,661
Number of nodes	2,360
Number of edges	2,452
Average degree	2.0780
Directed	No
Radius	7
Center	'Organic acids and derivatives'
Density	0.000880867359768934
Depth	5

Table 3.2: Graph information for both cases: unweighted and weighted.

- **Density.** The density is 0 for a graph without edges and 1 for a complete graph. The density of multigraphs can be higher than 1.
- **Depth.** The different levels of a graph, in our case, the graph is formed by two trees, each of them with five levels (Classification field from DrugBank).

Please, note that the number of drugs we have used to build the graphs is 1,661, as we did for the measure of text based similarity. Although we could build a graph now with all the drugs, we just want to have comparable results of all our experiments. In the table 3.2 we show the values the properties of our graphs (weighted and unweighted).

3.3.2 Similarity measure

Once we have built a graph, it is necessary to use it to compute a similarity among our nodes, specifically, we use a distance measure. In the mathematical field of graph theory, the distance between two vertices in a graph is the number of edges in a shortest path (also called a graph geodesic) connecting them. This is also known as the geodesic distance. Notice that there may be more than one shortest path between two vertices. If there is no path connecting the two vertices, i.e., if they belong to different connected components, then conventionally the distance is defined as infinite (in our case, it is set to -1).

In the case of a directed graph the distance $d(u, v)$ between two vertices u and v is defined as the length of a shortest directed path from u to v consisting of arcs, provided at least one such path exists. Notice that, in contrast with the case of undirected graphs, $d(u, v)$ does not necessarily coincide with $d(v, u)$, and it might be the case that one is defined while the other is not. In our case, we work with an undirected graph.

We work with two different sorts of graphs: weighted and unweighted. On the one hand, a weighted graph uses differentiates the cost of moving from one node to another giving different weights to the edges. The closer we are to the root node, the bigger is the weight of the edges.

Unweighted Graph

This is the simplest example of graph in which each edge has the same value (weight) when computing the shortest path. So basically, in order to find the shortest path, we just focus on the length of it.

Using the small taxonomy showed before, we have created another image in which we show a possible path between two drugs: *Acetaminophen (paracetamol)* and *Acetylsalicylic Acid (aspirin)* (see Figure 3.3). The path is highlighted in green (see edges) and in blue circles we have written the weight of every edge, in this case, it is always one.

The length of the path (considering all the edges and their weights) is 7. Please, note that the node corresponding to the 'Direct Parent' of the *Acetaminophen* is not included in our graph (since it is equal to the Subclass).

Weighted Graph

In this case, we work with weighted graphs which means that now not only the length of the path but also the weight of the edges are important to compute the shortest path.

Using the small taxonomy showed before, we have created another image in which we show a possible path between two drugs: *Acetaminophen (paracetamol)* and *Acetylsalicylic Acid (aspirin)* (see Figure 3.3). The path is highlighted in green (see edges) and in blue circles we have written the weight of every edge. The weight is a parameter of the function we have implemented to build the graph. However, we have set a default value for the weights: 1, 10, 20, 25 and 50. We research about what values to use and since it is not crucial, we just made the weights increase the deeper is the edge. A good alternative would be to use the inverse of the depth as the weight for an edge.

The length of the path (considering all the edges and their weights) is 111. Please, note that the node corresponding to the 'Direct Parent' of the *Acetaminophen* is not included in our graph (since it is equal to the Subclass).

Distance to Similarity

We are interested in measuring the similarity between drugs, however, with the shortest path computation we obtain a distance. Thus, it is necessary to translate that distance into a similarity, which is not trivial this time. In the text based similarity, we could obtain the distance from the similarity and vice-versa just calculating the unit minus one of them. Nevertheless, in this case we need to use other kind of transformation.

There exist several ways of turning those path distances into similarities though, we have chosen the method proposed by Leacock and Chodorow [Leacock and Chodorow, 1998]. The Leacock and Chodorow Similarity between two nodes of a graph (drugs, in this case, $d1$ and $d2$) is as follows:

$$Sim(d1, d2) = -\log\left(\frac{length}{2D}\right)$$

Where length is the length of the shortest path between the two concepts (using node-counting) and D is the maximum depth of the taxonomy. Based on this measure, the shortest path between two concepts of the ontology restricted to taxonomic links is normalized by introducing a division by the double of the maximum hierarchy depth.

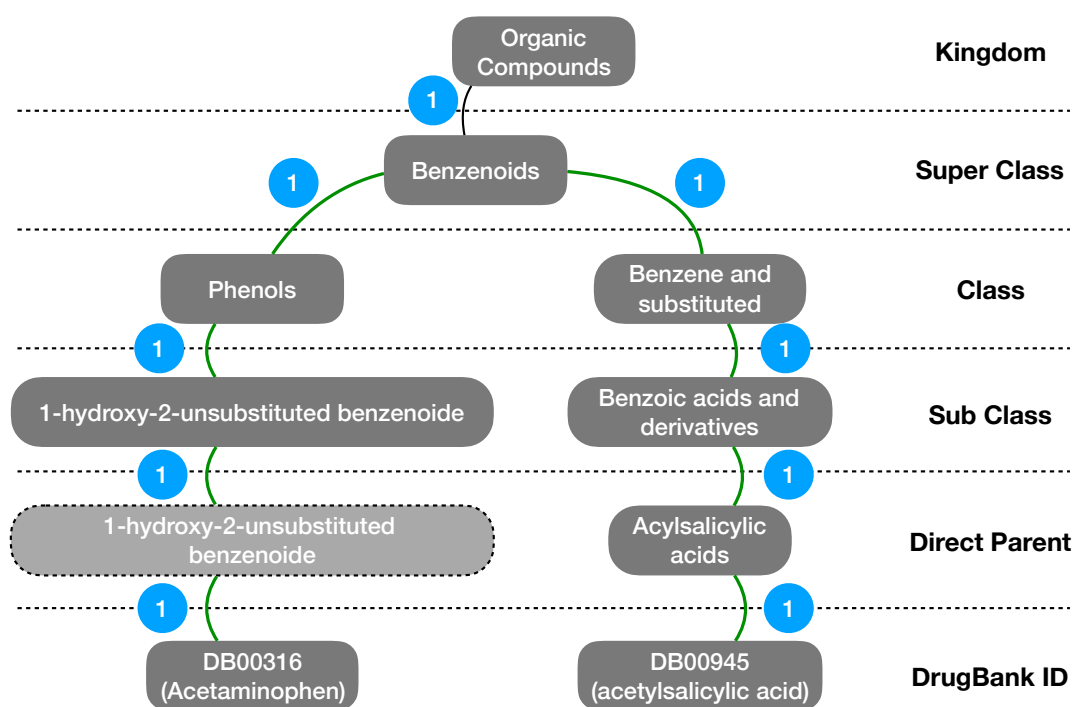


Figure 3.3: Example of a subgraph of the total graph built in our project in which an unweighted distance path is computed between two drugs. The drugs used for this example are: Acetaminophen and Acetylsalicylic acid

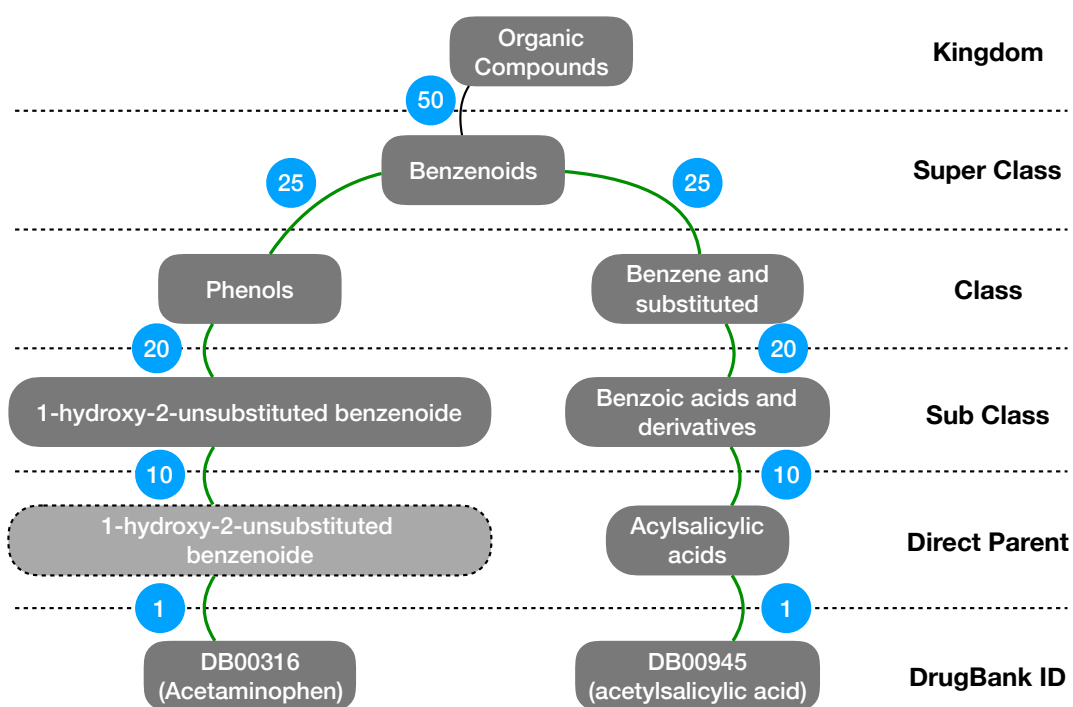


Figure 3.4: Example of a subgraph of the total graph built in our project in which a weighted distance path is computed between two drugs. The drugs used for this example are: Acetaminophen and Acetylsalicylic acid

3 4**Molecular Structure Based Similarity**

Measurements of structural similarity play an important role in chemoinformatics for applications such as similarity searching, database clustering and molecular diversity analysis. Thus, computing similarity among chemical structures is a current trend in the domain and it is used for tasks such as Drug-Drug Interaction Prediction [Liu and Zhao, 2016, Vilar and Hripesak, 2016, Wang et al., 2016, Takeda et al., 2017].

The importance of structural similarity derives in large part from the Similar Property Principle, which states that molecules that are structurally similar are likely to have similar properties [Johnson and Maggiora, 1990]. Actually, most of the drug/chemical compounds databases use the molecular structure for different applications. For instance, DrugBank, PubChem [Bolton et al., 2008] and ChEMBL [Gaulton et al., 2011], have a search engine in which, if we have the molecular structure of a compound, we can find other similar ones. Another example is STICH [Kuhn et al., 2007] (Search Tool for Interactions of Chemicals), a database which uses molecular structure similarities to predict relations between chemicals.

The main three elements of any similarity measure based on Molecular Structure are:

- **Representation or Descriptor.** It is used to characterize the two molecules that are being compared. Among all the possible descriptors we use the 2D fingerprints⁶, for more detail, see Section 3.4.2.
- **Weighting Scheme.** It is used to reflect the relative importance of different parts of the representation. No weights are used in this project.
- **Similarity Coefficient.** It is used to quantify the degree of resemblance between two appropriately weighted structural representations. In our case, we use the Tanimoto Coefficient (see Section 3.4.3).

Typical 2D and 3D representations of a drug can be seen in Figures 3.5 and 3.6, where we show the molecular structure of the *Acetaminophen* drug.

3 4 1 Data Format

Drugbank database contains a lot of data which is rather heterogeneous, however, it is possible to download a complete database, written in XML, which includes all the fields for the covered drugs. For other experiments we have done in this project, it was better to use that generic file. Nevertheless, in this concrete experiment, we are using a specific file which is devoted to work with the molecular structure information of the drugs.

In every Drugbank release there are several different documents to download (see Section 3.1.2). Thus, we can ensure that the drugs within the different documents are the same. This is just to clarify that we are using the same information than

⁶A fingerprint is a vector, each element of which describes the presence of one or more substructures in a molecule, with typical fingerprints containing a few hundred or a few thousand elements, and with two molecules being considered to be similar if their fingerprints share common values for many of the constituent elements.

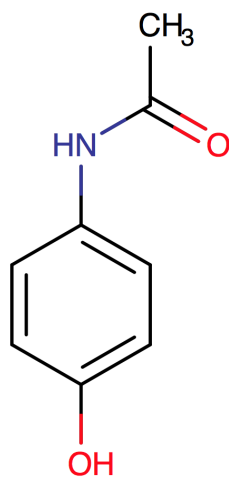


Figure 3.5: Molecular Structure in 2D of the Acetaminophen drug.

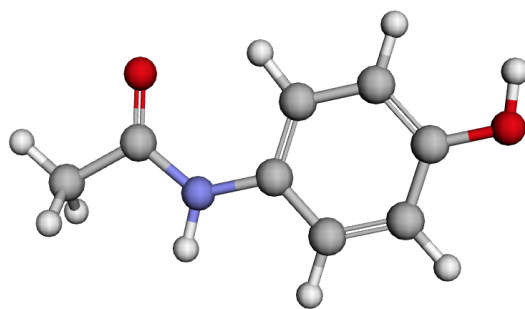


Figure 3.6: Molecular Structure in 3D of the Acetaminophen drug.

in the rest of performed experiments. Readers and users of this document could get confused about this, since in the present experiment we use more drugs than in others. The reason is that here we have used all the drugs which have a complete molecular structure, 8,738 from a total of 10,562 included in the used release.

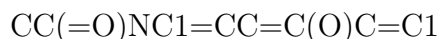
The format of the file which contains the specific information related to Molecular Structures is SDF, which is one of a family of chemical-data file formats developed by MDL⁷; it is intended especially for structural information. "SDF" stands for structure-data file, and SDF files actually wrap the molfile (MDL Molfile) format. A feature of the SDF format is its ability to include associated data (not only molecular, but also general purpose data like 'drug id'). The reason why we decided to use this document was because we discovered there was a specific library for Python to deal with this sort of document. That tool (RDKit) was perfectly aligned to the work we wanted to do since it includes several useful functions. Some of them to generate fingerprints automatically (from a SDF file) and to compute typical similarity coefficients commonly used in this domain.

RDKit

RDKit [Landrum et al., 2006] is an Open source toolkit for cheminformatics and Machine Learning which contains a lot of methods and functionalities to deal with Molecular Structure data from chemical compounds (including similarity measures). It is written in Python (2 and 3) and has been used in several relevant open source projects (ChEMBL Beaker [Nowotka et al., 2014], myChEMBL [Ochoa et al., 2013], etc.).

3.4.2 Data representation

An important issue is how to actually represent the Molecular Structure of a chemical compound so that a computer can process it efficiently. Normally, the Molecular Structure is represented by well-known methods like: InChi Key [McNaught, 2006] or SMILES [Weininger, 1988]. For instance, the SMILES representation for the *Acetaminophen*, molecular structure showed in Figures 3.5 and 3.6, is:



Even though both, SMILES and InChi Key, are included within the fields of DrugBank, we cannot use those sorts of representation to compute similarity between drugs, a more efficient representation is needed. In this work, we have used 2D fingerprints, a list of binary values (0 or 1) which characterize a molecule.

A complete analysis of similarity measures based on molecular structure is provided in [Nikolova and Jaworska, 2003] and [Willett, 2014], where different ways of representation are studied. In both reviews, is claimed that, even though 2D Fingerprints compress the molecular information (losing part of it) are preferable to other complicated representations. While the improvement provided by those other complex representations is not significant, the efficiency and simplicity of 2D Fingerprints is noteworthy. Indeed, 2D Fingerprints are the state of the art in this domain [Yu et al., 2015, Cereto-Massagué et al., 2015, Muegge and Mukherjee, 2016].

⁷MDL Information Systems, Inc. was a provider of R&D informatics products for the life sciences and chemicals industries. The company was launched as a computer-aided drug design firm (originally named Molecular Design Limited, Inc.) in January 1978 in Hayward, California.

Fingerprints are lists of binary values which characterize a molecule. Obviously, the more bits we use, the more precise the representation is. In this project, we have explored two of the most well-known types: MACCS [Keys, 2011] and ECFPs [Rogers and Hahn, 2010].

MACCS

The MACCS keys are a set of questions about a chemical structure. Here are some of the questions: Are there fewer than 3 oxygens? Is there a S-S bond? Is there a ring of size 4? Is at least one F, Cl, Br, or I present?

The result of this is a list of binary values – either true (1) or false (0). This list of values for a given chemical structure is called the MACCS key fingerprint for that structure.

ECFP

Extended-Connectivity Fingerprints (ECFPs) are circular topological fingerprints designed for molecular characterization, similarity searching, and structure-activity modeling. They are among the most popular similarity search tools in drug discovery and they are effectively used in a wide variety of applications.

The length for the fingerprints we are using in this project are: 167 bits for MACCS and 1,024 bits for ECFP. This fact has an effect on the result of the computed similarities. The values of similarities when using MACCS are higher in general. ECFP has more precision (more bits) but that leads us to smaller values of similarity, obviously, the more bits we have, the more difficult is for the pairs of fingerprints to be similar. We have studied the correlation between the similarities computed using MACCS and ECFPs in order to see if we can just choose one of them. However, the value of Pearson Correlation is around 0.6, which is not enough to say that they are really correlated, so we use both for the evaluation.

3 4 3 Similarity measure/coefficient

The selection of a similarity coefficient is made a condition of the sort of chosen representation, in this case, 2D fingerprints. The most well-known coefficient used with fingerprints is the Tanimoto Coefficient (also known as Jaccard Index). The computation of the Tanimoto Coefficient for two binary vectors (a and b) of length k is defined as:

$$\frac{\sum_{j=1}^k a_j \times b_j}{\sum_{j=1}^k a_j^2 + \sum_{j=1}^k b_j^2 - \sum_{j=1}^k a_j \times b_j} \quad (3.3)$$

However, there are other coefficients to be used in this case. One instance is the Dice Coefficient, which is actually, quite similar to Tanimoto. The computation of the Dice Coefficient for two binary vectors (a and b) of length k is defined as:

$$\frac{2 \sum_{j=1}^k a_j \times b_j}{\sum_{j=1}^k a_j^2 + \sum_{j=1}^k b_j^2} \quad (3.4)$$

As we see, both expressions are rather similar. Even so, we wanted to prove that relation which exists between them, so that we studied the correlation between the

obtained similarity matrices in both cases. We used the Person's Correlation Coefficient (see 3.5.2 and the value was around 0.98. Person's Correlation Coefficient denotes total positive correlation when the value is 1, so we can say that both coefficients (Tanimoto and Dice) are totally correlated. Note that this study was done for the two sorts of fingerprints we use (MACCS and ECFP) and it was similar in both cases.

A whole study of the convenience of using the Tanimoto Coefficient is provided in [Bajusz et al., 2015]. The conclusion of that study claims that the Tanimoto index, Dice index, Cosine coefficient and Soergel distance were identified to be the best (and in some sense equivalent) metrics for similarity calculations. The similarity metrics derived from Euclidean and Manhattan distances are not recommended on their own, although their variability and diversity from other similarity metrics might be advantageous in certain cases (e.g. for data fusion).

As said before, in our approach, two different coefficients were used: Tanimoto (Jaccard) and Dice. The correlation between the similarity values that both provided was studied using the Pearson Correlation. The value of correlation was about 0.99 for both sorts of fingerprints: MACCS and ECFP. For this reason and for the evidences about the benefits of using Tanimoto Coefficient proved in other works, we decided just to use it to cluster and evaluate.

3 5

Evaluation Setup

There are two main sorts of evaluation: direct and indirect. On the one hand, a direct evaluation is the one performed directly over the result you want to study. On the other hand, an indirect evaluation is the one in which you use the obtained result to solve a task and then you evaluate the performance of it over the task. Normally, the ideal evaluation is a direct one, in which the result is compared with a '*golden standard*'. However, it is difficult to evaluate our work since there is not any clear '*golden standard*' to compare our results with.

In this project, we have performed to different evaluations over the computed similarities:

- **Clustering.** This is an example of indirect evaluation. We have used the similarities to cluster the drugs into groups. Then, we study the ATC Code distribution of those clusters in order to check if our similarity measurements are good.
- **Ground Truth.** This evaluation is a small direct evaluation we have done with a ground truth annotated by experts in the domain. The similarity of a list of 100 pairs of drugs were annotated by experts. We have taken it from [Franco et al., 2014] and modified and adapted to our convenience. We compare the similarities computed within this project with the similarity following the experts's opinion. In Section 3.5.2, we talk more about this.

3 5 1 Clustering

As said before within the present section, the similarity measurements we have implemented are used to cluster the used drugs. This is meant to have an evaluation method to measure the quality of the computed similarities.

Method name	Parameters	Scalability	Usecase	Geometry (metric used)
K-Means	number of clusters	Very large n_{samples} , medium n_{clusters} with MiniBatch code	General-purpose, even cluster size, flat geometry, not too many clusters	Distances between points
Affinity propagation	damping, sample preference	Not scalable with n_{samples}	Many clusters, uneven cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Mean-shift	bandwidth	Not scalable with n_{samples}	Many clusters, uneven cluster size, non-flat geometry	Distances between points
Spectral clustering	number of clusters	Medium n_{samples} , small n_{clusters}	Few clusters, even cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Ward hierarchical clustering	number of clusters	Large n_{samples} and n_{clusters}	Many clusters, possibly connectivity constraints	Distances between points
Agglomerative clustering	number of clusters, linkage type, distance	Large n_{samples} and n_{clusters}	Many clusters, possibly connectivity constraints, non Euclidean distances	Any pairwise distance
DBSCAN	neighborhood size	Very large n_{samples} , medium n_{clusters}	Non-flat geometry, uneven cluster sizes	Distances between nearest points
Gaussian mixtures	many	Not scalable	Flat geometry, good for density estimation	Mahalanobis distances to centers
Birch	branching factor, threshold, optional global clusterer.	Large n_{clusters} and n_{samples}	Large dataset, outlier removal, data reduction.	Euclidean distance between points

Figure 3.7: A comparison of the clustering algorithms in scikit-learn library (Python)

In a nutshell, our approach is to cluster the drugs in 14 groups. The number of clusters is chosen from the ATC Code classification, specifically, the first level, which classifies the drugs into fourteen different groups. During this section we explain why we decided to use that number. Then, we study the ATC first level distribution of all the drugs within the clusters, in order to see if actually we are grouping drugs which belong to the same ATC First Level group. We evaluate this using histograms.

In this section we talk about all the different aspects of this evaluation: decision making about the type of clustering and number of clusters.

Clustering Technique

An interesting library of Python named *scikit-learn* has several ready to use techniques to do clustering. We studied the properties of the clustering algorithms included in that library and we thought which one would fit better with our problem. In the Figure 3.7 we show a table from the main page of *scikit-learn* in which all possible techniques are compared.

The type of clustering we are using is Spectral Clustering, the reasons of our decision are the followings:

1. We want to evaluate if our similarities group correctly the drugs by their ATC Codes. ATC Codes are a classification of drugs which can be think as a graph. *Scikit-learn* offers two different techniques in which the geometry is based on a graph distance: Affinity Propagation and Spectral Clustering. We think that one of those two options would be good for our problem.
2. Affinity Propagation is normally used for cases in which the number of clusters is large while Spectral Clustering is used for small number of clusters (see column 'Usecase' in Figure 3.7). In addition, the only needed parameter for Spectral

Clustering is the number of clusters, while the needed parameters for Affinity Propagation are two: damping and sample preference. In our case, it is easier for us to choose the number of clusters and we do not want to have many clusters (because then, the evaluation using the histograms would be more difficult and meaningless).

Spectral clustering techniques make use of the spectrum (eigenvalues) of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions. The similarity matrix is provided as an input and consists of a quantitative assessment of the relative similarity of each pair of points in the dataset.

Spectral Clustering needs as input argument the number of clusters, therefore, we need to choose that number.

Number of Clusters

The Anatomical Therapeutic Chemical (ATC) Classification System is used for the classification of active ingredients of drugs according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties. It is controlled by the World Health Organization Collaborating Center for Drug Statistics Methodology (WHOCC), and was published in 1976.

The system has a total of 5 levels, and the code consists of 7 alphanumerical characters. The more characters we use, the more specific is the ATC code, until we reach the final pharmacological compound. The five levels are as follows:

- **First level.** The first level of the code indicates the anatomical main group and consists of one letter (character 0). For instance: **C** Cardiovascular System.
- **Second level.** The second level of the code indicates the therapeutic subgroup and consists of two digits (characters 1-2). For instance: **C03** Diuretics.
- **Third level.** The third level of the code indicates the therapeutic/pharmacological subgroup and consists of one letter (character 3). For instance: **C03C** High-ceiling diuretics.
- **Fourth level.** The fourth level of the code indicates the chemical/therapeutic/pharmacological subgroup and consists of one letter (character 4). For instance: **C03CA** Sulfonamides.
- **Fifth level.** The fifth level of the code indicates the chemical substance and consists of two digits (character 5-6). For instance: **C03CA01** Furosemide.

Although each level has its significance (as shown before), we have decided to focus on the first one of the system, which determines the anatomical main group and consists of 14 categories, as shown in the Table 3.3. This first level is the most relevant one, since it indicates the anatomical parts of the body in which the drug could act.

Clustering Evaluation Measurement

Along the Section 2.3, some clustering evaluation techniques are formally explained. In this project, we can evaluate using an external set of clusters: the real ATC Code distribution of the drugs. Specifically, we use *Purity*, which is a measure of the extent

Code	Contents
A	Alimentary tract and metabolism
B	Blood and blood forming organs
C	Cardiovascular system
D	Dermatologicals
G	Genito-urinary system and sex hormones
H	Systemic hormonal preparations, excluding sex hormones and insulins
J	Antiinfectives for systemic use
L	Antineoplastic and immunomodulating agents
M	Musculo-skeletal system
N	Nervous system
P	Antiparasitic products, insecticides and repellents
R	Respiratory system
S	Sensory organs
V	Various

Table 3.3: First Level ATC-code Meaning

to which clusters contain a single class. In our case, the correct classified classes are each of the fourteen possible ATC Codes.

Usually, the calculation of *Purity* can be thought of as follows: For each cluster, count the number of data points from the most common class in said cluster. Now take the sum over all clusters and divide by the total number of data points. This is a quantitative evaluation measurement, however, in this work we have studied the *Purity* in a more qualitatively way. Specifically, we have obtained the histograms of the ATC Codes distribution in each cluster and we have analyzed them.

The reason why we have not computed the *Purity* directly, is because the result of the clustering is not really good. Thus, we have preferred just to perform the evaluation with a visual approach more than with the exact value of *Purity*. Actually, *Purity* evaluation is not perfect, since high purity is easy to achieve when the number of clusters is large - in particular, purity would be 1 (maximum value) if each document (ATC Code) gets its own cluster. Thus, we cannot use purity to trade off the quality of the clustering against the number of clusters. Even so, the principle we follow, it is the same.

3.5.2 Ground Truth

The external (direct) evaluation consisted of comparing the computed similarities values with the similarity between 100 pairs of drugs which were annotated by experts. That annotated data has been taken from [Franco et al., 2014] and modified and adapted to our convenience. In the Figure 3.8, we can see how the original file provided in [Franco et al., 2014] looks like.

Specifically, the ground truth was built using the opinion of 143 experts, who provided Yes/No decisions on set of 100 DrugBank 3.0 [Knox et al., 2010] molecule-

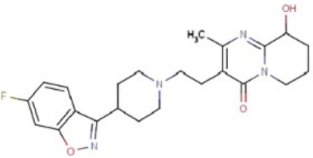
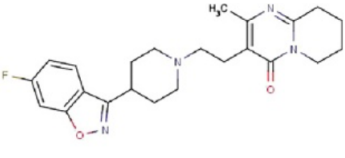
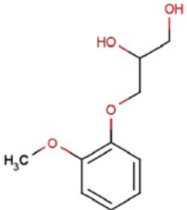
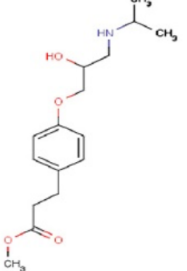
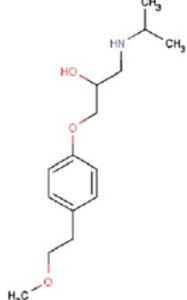
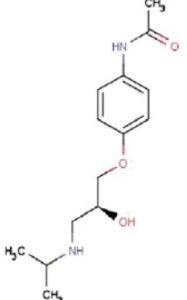
Molecule A	Molecule B	Yes	No	Similarity
		0.93	0.07	0.865
		0.14	0.86	0.432
		0.59	0.41	0.595

Figure 3.8: Three molecule-pairs with the corresponding fractions of YesNo responses to the question: 'Are these molecules similar?' The similarity values in the right-hand column were computed by the authors using the Tanimoto coefficient and ECFP4 fingerprints.

DB00281	DB00296	0.4688
DB01267	DB00734	0.9375
DB00874	DB00187	0.1406
DB00706	DB00499	0.0313
DB01576	DB00191	0.8828
DB00867	DB00884	0.0078
DB00249	DB00432	0.9055
DB01522	DB04552	0.0156
DB00264	DB01297	0.5669
DB00353	DB01253	0.9844

Figure 3.9: Extract of the CSV file used to evaluate our similarities against the proposed ground truth.

pairs. Basically, all those experts were asked to answer with Yes/No to the question: “Are these molecules similar?”. The answers were collected and a distribution of Yes/No answers was computed. In this project, we use the proportion (percentage) of ‘Yes’ answers as degree of similarity. Of course, the reader should note that the experts were not asked about the degree of similarity among the drugs but if the molecules were similar or not.

Even though the original file contains 100 pairs of drugs, we have been able to use just 97 pairs. We have not been able to find some of the names of the drugs. In the original file provided by the authors of the cited paper, we find the fields showed in Figure 3.8 and the SMILES representation of the molecules. Making use of different search tools and taking the 2D structure and the SMILES representation of the drugs, we were able to find the DrugBank IDs of just 97 pairs of drugs. That is the reason why our final ground truth is shorter than the original.

We built a new file, a .csv file, which contains, on the one hand, three columns (the two DrugBank IDs and the similarity), and on the other hand, 97 rows (all the pairs). In the Figure 3.9 we show ten of the rows of that file (which can be found in our GitHub repository).

In order to evaluate how our similarity measures are related to the ground truth values, we have studied three different aspects:

- Correlation between the ground truth and our measurements considering the values of the similarities.
- Correlation between the ground truth and our measurements considering the order inferred by the similarity values (from the most different pair to the most similar one).
- Classification of the drugs as similar or non-similar using a threshold.

Note that all those evaluations have been performed individually for each of the three similarities computed in this project: text, taxonomy and molecular structure based similarity.

Of course, not all the pairs of drugs which exist in the ground truth are used in our experiments. On the one hand, for the cases of text and taxonomy based similarity,

we use just the drugs which contain non-empty data in the fields we are interested in. This, reduces a lot the number of drugs we use. On the other hand, in the molecular structure similarity measurement, we work with more drugs, so it is more likely to find all the pairs of the ground truth.

Value

Each pair of drugs has associated a similarity value. In the case of the ground truth, that value is the percentage of experts who said that the pair of molecules is similar. In the case of our measurements, is the similarity value computed with each of our three different approaches.

We study the correlation between the value of the ground truth and our measurements (individually). For that, we have used Pearson's Correlation Coefficient.

Pearson's Correlation Coefficient has a value between +1 and -1, where 1 means total positive correlation, 0 is no linear correlation and -1 is total negative correlation.

Pearson's correlation coefficient when applied to a sample (as it is in this case) is commonly represented by the letter r and may be referred to as the sample correlation coefficient or the sample Pearson correlation coefficient.

Let x_i and y_i with $i \in [1, n]$ be a set of observations of the variables X and Y . We can obtain r from the next formula:

$$r = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sqrt{(\sum x_i^2 - n \bar{x}^2)} \sqrt{(\sum y_i^2 - n \bar{y}^2)}} \quad (3.5)$$

where \bar{x} and \bar{y} are the sample mean.

For us, X and Y are two lists which contain the similarity values for each of the drug pairs. One of them is for the ground truth, the another one is for one of our similarity measurements.

Order

We have ordered the pairs by the value of their similarity, from the most different to the most similar, in both cases (our similarity measurement and the ground truth). Our aim is to study how each of our similarities and the ground truth are correlated, in this case, we study the rank inferred using the similarity values. We use one of the most well-known rank correlation methods: *Kendall's Tau*.

Kendall's τ has a value between +1 and -1, where 1 means total positive correlation, 0 is no linear correlation and -1 is total negative correlation.

Let x_i and y_i with $i \in [1, n]$ be a set of observations of the variables X and Y . Any pair of observations (x_i, y_i) and (x_j, y_j) , where $i \neq j$, are said to be *concordant* if the ranks for both elements (more precisely, the sort order by x and by y) agree. That is, if both $x_i > x_j$ and $y_i > y_j$; or if both $x_i < x_j$ and $y_i < y_j$. They are said to be *discordant*, if $x_i > x_j$ and $y_i < y_j$; or if $x_i < x_j$ and $y_i > y_j$. If $x_i = x_j$ or $y_i = y_j$, the pair is neither concordant nor discordant.

The Kendall τ coefficient is defined as:

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n(n-1)/2} \quad (3.6)$$

For us, X and Y are two lists which contain the pairs of drugs ranked by their similarity values. One of them is for the ground truth, the another one is for one of our similarity measurements.

Threshold

We have selected a threshold to classify the pairs of drugs into two different categories: similar and non-similar. If the similarity value is greater than the threshold, then, the drugs are similar. The threshold we have chosen is 0.85. The reason is because one of our similarity measures, the Tanimoto Coefficient (see Section 3.4.3), is considered relevant from that value. Then, we compute the precision and the recall of the classification process.

EXPERIMENTS AND ANALYSIS

Three different experiments have been developed within this work. Each of them is devoted to evaluate the different similarity measures we have computed. In this chapter we talk about those experiments and the obtained results for the evaluation of the measurements. In particular, we have divided the chapter into four different sections:

- General Experimental Setup
- Text Based Similarity
- Taxonomy Based Similarity
- Molecular Structure Based Similarity

As a reminder, the implementation of this similarities, can be found on a free access repository on GitHub created by the author of this thesis¹. In there, there is a folder named *'notebooks'* in which the three experiments appear.

4 1

General Experimental Setup

There are some general aspects which are shared among the three experiments. This section is dedicated to set the general framework in which the experiments have been done.

As said in previous chapters, we are using the previous most updated release of DrugBank 5.0. [Wishart et al., 2017]. Please, note that the used version is not the latest one. There exists a new version which was published at the beginning of April, obviously, we had not time to use it and change all our analysis of the results. However, the changes are minimal, instead of 11,002 drugs now there are 11,037.

Specifically, we use two different files: the complete database (a .xml file) and the molecular structure information (a .sdf file). The first one is used in Sections 4.2 and 4.3 while the second one is used in Section 4.4. The previously cited release, is said to

¹<https://github.com/albertoOA/Medical-Entities-Similarity-Measurements>

contain a total of 11,002 drugs, however, we have been able to read just 10,562 drugs from the .xml file, because, actually, there are just that number of drugs².

One of the evaluations we do is based on clustering the drugs using our implemented similarity measures (see Section 3.5.1). Considering that DrugBank is not totally complete, some drugs do not contain all the fields, we decided to see what number of drugs had ATC Code in DrugBank. From the total number of drugs, just 2,287 has a non-empty ATC Code. Thus, we are going to be limited by that number in our evaluation. Note that each drug can have more than one ATC Code, so that the number of ATC Codes is not 2,287 but 3,876. In the Figure 4.1, we can see a distribution of the First Level of all available ATC Codes. ATC Code has five levels, the reason why we are just interested in the first one is because we use the number of classes of that level (fourteen) as number of clusters. The classes are alphabetical characters. Please, for further information about ATC Code, see Section 3.1.4.

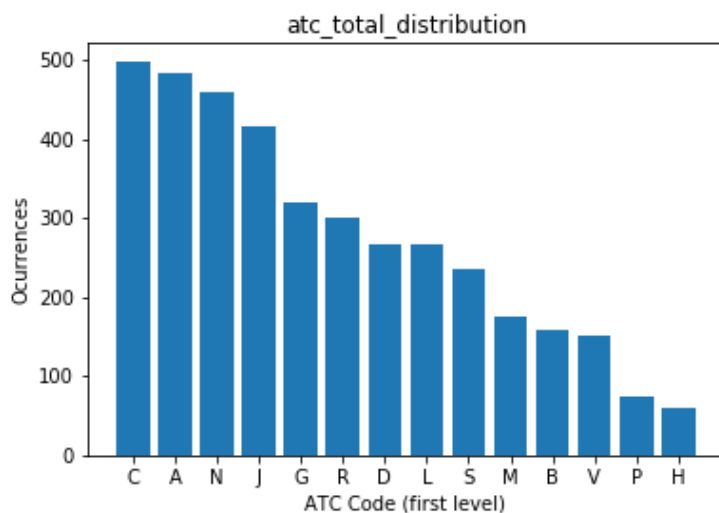


Figure 4.1: Distribution of the first level of all ATC Codes contained in DrugBank.

As we see in the the Figure 4.1, the distribution of the fourteen groups is not balanced. This will affect the performance of the clustering, since the classes with less samples are more difficult to be discriminated from the rest.

4 2

Text Based Similarity

In this section, we explain the experimental setup and the results of the experiment for the text based similarity. Note that in the upcoming paragraphs we just talk about the experiment itself, avoiding some details. For more detailed description about how the similarity measure is computed, please, see Section 3.2.

²It is strange, but we have downloaded several times the cited version and checked that the number of times that the xml field *'drug'* appears into the .xml file is 10,562. So that it is not a problem of our accessor.

4.2.1 Experimental Setup

Our aim is to measure similarity between drugs, in particular, using the analysis of textual information about those drugs. The textual information of each drug is gathered from the complete DrugBank database file (.xml format). The textual fields for each drug we use are: *description*, *indication*, *pharmacodynamics*. Note that other fields are also collected: name, synonyms, ATC Code, etc. Some are used to other parts of this project, others might be useful in future versions of our code.

When reading the information of the drugs, we make sure that, for each collected drug, the three textual fields and the ATC Code are not empty. This reduces the number of drugs we use during the experiment from 10,562 to 1,661.

4.2.2 Similarity Matrix

The similarity between all pairs formed by the 1,661 drugs is computed using our measurement and saved into a redundant square matrix. We have a $n \times n$ matrix (where n is the number of drugs), then each pairwise combination from the set exists twice, once in each order, the similarity between the drug x to the drug y and vice-versa. However, those two values of similarities are the same.

In the Figure 4.2, we can see a heat map of the similarity matrix used to cluster the drugs. Please, note that, as explained in the Section 3.2.3, in this experiment we have reduced the dimensionality of the data using LSA. The main parameter of that technique is the number of final components you want, k . In this work we have used three different values: 500, 200 and 100. With them, we have computed three different similarity matrices which have been used to evaluate our similarity measurement against the ground truth (see Section 4.2.4). We did not want to cluster for the three cases, so we chose one of the three similarity matrices. For the case in which k was equal to 100, we got the best result in that evaluation against the ground truth, so we decided to use it to do the indirect evaluation based on clustering.

4.2.3 Indirect Evaluation: Clustering

Using the similarity matrix showed in the Figure 4.2 we have clustered the drugs into fourteen clusters with a Spectral Clustering. For further information about the used clustering technique or other aspects about the clustering process, like the number of clusters, please, see Section 3.5.1.

Ordered Similarity Matrix

With the obtained clusters we have ordered the columns and rows of the similarity matrix, the result is showed in the Figure 4.3. The clusters are the green/yellow squares around the diagonal of the matrix. Even though it is obvious we would like to note that the diagonal is completely drawn in yellow because it corresponds to the similarity to each drug to itself.

Clusters - Overview

As we have done before with the total number of ATC Codes. We have studied the First Level ATC Codes distribution for the number of drugs used in this experiment (1,661). In this case, the total number of ATC Codes is 3,007 and the distribution of

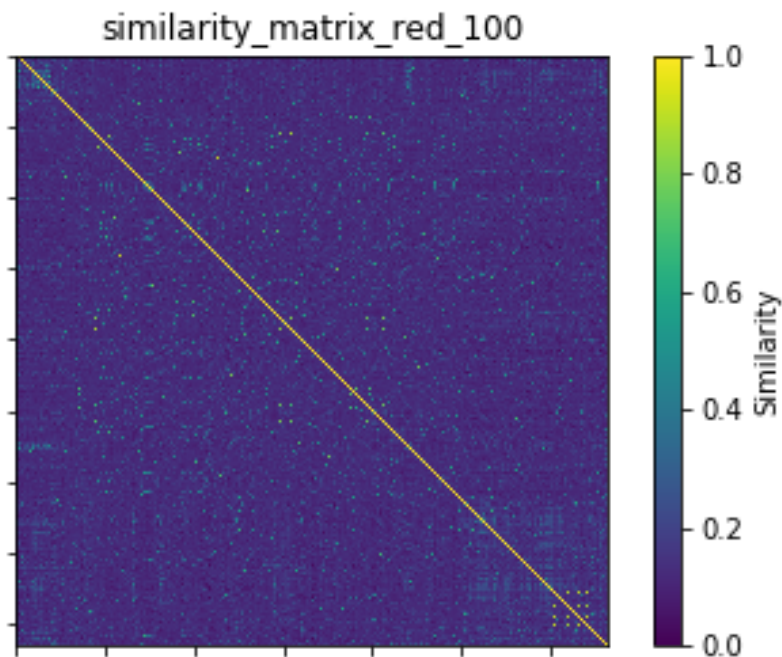


Figure 4.2: Similarity matrix based in text mining. The textual data has been reduced from original number of features to 100 using LSA.

drugs for ATC code is shown in the Figure 4.4. We have ordered the histogram from the most to the least common ATC Code.

We have fourteen clusters because we have fourteen classes of First Level in the ATC Codes. The ideal scenario, would be to see that all the drugs with one specific first level of ATC Code were together in the same cluster. However, the reality is quite different to that.

As said before, the result of the clustering is not really good. In this section, we try to figure out which could be the reasons (at least some) of this fact.

The first level of the ATC Code indicates the anatomical main (not only) group in which the drug is supposed to act. The similarity measure computed for this experiment uses textual information of three DrugBank fields:

- **Description.** Description of the drug describing general facts, composition and/or preparation.
- **Indication.** Description or common names of diseases that the drug is used to treat.
- **Pharmacodynamics.** Description of how the drug works at a clinical or physiological level.

As we see, Indication and Pharmacodynamics provide information which could be related to the anatomical group in which the drug acts. Even so, the description can include a lot of information which could be misleading to cluster the drugs into the fourteen classes of first level of ATC Codes. This could be one of the reasons why the

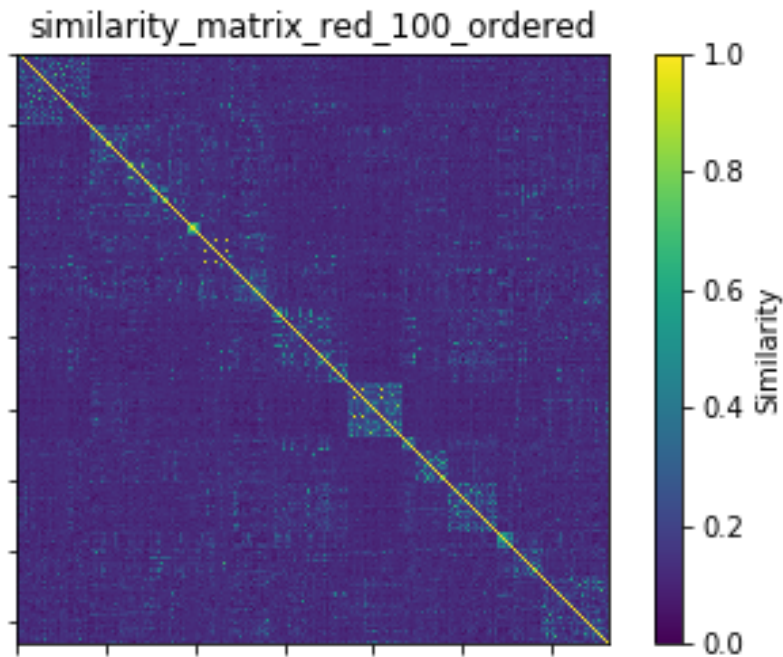


Figure 4.3: Similarity matrix based on text mining ordered using the clusters. The textual data has been reduced from original number of features to 100 using LSA.

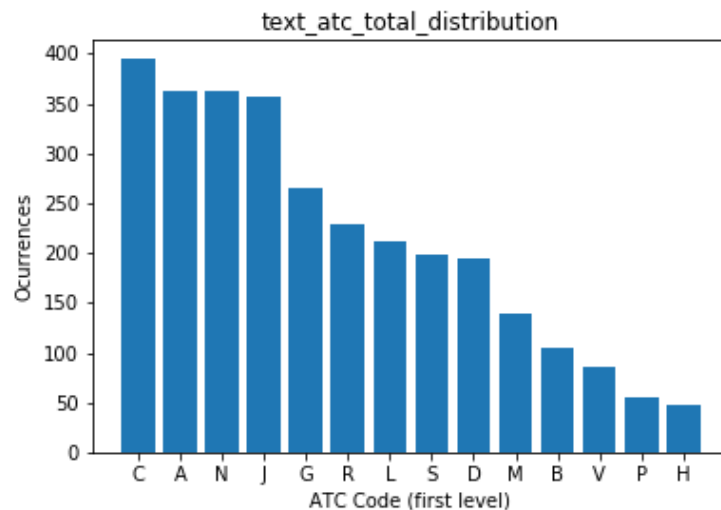


Figure 4.4: Distribution of the first level of all ATC Codes for the 1,661 drugs used in the textual mining experiment.

performance of the clustering is not perfect. In addition, there is another reason which comes from the nature of the data. As we have seen in the Figure 4.18, the distribution of our classes is really unbalanced, what increases the difficulty of clustering properly.

In spite of everything we have exposed above, we claim that we have not totally failed. Our text based similarity measure still has value, actually, we got the best clustering results for the text mining similarity experiment, as we will see later.

It is relevant to remind that our principal aim (task) was not to cluster the drugs into the fourteen chosen classes but to evaluate if the computed similarity measure was good or not, and we can say that, definitely, it is not bad. Of course, if the performance of the clustering had been really good, our conclusion about the quality of our measurement would be more positive.

Clusters - Deep Analysis

Now it is time for us to actually analyze more in depth the obtained clusters. As we have said before, the similarity measure computed in this experiment is not perfect, since the drugs are not totally grouped by their ATC Code. However, in some cases we actually have got good results. Note that even though Spectral Clustering is not completely deterministic, it is more stable than K-means. If someone runs our experiment several times, slightly different clusters may be found, but we have deprecated this and just run it once. In the Figures 4.5 and 4.6, we can see the distribution for the fourteen obtained clusters.

Looking at the distribution of all the clusters, we have divided the clusters in three different groups:

- Clusters in which the most common ATC Code represents a good percentage of the total number of occurrences of the ATC Code within the complete set of used data. This could be equivalent to the notion of *Purity* we have talked before along the Section 3.5.1, meaning, how good we are grouping in the same cluster all instances of drugs with a certain ATC Code. The clusters which show this behavior are: 0, 7, 9 and 13. The most evident example is the Cluster number 0, which includes around the 75% of the drugs with ATC Code 'C'. The number of occurrences within the cluster is around 300 (see Figure 4.5) while the total number of occurrences of that ATC Code inside the total data set we used is around 400 (see Figure 4.4). The other three examples (clusters 7, 9 and 13) show around the 50% of the instances of the ATC Codes 'J', 'G' and 'D', respectively.
- Clusters in which the most common ATC Code appears clearly more times within the cluster than the rest of ATC Codes included in the cluster. Even though it is not exactly the same, this is somehow related to the notion of *Purity*. The clusters which show this behavior are: 0, 3, 5, 6, 7, 8, 9, 10, 11 and 12. Maybe the clusters which shows better this are numbers: 7, 8, 10 and 12. Note that two of the clusters included in the previous group (7 and 9), are also here. Because they have the characteristics of both groups: the most common ATC is predominant in the cluster and it is also a good representation of the total of drugs of that ATC code.
- Clusters which are a bit meaningless for us because either they cannot be included in one of the previous cases or because the number of drugs within the cluster is too small. The clusters which show this behavior are: 1, 2 and 4. For instance, in clusters 2 and 4 we do not have a predominant ATC Code, which make difficult for us to give to the distribution a meaning. Please, note that we have not strongly claimed that those clusters are meaningless, we just cannot extract a clear conclusion from them. Maybe, with the help of some experts in the domain we could have a better understanding of our results. Might be possible to discover

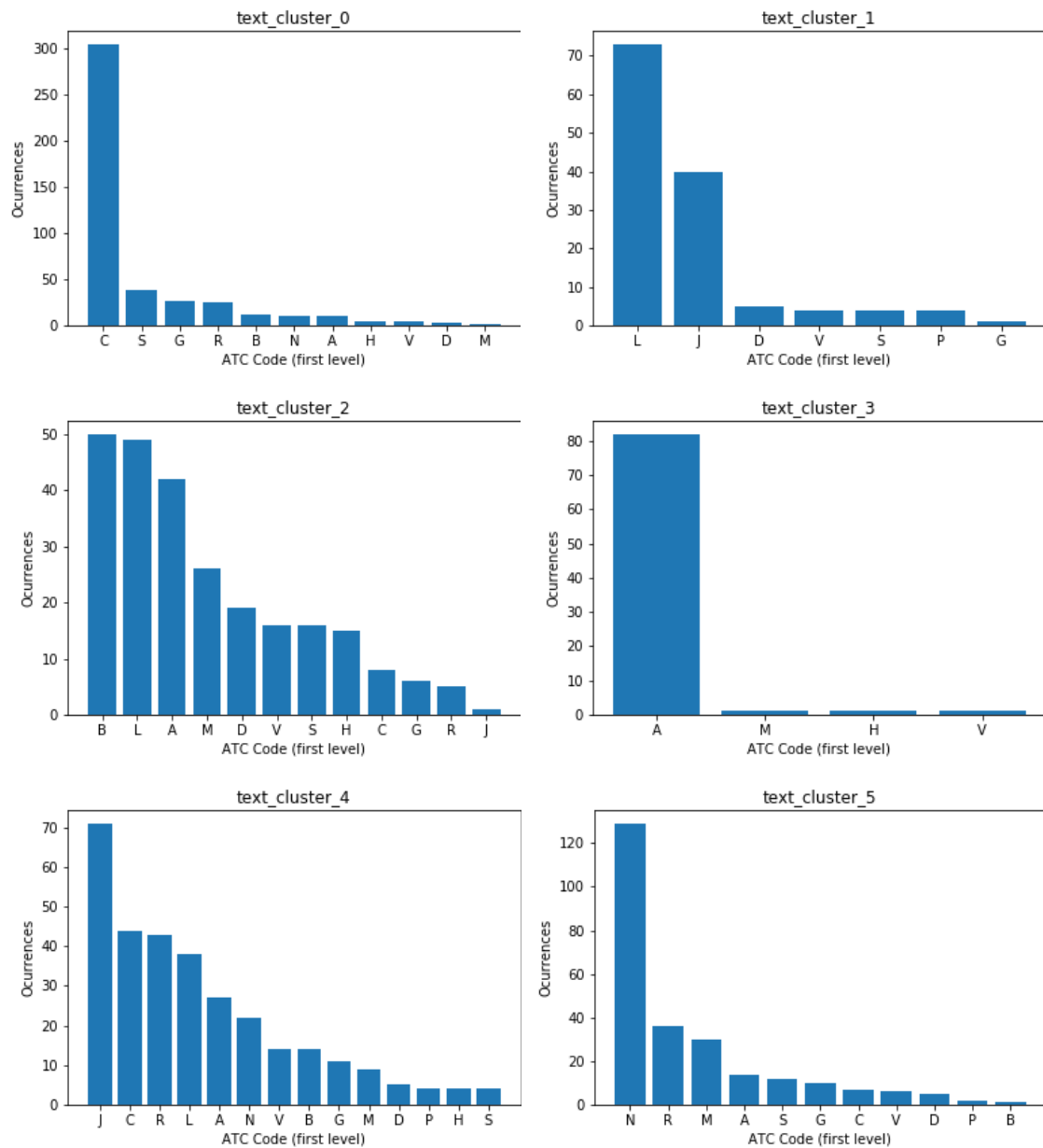


Figure 4.5: Distribution of the first level of all ATC Codes for the drugs contained within the Clusters 0-5 for the text experiment.

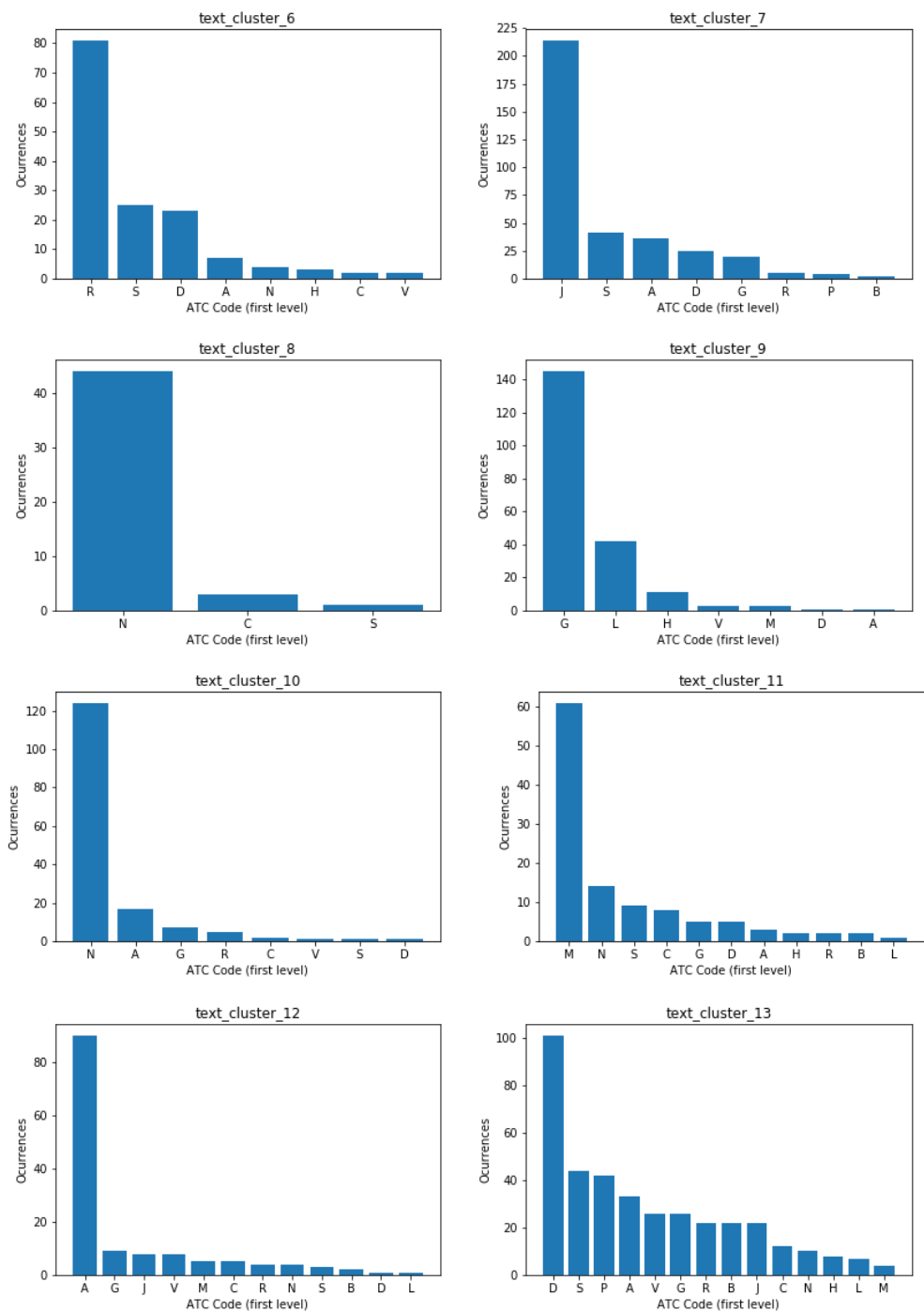


Figure 4.6: Distribution of the first level of all ATC Codes for the drugs contained within the Clusters 6-13 for the text experiment.

that it actually makes sense to have a cluster like the number 2, in which the number of ATC Codes 'B', 'L' and 'A' is closely the same. Might be that those ATC Codes are somehow related.

Conclusion

As a general note, we would like to say that from the fourteen original ATC Codes, just eight of them were detected: 'N', 'A', 'R', 'M', 'C', 'J', 'G' and 'D'. Actually, six of them are the six most common ATC Codes of the used data (see Figure 4.4). Thus, our expectations about the unbalanced nature of the data were true, those ATC Codes with less instances were more difficult to cluster inside a unique group. What 'detected', we mean that those ATC Codes were predominant in at least one of the clusters explained above.

4.2.4 Direct Evaluation: Ground Truth

This evaluation is explained in the Section 3.5.2, please, for detailed information, read that section. Basically, we use 100 pairs of drugs which have been annotated by 143 experts (henceforward, ground truth). They were asked if the two molecules were or not similar. Our aim is to see if our approach actually gives a similar answer to the one provided by those experts. Keeping in mind that aim, we have decided to evaluate our similarity measurements against the ground truth in three different dimensions or aspects:

- **Value of similarity.** Study of the correlation between the value of similarity computed by our measure and the similarity from the ground truth. For this experiment we use Pearson's Correlation Coefficient. Please, note that the value of similarity provided by the ground truth is not a degree of similarity between the drugs, but the percentage (from 0 to 1) of the experts who said that the pair of drugs were similar.
- **Order.** Study of the order or rank inferred by the value of similarity. We have ordered the pairs from the least to the most similar and then studied the rank correlation using Kendall's τ Correlation Coefficient.
- **Threshold.** We have set a threshold in order to classify our pairs of drugs into two classes: similar and non-similar. The threshold is 0.85, because the Tanimoto Coefficient, one of the similarities we use (see Section 3.4) has shown to indicate similarity between two molecules from that value. Once we have classified the pairs using our similarity measure and the ground truth, we compute the *accuracy* and the *recall*.

In this case, we have three different experiments, one for each of the values of number of components for LSA: 100, 200 and 500. In the Table 4.1, we can see all relevant information for those experiments, including all the evaluation coefficients explained above. Note that we have also included the number of pairs we have from the ground truth (97) and the pairs which are among our computed similarities (65). Originally, the ground truth is a set of 100 pairs, however, we have not been able to find all the names of some drugs from the original paper [Franco et al., 2014]. The reason is that the authors just published the molecular structure of the pairs, the SMILE representation and the decision of the experts, but not the names of the drugs. Thus,

we needed to search for the structure and the SMILE representation on different webs, and we could not find some of them. The file we have used for this evaluation is a .csv file which looks like it is shown in the Figure 3.9.

Of course, since we even have less pairs among the computed similarities (65), we just evaluate considering those pairs.

Number of components for LSA	100	200	500
Pairs in ground truth	97	97	97
Pairs in computed similarity 65	65	65	
Kendall's τ	0.2327	-0.0269	0.0125
Pearson's Correlation	0.7920	0.7385	0.6875
Accuracy	0.7385	0.7385	0.7385
Recall	0.0556	0.0556	0.056

Table 4.1: Direct Evaluation against a ground truth of the Text Based Similarity

The *recall* is really bad value in the three experiments, which means that a really small portion of the similar drugs are classified as similar. This can be because the threshold is too high and our similarity measures are below it. However, the accuracy is not bad (around 0.75), so a good portion of drugs which are classified as similar, actually, are classified properly. Those values are equal for the three experiments, so they do not give us a lot of information about choosing one of them to cluster.

As said implicitly above, we would like to identify which case of LSA reduction is better, in order to just use that one for the clustering. In order to make a decision, we chose the one with better Pearson's and Kendall's τ Correlations, LSA with 100 components. Even so, we have to say that, while Pearson's correlation values could be considered as good, it is not the case with Kendall's correlation.

As a matter of fact, we have discovered that our similarity measure based on text mining techniques, does not seem to be very useful to infer a rank of the similarity among drugs (compared to the ground truth). Actually, as a conclusion, based on the values of Accuracy, and correlations, we could say that our method is relatively good to infer the similarity between a pair of drugs. However, is not really good to infer the degree of similarity between two drugs taking into account how similar are other pairs of drugs. So that, the measure of similarity is good locally, but we cannot say that it is good globally. Even so, it is possible that this fact is not only caused by the quality of our measure. It could be caused because, as said before, the ground truth gives us information about how many experts said that a pair of drugs is similar or not. Nevertheless, the experts were not asked about which degree of similarity have those drugs, neither they were asked to say how similar are two drugs in comparison to another two other ones.

4 3

Taxonomy Based Similarity

In this section, we explain the experimental setup and the results of the experiment for the taxonomy based similarity. Note that in the upcoming paragraphs we just talk

about the experiment itself, avoiding some details. For more detailed description about how the similarity measure is computed, please, see Section 3.3.

4.3.1 Experimental Setup

Our aim is to measure similarity between drugs, in particular, using taxonomic information about those drugs. The taxonomic structure of the drugs is built using the information from the complete DrugBank database file (.xml format). Specifically, we use all the fields related to the classification tag: kingdom, superclass, class, etc.

It could be possible to build a whole graph of the DrugBank, with all the drugs (10,562), however, we have chosen to use the same number as we used in the text experiment (see Section 4.2). The reason is that if we used all the drugs, the cost of computation of the similarity matrices would be really high. Actually, we consider that using the same amount of drugs between two experiments could be interesting, in order to see how two distinct approaches address the exact same problem.

When reading the information of the drugs, we make sure that, for each collected drug, the three textual fields (used in the previous section) and the ATC Code are not empty. This reduces the number of drugs we use during the experiment from 10,562 to 1,661.

4.3.2 Similarity Matrix

The similarity between all pairs formed by the 1,661 drugs is computed using our measurement (see below) and saved into a redundant square matrix. We have a $n \times n$ matrix (where n is the number of drugs), then each pairwise combination from the set exists twice, once in each order, the similarity between the drug x to the drug y and vice-versa. However, those two values of similarities are the same.

Just as a reminder, for this experiment, we have built two different sorts of graphs: unweighted and weighted. For each graph, we compute the distance between the drugs (as the shortest path in the taxonomic structure) and then the similarity using the method proposed by Leacock and Chodorow [Leacock and Chodorow, 1998]. Please, note that all of this, is explained in the Section 3.3. The similarities are put into two matrices, which can be visually inspected in the Figures 4.7 and 4.8.

As it is shown in the Figures 4.7 and 4.8, the weighted similarity matrix looks darker, that is why it contains smaller values of similarity (closer to zero). This totally reasonable, because its values of distances between drugs are greater (paths are weighted), thus, when we normalize the distance and compute the the similarity the result is smaller for this case.

4.3.3 Indirect Evaluation: Clustering

Making use of the similarity matrices showed in the Figures 4.7 and 4.8 we have clustered the drugs into fourteen clusters with a Spectral Clustering. For further information about the used clustering technique or other aspects about the clustering process, like the number of clusters, please, see Section 3.5.1.

Ordered Similarity Matrix

With the obtained clusters we have ordered the columns and rows of the similarity matrices, the result is showed in the Figures 4.9 and 4.10. The clusters are the

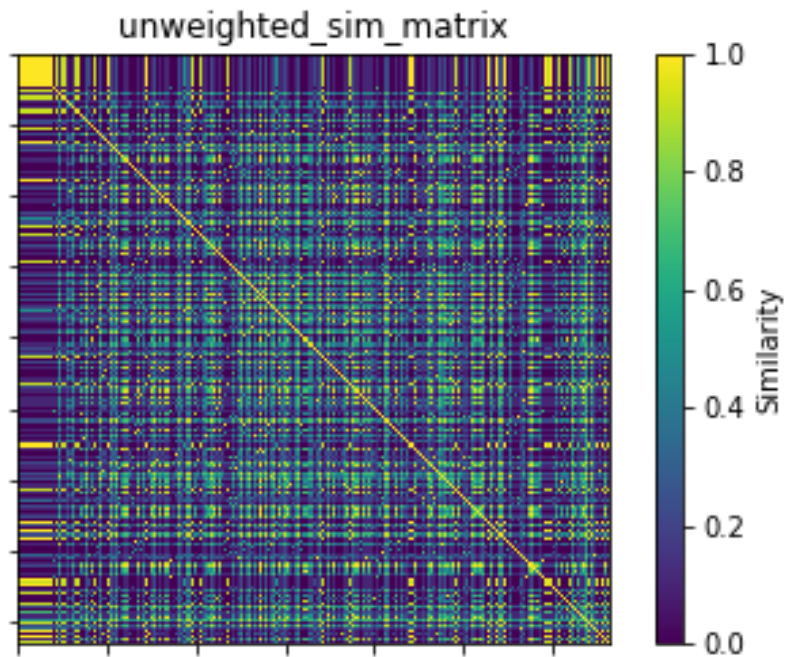


Figure 4.7: Similarity matrix based on taxonomy for the case: unweighted graph.

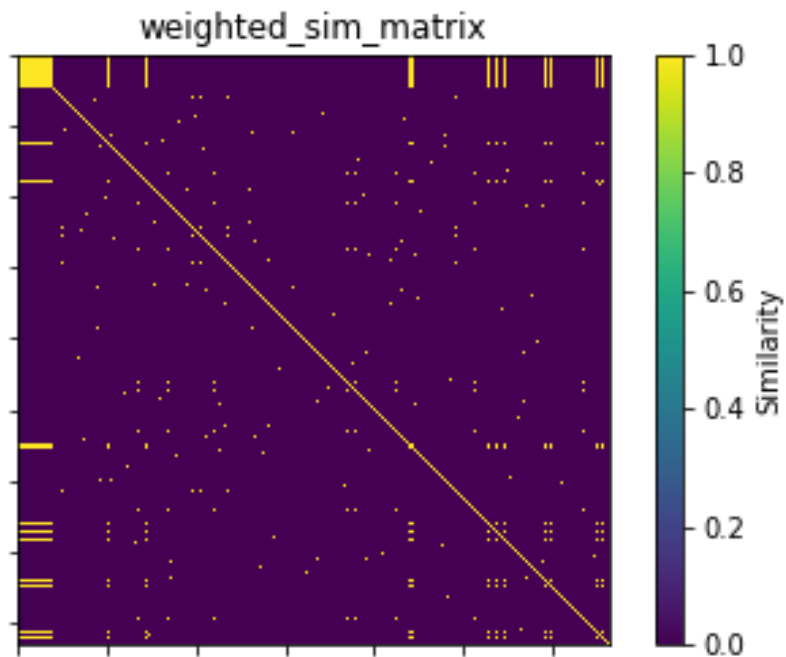


Figure 4.8: Similarity matrix based on taxonomy for the case: weighted graph.

green/yellow squares around the diagonal of the matrix. Even though it is obvious we would like to note that the diagonal is completely drawn in yellow because it

corresponds to the similarity to each drug to itself.



Figure 4.9: Similarity matrix based on taxonomy ordered using the clusters for the case: unweighted graph.

Clusters - Overview

As in the previous experiment, we have studied the First Level ATC Codes distribution for the number of drugs used in this experiment (1,661). As in the previous experiment, the total number of ATC Codes is 3,007 and the distribution is shown in the Figure 4.11. We have ordered the histogram from the most to the least common ATC Code.

We have fourteen clusters because we have fourteen classes of First Level in the ATC Codes. The ideal scenario, would be to see that all the drugs with one specific first level of ATC Code were together in the same cluster, nevertheless, it is not the case. In the upcoming paragraphs, we try to understand which could be the cause of that fact.

The first level of the ATC Code indicates the anatomical main (not unique) group in which the drug is supposed to act. The similarity measure computed for this experiment uses taxonomic information of the DrugBank Classification. In principle, there is not a clear relationship between the DrugBank Classification and the ATC Code Classification, so the performance of the clustering is not expected to be perfect. In addition, there is another reason which comes from the nature of the data. As we have seen in the Figure 4.11, the distribution of our classes is really unbalanced, what increases the difficulty of clustering properly.

Nevertheless, our principle aim is not cluster but just to evaluate if our similarity measurement is good. Of course, the better we cluster, the better could be our measure, but it is important to note that our task is not to cluster perfectly.

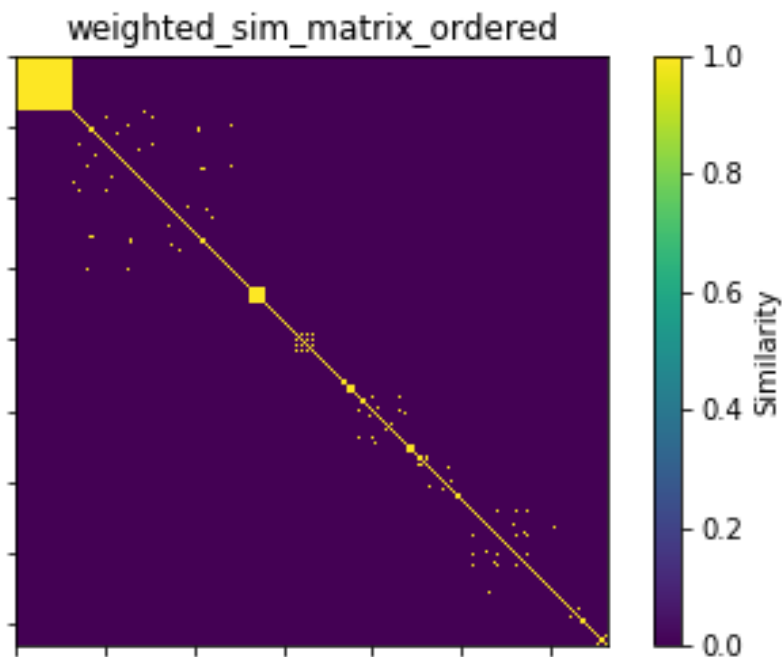


Figure 4.10: Similarity matrix based on taxonomy ordered using the clusters for the case: weighted graph.

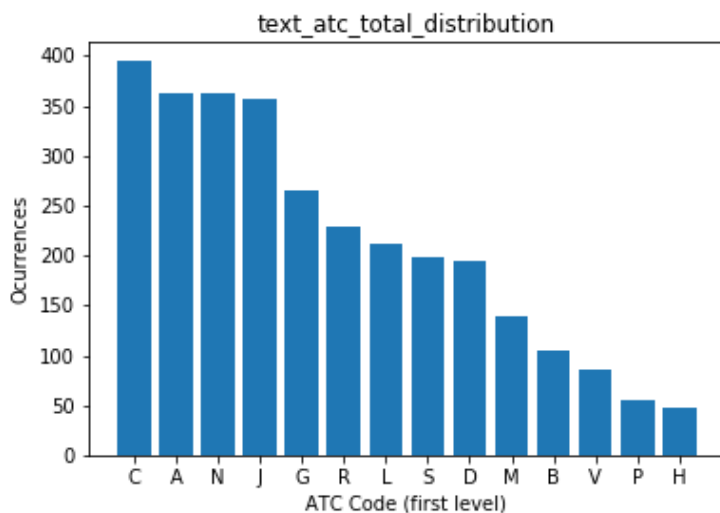


Figure 4.11: Distribution of the first level of all ATC Codes for the 1,661 drugs used in the taxonomic experiment.

Clusters - Deep Analysis

As we have said before, the similarity measure computed in this experiment is not perfect, since the drugs are not totally grouped by their ATC Code. Even so, in some cases we actually have got good results.

In this experiment, we have used two different graphs, weighted and unweighted,

however, the clustering results are equal for both of them. This actually make sense since our clustering algorithm is relatively stable (not deterministic though) and with the weights we have just scaled the similarities. Consequently, we have decided to show and analyze just one of the cases: the weighted. In the Figures 4.12 and 4.13, we can see the distribution for the fourteen obtained clusters.

Please, note that the clusters will remain the same if the used data does not change a lot, since the Spectral Clustering is more or less stable (not deterministic though). However, if someone ran our experiment several times, slightly different clusters could be found, but we have deprecated this and just run it once.

As we did in the previous experiment, we have divided the clusters in three different types:

- Clusters in which the most common ATC Code represents a good percentage of the total number of occurrences of the ATC Code within the complete set of used data. This could be equivalent to the notion of *Purity* we have talked before (see Section 3.5.1), meaning, how good we are grouping in the same cluster all instances of drugs with a certain ATC Code. The only cluster we have which this characteristics is the number 12, in which we have grouped around the 50% of the instances of the ATC Code 'G'.
- Clusters in which the most common ATC Code appears clearly more times within the cluster than the rest of ATC Codes included in the cluster. Even though it is not exactly the same, this is somehow related to the notion of *Purity*. Just one cluster could be included in this category, the number 6. In that cluster, the most common ATC Code, 'J', appears around four times the sum of all the rest of the ATC Codes which are inside the cluster. Note that Cluster 12 could be also included in this group.
- Clusters which are a bit meaningless for us because either they cannot be included in one of the previous cases or because the number of drugs within the cluster is too small. The rest of the clusters would be part of this last group. We cannot easily extract conclusions. Just an interesting additional comment, we have found three clusters (1, 3 and 4) in which the four most common ATC Codes are 'J', 'N', 'C' and 'A'. They do not appear in the same order, but it is clear that our similarity measure has found a connection between those drugs. It would be really interesting to study this fact with experts, maybe, some drugs of those ATC Codes are actually similar because medical reasons we are not aware of. Or might be caused because those four ATC Codes are the most common within the data we work with (see Figure 4.11).

Conclusion

For this experiment, we did not get as good results from the clustering as in the text based experiment (Section 4.2). However, we can at least see that our similarity measurement has good properties and could be useful, maybe not individually but used together with other measures.

4.3.4 Direct Evaluation: Ground Truth

This evaluation is explained in the Section 3.5.2, please, for detailed information, read that section. Basically, we use 100 pairs of drugs which have been annotated by 143

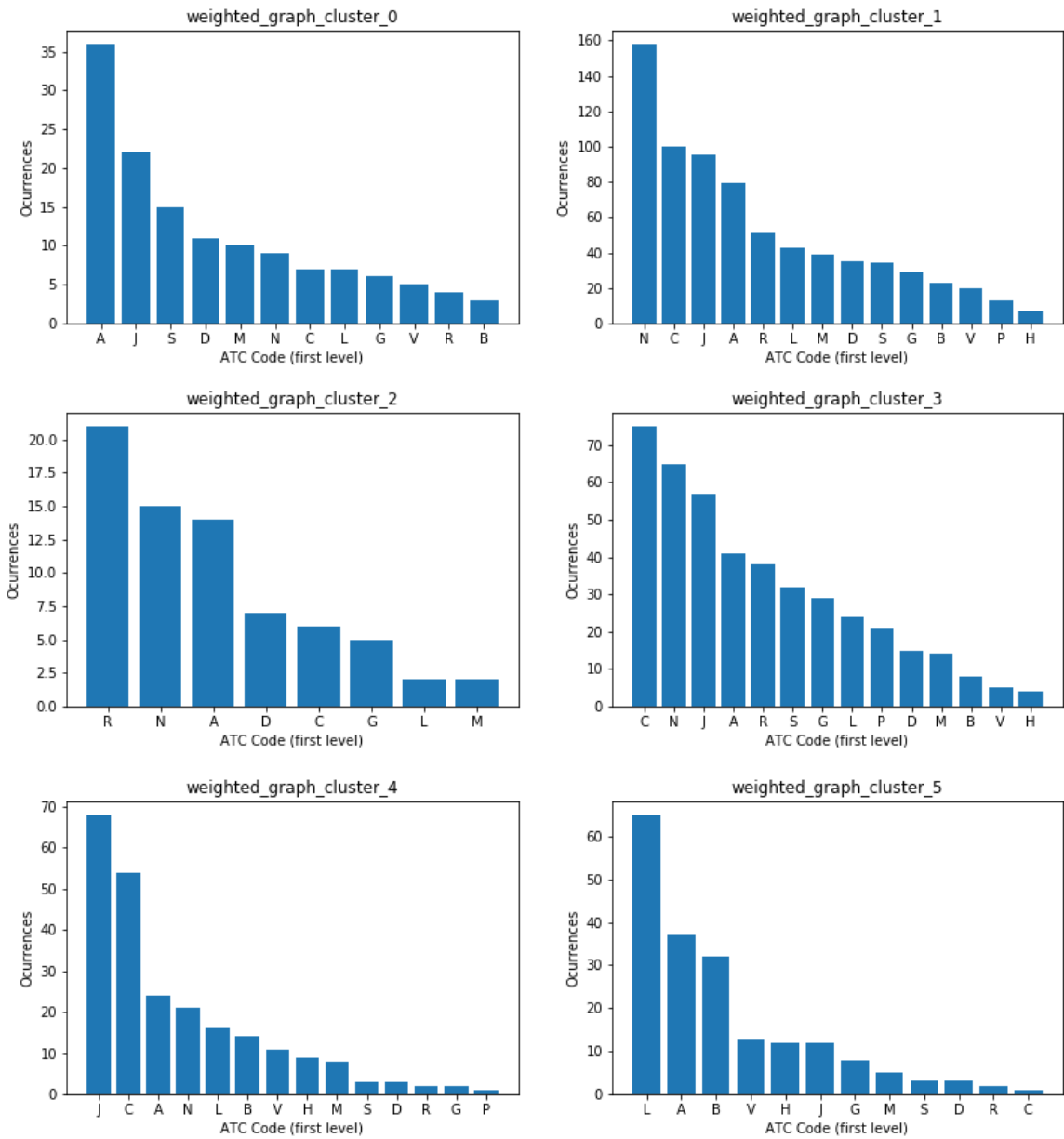


Figure 4.12: Distribution of the first level of all ATC Codes for the drugs contained within the Clusters 0-5 for the taxonomy experiment.

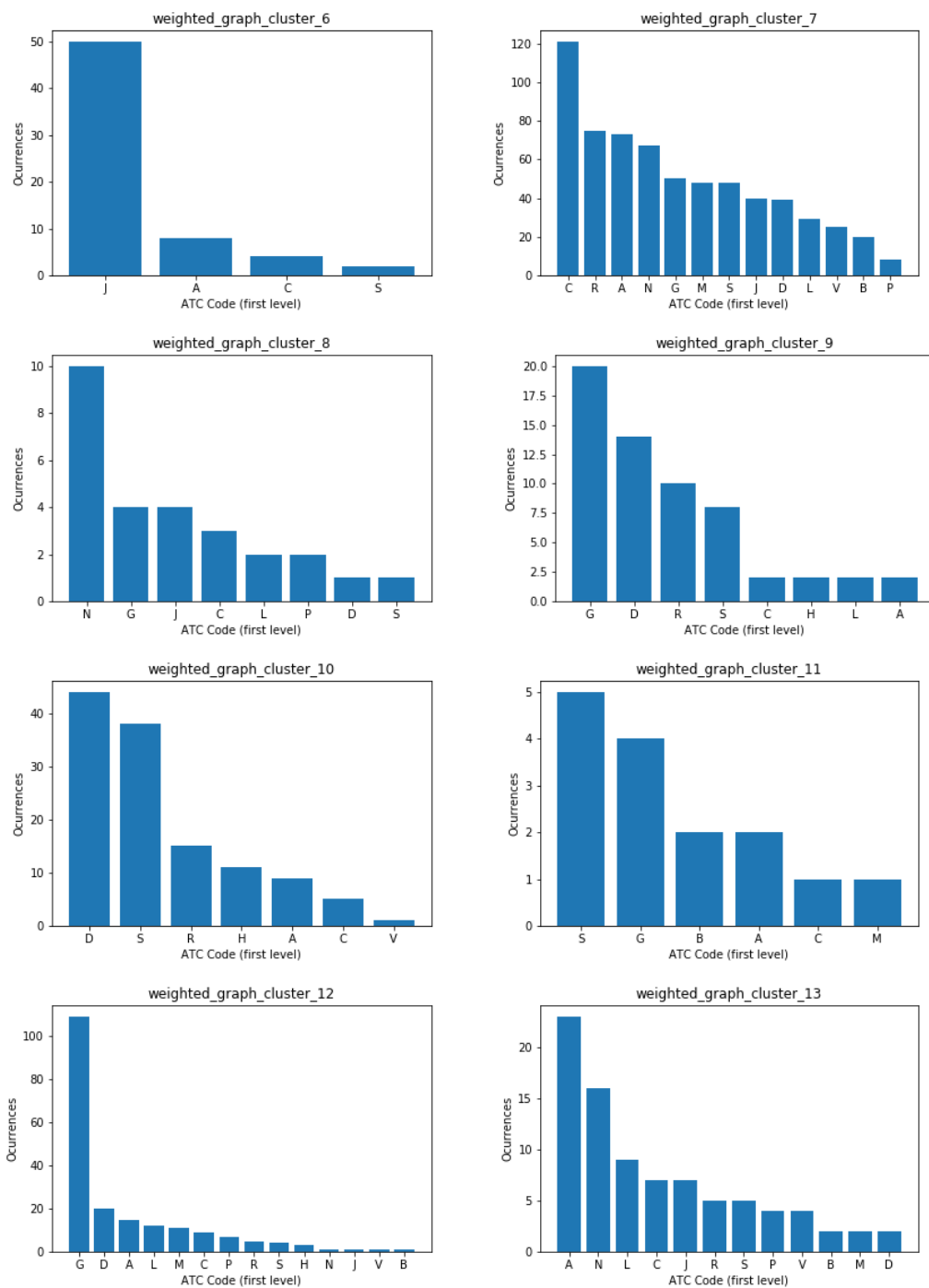


Figure 4.13: Distribution of the first level of all ATC Codes for the drugs contained within the Clusters 6-13 for the taxonomy experiment.

experts (henceforward, ground truth). They were asked if the two molecules were or not similar. Our aim is to see if our approach actually gives a similar answer to the one provided by those experts. Keeping in mind that aim, we have decided to evaluate our similarity measurements against the ground truth in three different dimensions or aspects:

- **Value of similarity.** Study of the correlation between the value of similarity computed by our measure and the similarity from the ground truth. For this experiment we use Pearson’s Correlation Coefficient. Please, note that the value of similarity provided by the ground truth is not a degree of similarity between the drugs, but the percentage (from 0 to 1) of the experts who said that the pair of drugs were similar.
- **Order.** Study of the order or rank inferred by the value of similarity. We have ordered the pairs from the least to the most similar and then studied the rank correlation using Kendall’s τ Correlation Coefficient.
- **Threshold.** We have set a threshold in order to classify our pairs of drugs into two classes: similar and non-similar. The threshold is 0.85, because the Tanimoto Coefficient, one of the similarities we use (see Section 3.4) has shown to indicate similarity between two molecules from that value. Once we have classified the pairs using our similarity measure and the ground truth, we compute the *accuracy* and the *recall*.

In this case, we have two different experiments: unweighted and weighted graph. In the Table 4.2, we can see all relevant information for those experiments, including all the evaluation coefficients explained above. Note that we have also included the number of pairs we have from the ground truth (97) and the pairs which are among our computed similarities (65). Originally, the ground truth is a set of 100 pairs, however, we have not been able to find all the names of some drugs from the original paper [Franco et al., 2014]. The reason is that the authors just published the molecular structure of the pairs, the SMILE representation and the decision of the experts, but not the names of the drugs. Thus, we needed to search for the structure and the SMILE representation on different webs, and we could not find some of them. The file we have used for this evaluation is a .csv file which looks like it is shown in the Figure 3.9.

Of course, since we even have less pairs among the computed similarities (65), we just evaluate considering those pairs.

Graph	Unweighted	Weighted
Pairs in ground truth	97	97
Pairs in computed similarity	65	65
Kendall’s τ	0.2212	0.0673
Pearson’s Correlation	0.6721	0.6998
Accuracy	0.7538	0.7692
Recall	0.7222	0.7778

Table 4.2: Direct Evaluation against a ground truth of the Taxonomy Based Similarity

Following the data showed in the Table 4.2 we could claim:

1. In both cases, there exists certain positive correlation between the values of our computed similarities and the ground truth (see Pearson's Correlation).
2. The Classification using the threshold shows good results for both, weighted and unweighted, since the values of *accuracy* and *recall* are relatively good.
3. The weighted path shows that there is not rank correlation (Kendall's τ), while the unweighted improves that value.

As we see, the result of the correlation using the Kendall's τ is not good at all. It says that there does not exist any rank correlation. We expected to find correlation, but actually, the ground truth we use should not be expected to infer any order between the similarity of the pairs. The experts were not asked about which degree of similarity a pair of drugs had, neither they were asked to say how similar were two drugs in comparison to another two other ones. They were asked just to say if two drugs were or not similar. Then, a percentage of the experts saying 'yes' was calculated and used here in our experiment.

4 4

Molecular Structure Based Similarity

In this section, we explain the experimental setup and the results of the experiment for the molecular structure based similarity. Note that in the upcoming paragraphs we just talk about the experiment itself, avoiding some details. For more detailed description about how the similarity measure is computed, please, see Section 3.4.

4 4 1 Experimental Setup

Our aim is to measure similarity between drugs, in particular, utilizing the similarity between the molecular structure of those drugs. The molecular information of each drug is gathered from the specific DrugBank file devoted to it (.sdf format). There are several fields within the database which contain information about the molecular structure of the drugs: 2D and 3D structure, different sorts of representation, etc. We could access to the molecular structure using the complete database (.xml file), however, it is easier and more powerful to use the specific file devoted to the structure of the molecules (.sdf file). The reason is because we are using a Python library, RDKit, which includes several methods to extract exploit the potential of the .sdf files.

As said in previous sections, DrugBank is not totally complete, some drugs are missing part of their fields. In this case, we read the drugs from the .sdf file using a method of RDKit. That method is able to read the file and generate a list of molecules, however, some of those molecules are not complete and we discard them. This reduces the number of drugs we use during the experiment from 10,562 to 8,738, still, a really good number and quite greater than in the other two experiments.

4 4 2 Similarity Matrix

The similarity between all pairs formed by the 8,738 drugs is computed using our measurement and saved into a redundant square matrix. We have a $n \times n$ matrix (where n is the number of drugs), then each pairwise combination from the set exists

twice, once in each order, the similarity between the drug x to the drug y and vice-versa. However, those two values of similarities are the same.

RDKit library contains methods to compute automatically the fingerprints (binary vectors representing the molecular structures) from the molecules we have read before. We generate the two most well-known different sorts of fingerprints: MACCS (167 bits) and ECFP (1024 bits). We want to analyze the performance of both, which is expected to be different, since ECFP offers more precision (more bits).

The similarity matrices are built using the Tanimoto Coefficient, which measures how different those fingerprints (binary vectors) are from each other. Please, remember that we have used another coefficient, Dice, but we do not show the results here and we do not use it for clustering because it is extremely correlated to Tanimoto. The study of correlation between the similarity matrices obtained with each of the coefficients appears within the Python notebooks we have in our Git repository. The aspect of the two similarity matrices is shown in the Figures 4.14 and 4.15.

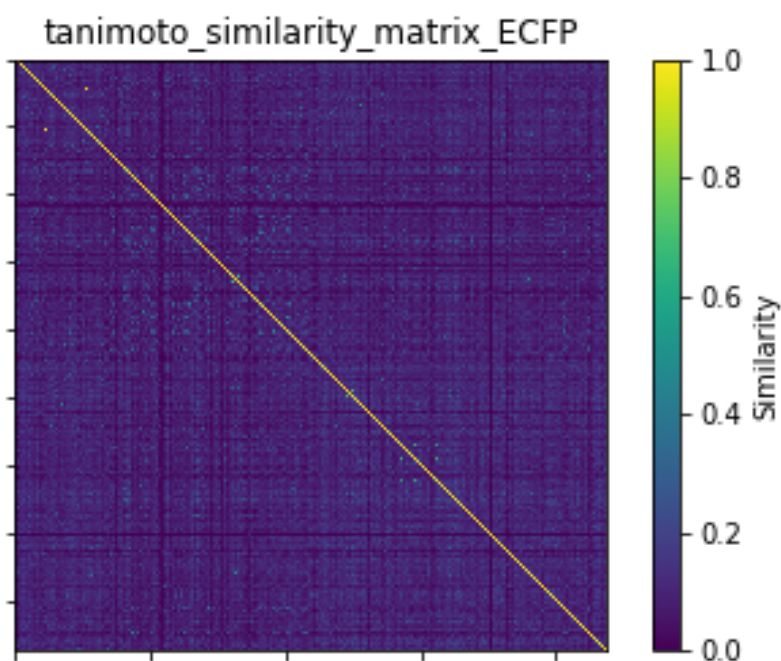


Figure 4.14: Similarity matrix based on molecular structure similarity. The molecular structure has been represented using ECFP fingerprints (1,024 bits).

We can observe in the visualization of the matrices that the values of similarity of the MACCS fingerprints are greater (closer to yellow) than in the case of using ECFP fingerprints. This is an expected behavior, since ECFP fingerprints have quite more bits, it is reasonable that the drugs are less similar between each other (since their representation is more specific/precise). We also studied the correlation between these two similarity matrices, but the correlation was around 0.6. Even though that value of correlation shows certain positive correlation, a value of correlation is considered as relevant from 0.7. Thus, we decided to continue the experiment (with the complete evaluation) using both similarity matrices.

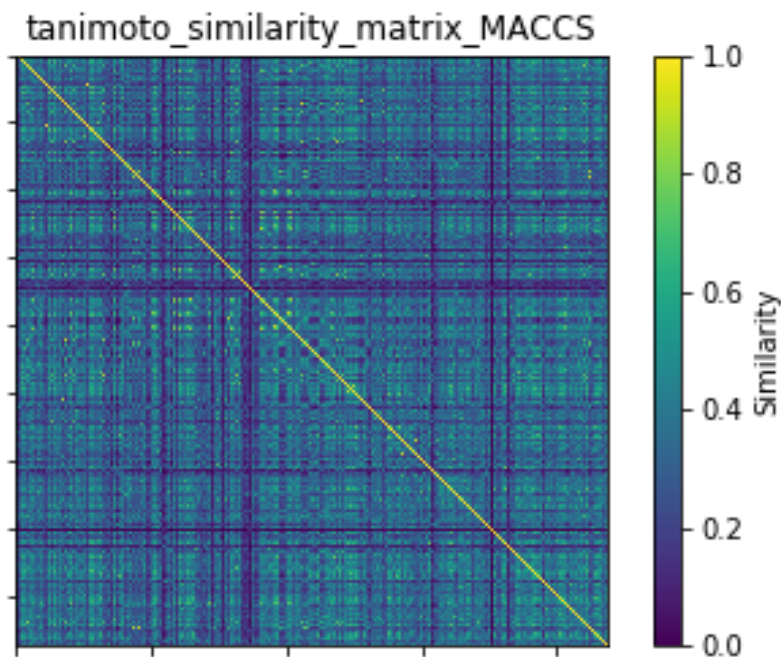


Figure 4.15: Similarity matrix based on molecular structure similarity. The molecular structure has been represented using MACCS fingerprints (167 bits).

4 4 3 Indirect Evaluation: Clustering

Using the similarity matrices showed in the Figures 4.14 and 4.15, we have clustered the drugs into fourteen clusters with a Spectral Clustering. For further information about the used clustering technique or other aspects about the clustering process, like the number of clusters, please, see Section 3.5.1.

Ordered Similarity Matrix

With the obtained clusters we have ordered the columns and rows of the similarity matrix, the result is showed in the Figure 4.3. The clusters are the green/yellow squares around the diagonal of the matrix. Even though it is obvious we would like to note that the diagonal is completely drawn in yellow because it corresponds to the similarity to each drug to itself.

Clusters - Overview

Even though we compute the similarity between pairs using 8,738 drugs, as we have explained in the Section 4.1, from the total number of drugs (10,562), just 2,287 has a non-empty ATC Code. In addition, in this experiment we are not using all but just 8,738 drugs, so the number of drugs with non-empty ATC Code are even less, specifically, 2,003 drugs. Thus, this evaluation is not going to be done over all the similarity values we have computed but just the ones with ATC Code. In particular, the total number of ATC Codes is 3,512 and the distribution is shown in the Figure 4.18. We have ordered the histogram from the most to the least common ATC Code.

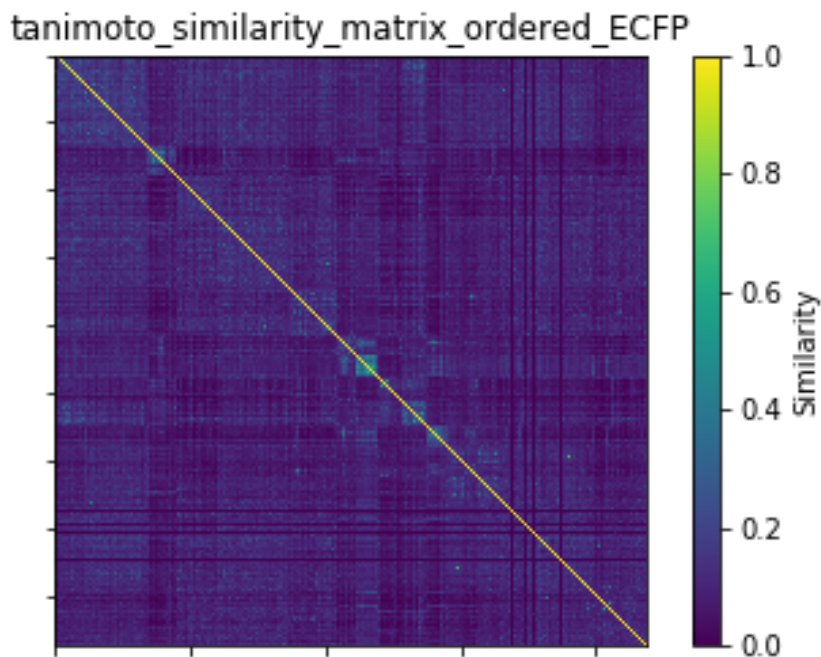


Figure 4.16: Similarity matrix based on molecular structure similarity ordered using the clusters. The molecular structure has been represented using ECFP fingerprints (1,024 bits).

We have fourteen clusters because we have fourteen classes of First Level in the ATC Codes. The ideal scenario, would be to see that all the drugs with one specific first level of ATC Code were together in the same cluster. However, the reality is obviously different to that. In this part of the document, we try to find possible causes of the bad performance of the clustering.

The first level of the ATC Code indicates the anatomical main (not only) group in which the drug is supposed to act. The similarity measure computed for this experiment uses the molecular information. Based on the similar property principle of Johnson and Maggiora, which states: *similar compounds have similar properties* [Johnson and Maggiora, 1990], we could say that drugs with similar molecular structure would have similar properties.

There is a drawback though, since that principle is not always true. Furthermore, the principle talks about the properties of a molecule (drug, in our case). However, the first level of the ATC Codes gives information about the main anatomical group in which the drug is meant to act, which not always has to be related to the properties of a drug. In addition, there is another reason which comes from the nature of the data. As we have seen in the Figure 4.18, the distribution of our classes is really unbalanced, what increases the difficulty of clustering properly.

For those reasons, we could justify the relatively good but not great performance of the clustering. Nevertheless, this fact does not mean we have totally failed. Our molecular structure based similarity measure still has value, as we will see later. Note that our principal aim was not to cluster the drugs into the fourteen chosen classes, but to build some similarity measures and evaluate if they were good or not, and we

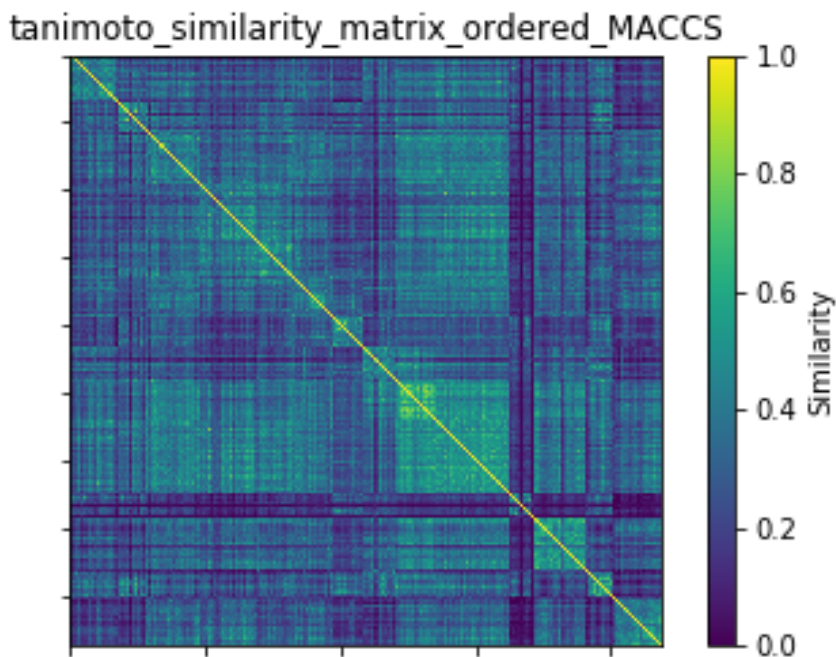


Figure 4.17: Similarity matrix based on molecular structure similarity ordered using the clusters. The molecular structure has been represented using MACCS fingerprints (167 bits).

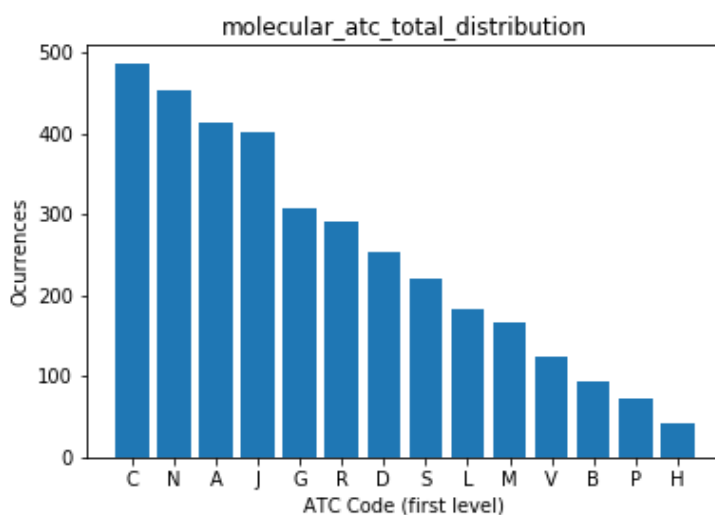


Figure 4.18: Distribution of the first level of all ATC Codes for the 8,738 drugs used in the molecular structure experiment.

can say that, definitely, in this case is not bad.

Clusters - Deep Analysis

In this section, we analyze more deep in detail the obtained clusters. As we have said before, the similarity measure computed in this experiment is not perfect, since the

drugs are not totally grouped by their ATC Code. However, in some cases we actually have got good results.

In this experiment, we have worked with two different sorts of fingerprints so that we have all the results in duplicate. Even so, the results and conclusions are not too different no matters if we utilize ECFP or MACCS. Therefore, we have decided to comment all the clusters at the same time (always referring our thoughts to the specific experiment).

Please, note that even though Spectral Clustering is not completely deterministic, it is more stable than K-means. If someone runs our experiment several times, slightly different clusters could be found, but we have deprecated this and just run it once. In the Figures 4.19 and 4.20, we can see the distribution for the fourteen obtained clusters when using the ECFP fingerprints. In the Figures 4.21 and 4.22, we can see the distribution for the fourteen obtained clusters when using the MACCS fingerprints.

In order to make easier the analysis of the results, we have divided the clusters in three different groups:

- Clusters in which the most common ATC Code represents a good percentage of the total number of occurrences of the ATC Code within the complete set of used data. This could be equivalent to the notion of *Purity* we have talked before (see Section 3.5.1), meaning, how good we are grouping in the same cluster all instances of drugs with a certain ATC Code. In this experiment, we have not been able to find clusters with this characteristics.
- Clusters in which the most common ATC Code appears clearly more times within the cluster than the rest of ATC Codes included in the cluster. Even though it is not exactly the same, this is somehow related to the notion of *Purity*. The clusters which show this behavior are: 1 and 7 (for ECFP) and 1 and 2 (for MACCS). In the ECFP clusters number 1 and 7 we have just one predominant ATC Code, 'N' and 'C', respectively. While in the clusters number 1 and 2 we have just one predominant ATC Code, 'N' and 'G', respectively.
- Clusters which are a bit meaningless for us because either they cannot be included in one of the previous cases or because the number of drugs within the cluster is too small. The rest of the clusters could be included within this group. Please, note that we have not strongly claimed that those clusters are meaningless, we just cannot extract a clear conclusion from them. Maybe, with the help of some experts in the domain we could have a better understanding of our results. Might be possible to discover that it actually makes sense to have clusters like the number 2 and number 6 (ECFP, Figures 4.19 and 4.20), in which the number of ATC Codes 'J', 'C', 'N' and 'A' is closely the same. Might be that those ATC Codes are somehow related. We are not experts in the medical domain, but it seems to be a relationship between those four ATC Codes, which are: *Antiinfectives for systemic use*, *Cardiovascular System*, *Alimentary tract and metabolism* and *Nervous System*. We observe a similar behavior in the MACCS clusters number 5 and 13 (Figures 4.21 and 4.22). Of course, this is just a conclusion extracted from our results, would be necessary to ask to several experts if this actually makes sense. In fact, those four ATC Codes, are the most common ones within our data, so maybe we group them just because there are more of them (see Figure 4.18).

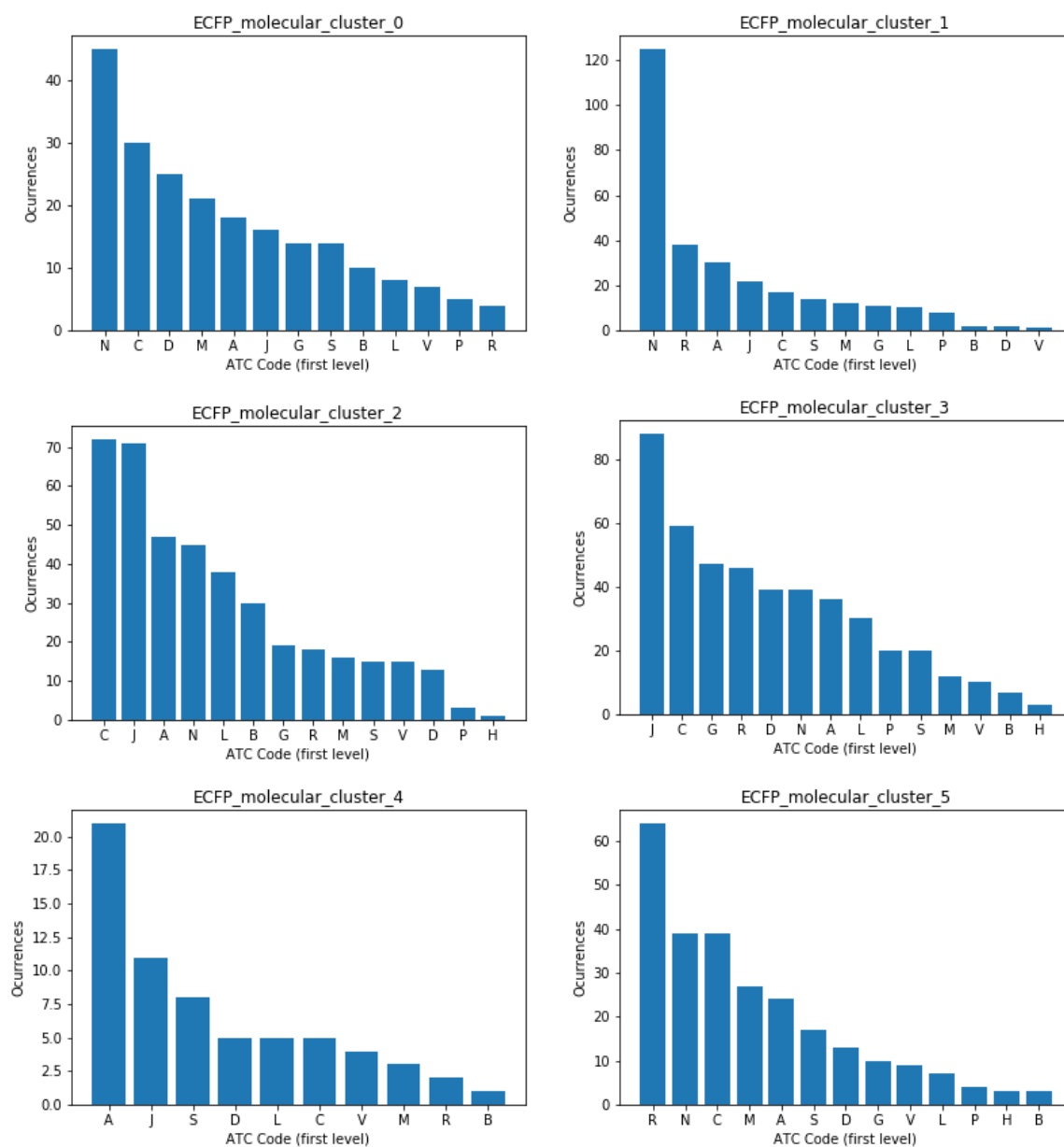


Figure 4.19: Distribution of the first level of all ATC Codes for the drugs contained within the Clusters 0-5 for the molecular structure experiment. The clustering was done using the similarity matrix computed with the ECFP fingerprints.

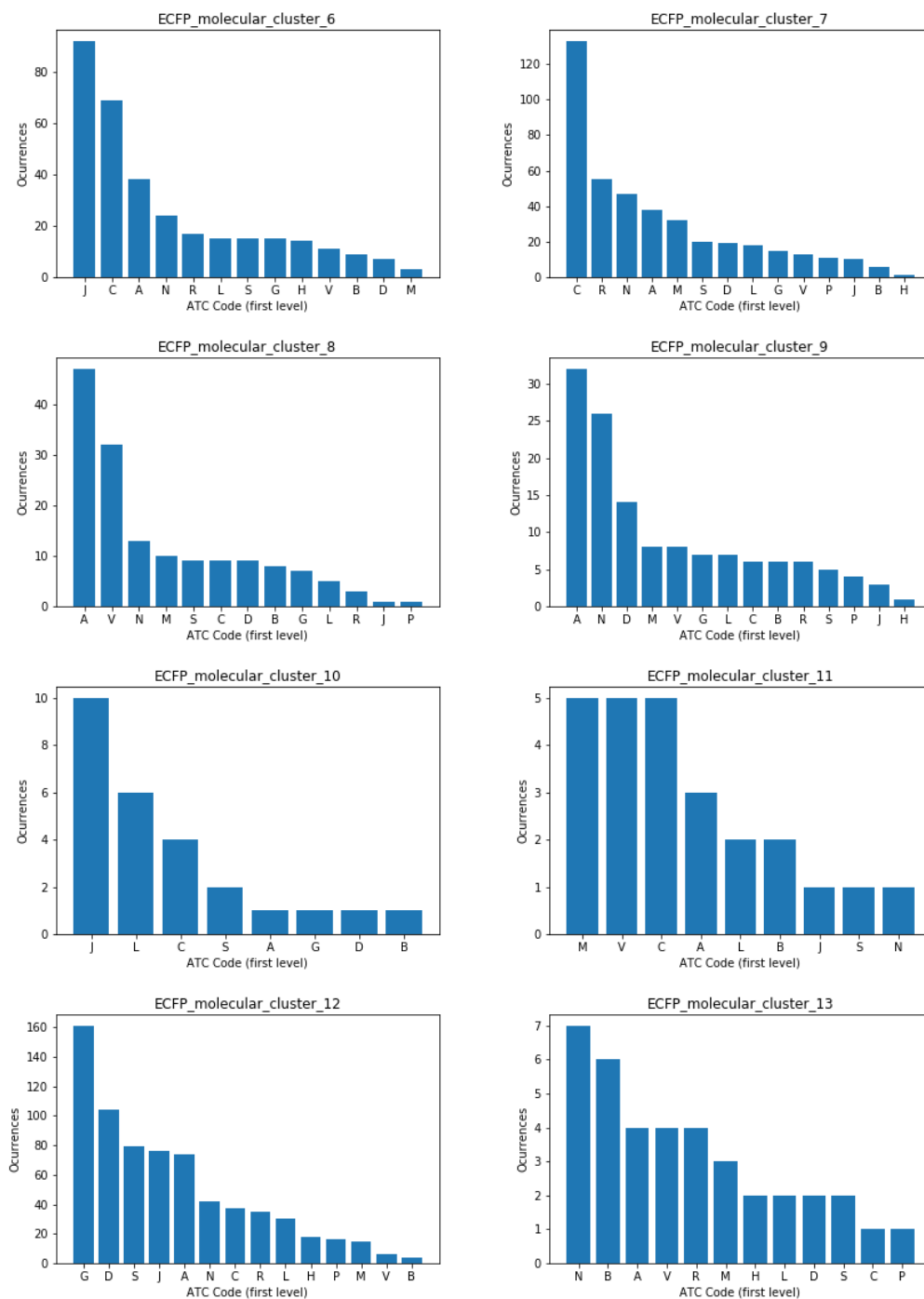


Figure 4.20: Distribution of the first level of all ATC Codes for the drugs contained within the Clusters 6-13 for the molecular structure experiment. The clustering was done using the similarity matrix computed with the ECFP fingerprints.

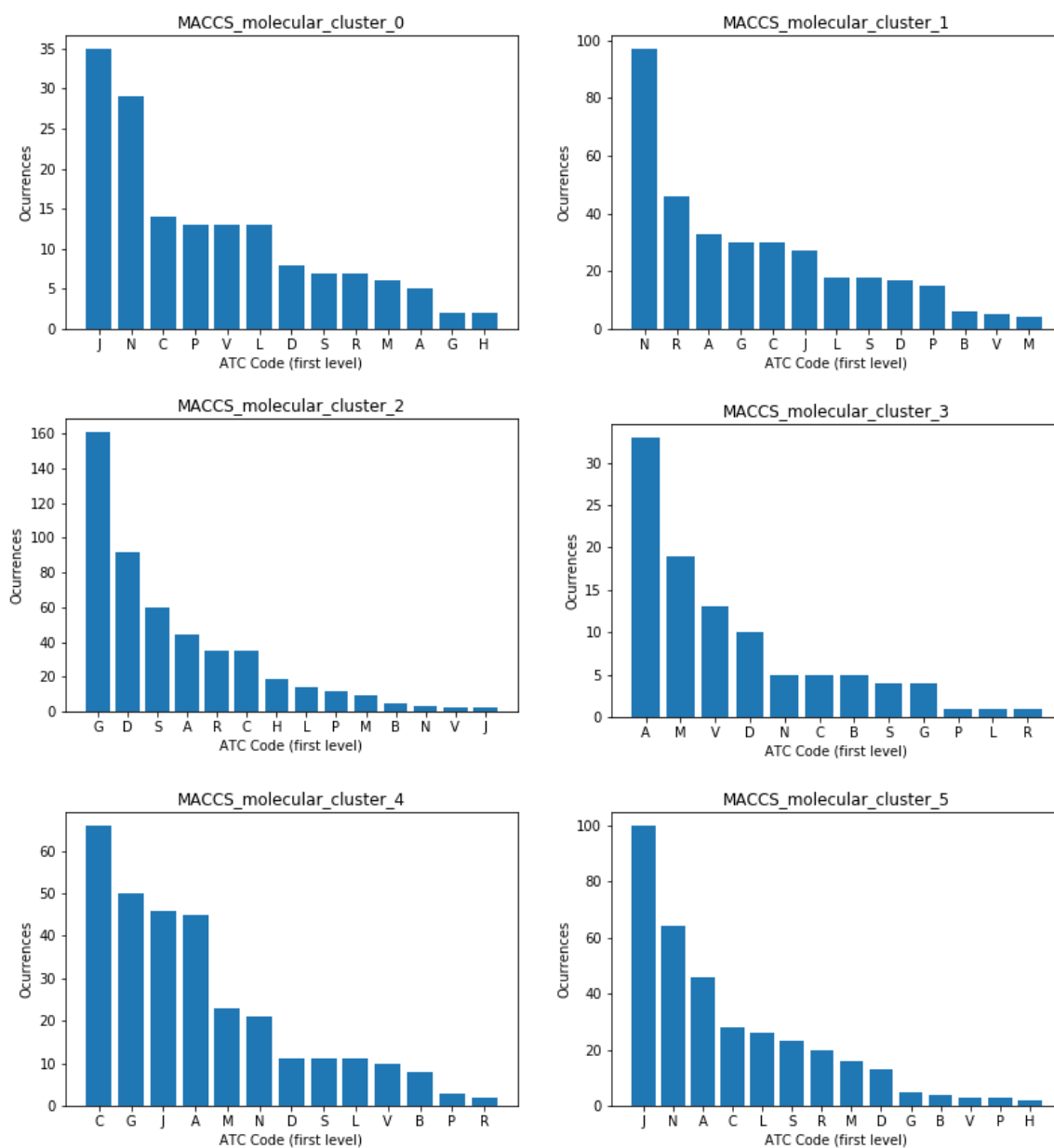


Figure 4.21: Distribution of the first level of all ATC Codes for the drugs contained within the Clusters 0-5 for the molecular structure experiment. The clustering was done using the similarity matrix computed with the MACCS fingerprints.

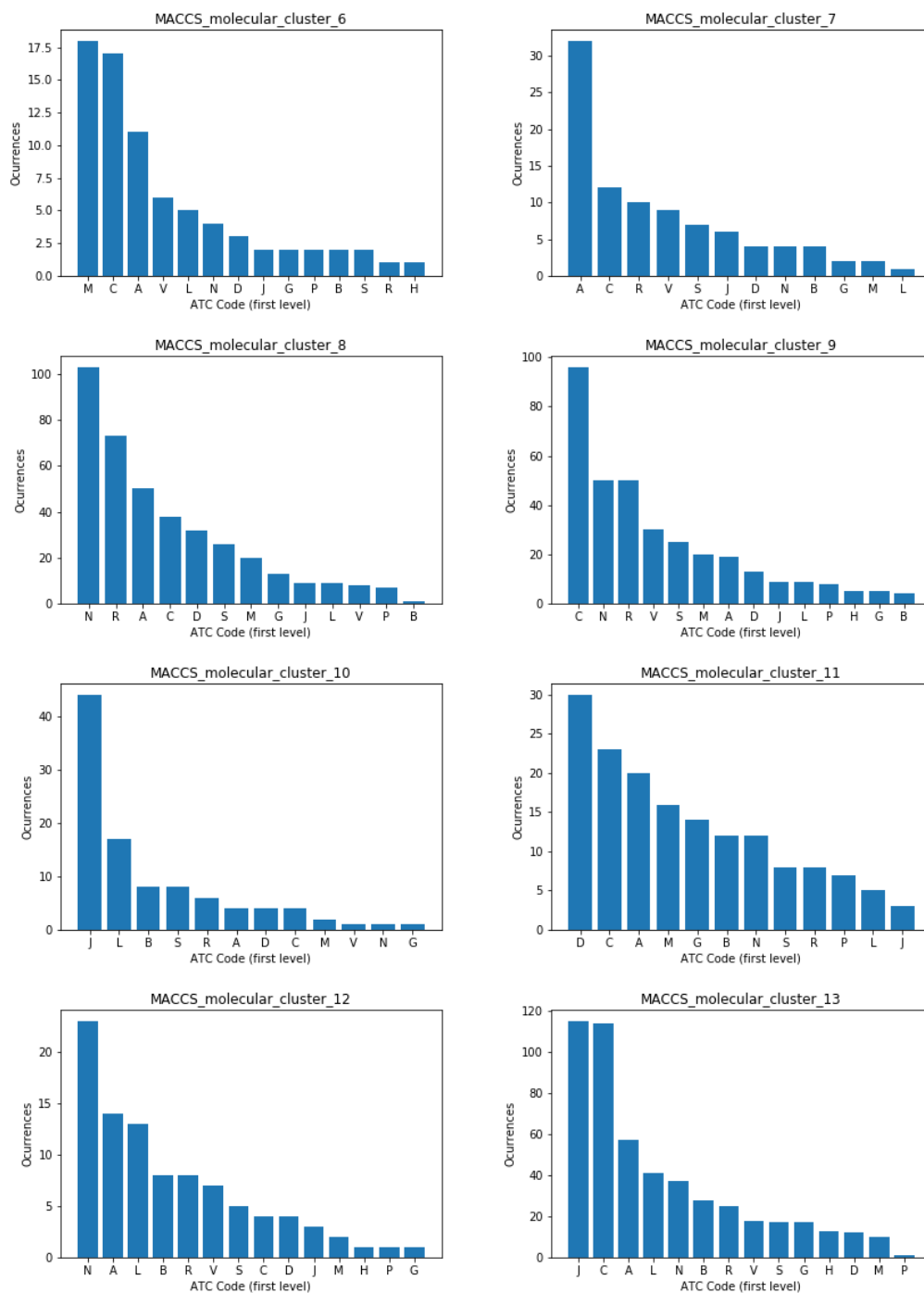


Figure 4.22: Distribution of the first level of all ATC Codes for the drugs contained within the Clusters 6-13 for the molecular structure experiment. The clustering was done using the similarity matrix computed with the MACCS fingerprints.

Conclusion

We have seen that the results of the clustering are quite less useful than for the previous cases. The conclusions we have extracted here are a bit meaningless in order to show a really good performance of the clustering. Even so, our conclusion is that further analysis of the results, with the help of experts, should be done. Because, as it is shown in the evaluation done using the ground truth (Section 4.4.4), the molecular structure based similarity has shown the most promising results. In fact, they are the most used similarity measurements among drugs in the state of the art.

4 4 4 Direct Evaluation: Ground Truth

This evaluation is explained in the Section 3.5.2, please, for detailed information, read that section. Basically, we use 100 pairs of drugs which have been annotated by 143 experts (henceforward, ground truth). They were asked if the two molecules were or not similar. Our aim is to see if our approach actually gives a similar answer to the one provided by those experts. Keeping in mind that aim, we have decided to evaluate our similarity measurements against the ground truth in three different dimensions or aspects:

- **Value of similarity.** Study of the correlation between the value of similarity computed by our measure and the similarity from the ground truth. For this experiment we use Pearson's Correlation Coefficient. Please, note that the value of similarity provided by the ground truth is not a degree of similarity between the drugs, but the percentage (from 0 to 1) of the experts who said that the pair of drugs were similar.
- **Order.** Study of the order or rank inferred by the value of similarity. We have ordered the pairs from the least to the most similar and then studied the rank correlation using Kendall's τ Correlation Coefficient.
- **Threshold.** We have set a threshold in order to classify our pairs of drugs into two classes: similar and non-similar. The threshold is 0.85, because the Tanimoto Coefficient, one of the similarities we use (see Section 3.4) has shown to indicate similarity between two molecules from that value. Once we have classified the pairs using our similarity measure and the ground truth, we compute the *accuracy* and the *recall*.

In this case, we have two different experiments, one for each of the sorts of fingerprints we use: ECFP and MACCS. In the Table 4.3, we can see all relevant information for those experiments, including all the evaluation coefficients explained above. Note that we have also included the number of pairs we have from the ground truth (97) and the pairs which are among our computed similarities (96). Originally, the ground truth is a set of 100 pairs, however, we have not been able to find all the names of some drugs from the original paper [Franco et al., 2014]. The reason is that the authors just published the molecular structure of the pairs, the SMILE representation and the decision of the experts, but not the names of the drugs. Thus, we needed to search for the structure and the SMILE representation on different webs, and we could not find some of them. The file we have used for this evaluation is a .csv file which looks like it is shown in the Figure 3.9.

Even though in this experiment the difference is not so high as before, since we even have less pairs among the computed similarities (96), we just evaluate considering those pairs.

Sort of Fingerprint	ECFP	MACCS
Pairs in ground truth	97	97
Pairs in computed similarity	96	96
Kendall's τ	-0.0404	0.0601
Pearson's Correlation	0.8886	0.9186
Accuracy	0.7708	0.8854
Recall	0.12	0.76

Table 4.3: Direct Evaluation against a ground truth of the Molecular Based Similarity

Even though we have clustered using both cases, ECFP and MACCS, still, based on the results showed in the table, we can say that MACCS seems to be a better measure.

Both experiments show similar values for the two correlations, so that we have determined that MACCS could be a better solution just based on the values of *accuracy* and *recall*. The *accuracy* of both cases is good but in the case of using MACCS is actually quite good (0.8854). Nevertheless, the value of the *recall* is quite better for the case of using MACCS. The measure *recall* gives us an idea of which portion of 'similar' drugs were classified as 'similar'. In the case of ECFP, the values of similarities are quite smaller than in the MACCS. This is something we have already talked about when we have shown the two similarity matrices. This fact comes from the length of both fingerprints, ECFP is quite larger, so it is more difficult for the drugs to be similar (since the representation of them is more specific and unique). ECFP is normally consider even better than MACCS, since it contains more bits and it is more precise. Even so, in this case, since we are using a high threshold, the similarity measure computed with ECFP seems to be worse. We are not going to claim which one is better, we just want to state that even though with ECFP we probably obtain better similarity values (more precise), the absolute values are lower, and this should be taken into account if you want to classify the drugs using a threshold.

As happened in the previous experiments, Kendall's τ Correlation says that there is not correlation at all. However, Person's Correlation is quite good in both cases, so that there is correlation considering the values of the similarities. Actually, as a conclusion, based on the values of Accuracy, and correlations, we could say that our method is relatively good to infer the similarity between a pair of drugs. However, is not really good to infer the degree of similarity between two drugs taking into account how similar are other pairs of drugs. So that, the measure of similarity is good locally, but we cannot say that it is good globally. Even so, it is possible that this fact is not only caused by the quality of our measure. It could be caused because, as said before, the ground truth gives us information about how many experts said that a pair of drugs is similar or not. Nevertheless, the experts were not asked about which degree of similarity have those drugs, neither they were asked to say how similar are two drugs in comparison to another two other ones.

CONCLUSION

The current chapter, is devoted to discuss, on the one hand, the obtained results and draw a final conclusion of the work done along this master thesis. On the other hand, we state the contributions of the project and a set of possible future lines. Note that, when needed, specific conclusions for each experiment have been explained in their respective section along the Chapter 4.

5 1

Statement and Contributions

In this project, three different similarity measurements between drugs from the DrugBank database have been developed. Each of those measures have been computed over one or more dimensions of the drugs: textual, taxonomic and molecular information. In order to study how good is each of the similarity measures, two different evaluations have been performed: indirect and direct. The indirect evaluation is based on clustering the drugs and evaluating how good the obtained clusters are. The direct evaluation is done over the similarities, comparing them with a ground truth provided by experts in the domain. This section is devote to list and explain which are the contributions of this thesis.

- A text similarity measure has been computed over some of the textual fields of the DrugBank database. Even though there exist several works devoted to the implementation of text similarity measures, we do not know other works in which they have used that sort of similarity over DrugBank.
- A taxonomy similarity measure has been computed over the main classification structure provided by DrugBank. This is the only approach we know in which graph similarity has been computed over the taxonomic structure of DrugBank.
- A molecular structure based similarity has been computed over pairs of drugs from the DrugBank database. This is a well known topic in the domain, so it is not a novel work. Nevertheless, our work provides the largest number of pairs of drugs in which the similarity has been already computed and it is ready to be used.

- To summarize the three previous items, we have computed three different similarity measures within the same framework. Usually, other works focus on just one of them, here we provide results for three different ones.
- A qualitative indirect evaluation of the similarities is performed over the performance of clustering drugs using the computed similarities. The results are rather good sometimes and interesting and useful conclusions arose.
- A quantitative direct evaluation is done using a small ground truth. The results provide conclusions which can be useful for the research community. One of the contributions we have done with this evaluation is to actually prepare a document of the ground truth which is ready to be used. The original version provided by the authors could not be interpreted by a computer.
- Several future lines of work are proposed, therefore the research effort done in this thesis had continuity.
- The implemented code and used resources have been uploaded to a open repository in GitHub under a MIT license¹. This fact adds value to our work since all the exposed results can be easily obtained by other researches. Furthermore, this together with the fact that we provide a list of future lines of work, makes more possible to have an extension of our work.

5 2

Conclusions

In this section we draw some conclusions about the results we have obtained in each of the tasks developed within our work (see Chapter 3). Along Chapter 4, we have already give some specific conclusions to the experiments when we have considered it necessary. A more global perspective of the conclusions is provided in the current section.

As it is stated several times along the document, two different evaluations have been performed: indirect (clustering) and direct (ground truth). Conclusions from each of them are extracted and analyzed separately in the upcoming subsections.

5 2 1 Clustering Evaluation: Conclusions

Considering the results exposed within the Chapter 4, we claim that the best results of the clustering are achieved when the similarity based on text mining is used (see Section 3.2). The purity of the clusters, even though it is studied qualitatively, seems to be much better in this case. In fact, we obtain three clusters in which the most predominant ATC Code represents a good percentage of the total number of drugs with that ATC Code. We also get several clusters in which a unique ATC Code is really predominant, so that the clusters can be understood as belonging to that ATC Code. Nevertheless, it is not a perfect result.

¹The MIT License is a permissive free software license originating at the Massachusetts Institute of Technology. As a permissive license, it puts only very limited restriction on reuse and has, therefore, an excellent license compatibility. See more information on: https://en.wikipedia.org/wiki/MIT_License. Last visit: April 2018.

In general, the picture is not totally optimistic, we cannot state that the clustering has provided a prominent result. The performance of the clustering is not good in the cases of using similarity measures based on taxonomic information and molecular structure. The main reason why we obtain this bad result is because the data is clearly unbalanced (see Figures 4.4, 4.11, 4.18). Spectral clustering (algorithm used in this project) and graph-based semi-supervised learning algorithms, in general, are well known to be sensitive to how graphs are constructed from data. In particular if the data has proximal and unbalanced clusters these algorithms can lead to poor performance. On the other hand, we think that since we are using different information of the drugs, the meaning of our similarity measures is different from each other. It is to say, the first level of the ATC Code (14 categories), used to evaluate the clusters, indicates the anatomical main (not only) group in which the drug is supposed to act. However, each of our similarities is based on different characteristics of the drugs (DrugBank taxonomy, textual information and molecular structure). Might be the case that our similarity measures are good to cluster the drugs following another structure, not the ATC Code classification. For instance, the taxonomic information extracted from DrugBank might have nothing to do with the anatomical main group (ATC Code first level).

To conclude, the indirect evaluation based on clustering has lights and shadows. We see some relatively favorable results but it is not enough to claim that our similarities are really good. We have found though, some possible causes of the bad performance of the clustering. Further research in this direction should be done.

5.2.2 Ground Truth Evaluation: Conclusions

In the Chapter 4, we show the results of the evaluation of each similarity against what we call 'ground truth'. The best result has been obtained for the case of Molecular Structure based Similarity. The reason is quite clear, the ground truth was built by asking to 143 experts if the molecular structure of several pairs of drugs were or not similar. This result makes even more evident something we have said implicitly before: each of the similarity measures computed along this thesis has its own meaning and thus, its possible field of application. We have not gone so far to analyze in which applications each of the similarities would be more suitable, but it is clear that each of them has a proper meaning different from the rest.

Just as a reminder, three different studies have been done inside this evaluation: correlation between the values, the order inferred by the similarities and how both, ground truth and our similarities, classify the drugs into similar or not. In the following paragraphs, some specific comments about those three studies are provided.

Correlation between the value

In most of the cases we have seen positive values of correlations which were close to or bigger than 0.7, so we can say there exists correlation. The worst case is found in the text mining experiment. This result can be interpreted as there is a certain correlation between the degree of similarity of our measurements and the percentage of experts who said that a pair of drug was similar. Please, note that we expressed that percentage between 0 and 1.

Correlation of the inferred order

In all the cases, this value is nearly 0, which means that there does not exist correlation at all. The order inferred using the values of similarity we have computed is not correlated to the order inferred using the ground truth. Even though we expected to find correlation, the bad result might be reasonable if we think about how the ground truth was built. The experts were not asked about which degree of similarity have those drugs, neither they were asked to say how similar are two drugs in comparison to another two other ones.

Result of the classification

The best performance in this case is found in the Molecular Structure based Similarity. Reasonable, since we used 0.85 as the threshold to decide if a pair of drugs was similar or not. It is exactly the value which is recommended to use when the Tanimoto (Jaccard) Coefficient (basis of the molecular structure based similarity experiment) is used.

As a general comment, it would be really necessary to extend the ground truth in order to give better conclusions. However, we have not found a larger option so this is the best we can do.

5 3

Future Work

The heterogeneity of the work developed in this project is doubtless. Therefore, several future lines of work have been opened with this thesis. In this section, we list and explain briefly some of those possible lines, of course, the list could be larger and since we have shared all our work (including code) we are open to possible contributions with other groups.

- Combine the result provided by each of the similarity measures. As we have shown, any of them gives an irrefutable result, thus, a good approach would be to combine them, linearly, for instance. This combination could be weighted or not, therefore, choosing the correct weights is another possible future work.
- Improvements in the text based similarity:
 - Just three textual fields from DrugBank have been used, we could use others or even detect which are the most useful for the task and weight the relevance of each of them. An easy way of weighting would be to just divide the *tf* value of the *tf-idf* by the length of the field. Thus, we were doing a Bag of Words not in the total but in each textual field.
 - Another possible improvement would be to use the name of the drug. As we explained in the first chapters, the name of a drug contains prefixes and suffixes which sometimes are related to the type of drug.
 - Finally, we could use Latent Dirichlet Allocation (LDA) instead of LSA. LSI or LSA learn latent topics by performing a matrix decomposition (SVD) on the term-document matrix. LDA is a generative probabilistic model, that assumes a Dirichlet prior over the latent topics. In practice, LSI is much faster to train than LDA, but has lower accuracy.

- Improvements in the taxonomy based similarity can be done in choosing the weights of the weighted graph. It would be good also to look for more sophisticated ways of computing the distances/similarities between the drugs.
- Improvements in the molecular structure based similarity are, for instance, using 3D representations of the drugs, not only 2D fingerprints.
- Share and evaluate our results with experts in the domain in order to get better conclusions.
- One of our ideas was to program a parametric function to compute similarity between two list of drugs. Possible parameters of that function could be: type or type of similarities, sorts of coefficients, used fields in the text similarity measure, fingerprints used in the molecular structure measure, etc. This task would be just to put together all the pieces of our code.
- Build an API for users without specific programming skills. Once the similarities have been tested more than in this project, it would be interesting to build an API for people like doctors or chemists, etc.
- Enlarge the ground truth.
- Apply the implemented similarities to other medical entities (e.g. body parts, illnesses, etc.). Of course, the similarity based on the molecular structure of the drugs cannot be used in much more other cases. However, we could use the textual and taxonomic similarities over ontologies and other resources about body parts or illnesses. The potential of having similarity measures among those three medical entities is inestimable. We could know if a drug could be used to treat a specific illness in a specific body part, for instance, if that drug has been used to treat the same specific illness in another part of the body which is similar to the target one.

BIBLIOGRAPHY

- [Arandjelović, 2015] Arandjelović, O. (2015). Clinical trial adaptation by matching evidence in complementary patient sub-groups of auxiliary blinding questionnaire responses. *PloS one*, 10(7):e0131524.
- [Arandjelović, 2017] Arandjelović, O. (2017). Strategies for informed sample size reduction in adaptive controlled clinical trials. *EURASIP Journal on Advances in Signal Processing*, 2017(1):75.
- [Bajusz et al., 2015] Bajusz, D., Rácz, A., and Héberger, K. (2015). Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7(1):20.
- [Bellman, 2015] Bellman, R. E. (2015). *Adaptive control processes: a guided tour*. Princeton university press.
- [Benik et al., 2012a] Benik, J., Chang, C., Raschid, L., Vidal, M.-E., Palma, G., and Thor, A. (2012a). Finding cross genome patterns in annotation graphs. In *International Conference on Data Integration in the Life Sciences*, pages 21–36. Springer.
- [Benik et al., 2012b] Benik, J., Palma, G., Raschid, L., Thor, A., and Vidal, M.-E. (2012b). Mining patterns from clinical trial annotated datasets by exploiting the nci thesaurus. In *Proceedings of the 2012th International Conference on Posters & Demonstrations Track-Volume 914*, pages 1–4. CEUR-WS. org.
- [Beykikhoshk et al., 2015] Beykikhoshk, A., Arandjelović, O., Phung, D., Venkatesh, S., and Caelli, T. (2015). Using twitter to learn about the autism community. *Social Network Analysis and Mining*, 5(1):22.
- [Bodenreider, 2004] Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.
- [Bolton et al., 2008] Bolton, E. E., Wang, Y., Thiessen, P. A., and Bryant, S. H. (2008). Pubchem: integrated platform of small molecules and biological activities. In *Annual reports in computational chemistry*, volume 4, pages 217–241. Elsevier.
- [Bowman et al., 2017] Bowman, S., Goldberg, Y., Hill, F., Lazaridou, A., Levy, O., Reichart, R., and Søgaard, A. (2017). *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*. Association for Computational Linguistics.

- [Campos et al., 2017] Campos, L., Pedro, V., and Couto, F. (2017). Impact of translation on named-entity recognition in radiology texts. *Database*, 2017.
- [Cereto-Massagué et al., 2015] Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., and Pujadas, G. (2015). Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63.
- [Cheng et al., 2004] Cheng, J., Cline, M., Martin, J., Finkelstein, D., Awad, T., Kulp, D., and Siani-Rose, M. A. (2004). A knowledge-based clustering algorithm driven by gene ontology. *Journal of biopharmaceutical statistics*, 14(3):687–700.
- [Deerwester et al., 1990] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.
- [Del Pozo et al., 2008] Del Pozo, A., Pazos, F., and Valencia, A. (2008). Defining functional distances over gene ontology. *BMC bioinformatics*, 9(1):50.
- [Demner-Fushman et al., 2009] Demner-Fushman, D., Chapman, W. W., and McDonald, C. J. (2009). What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, 42(5):760–772.
- [Donnelly, 2006] Donnelly, K. (2006). Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279.
- [Dumais et al., 1995] Dumais, S. T. et al. (1995). Latent semantic indexing (lsi): Trec-3 report. *Nist Special Publication SP*, pages 219–219.
- [Dunn et al., 2018] Dunn, A. G., Coiera, E., and Bourgeois, F. T. (2018). Unreported links between trial registrations and published articles were identified using document similarity measures in a cross-sectional analysis of clinicaltrials. gov. *Journal of clinical epidemiology*, 95:94–101.
- [Franco et al., 2014] Franco, P., Porta, N., Holliday, J. D., and Willett, P. (2014). The use of 2d fingerprint methods to support the assessment of structural similarity in orphan drug legislation. *Journal of Cheminformatics*, 6(1):5.
- [Galkin et al., 2017] Galkin, M., Collarana, D., Traverso-Ribón, I., Vidal, M.-E., and Auer, S. (2017). Sjoin: A semantic join operator to integrate heterogeneous rdf graphs. In *International Conference on Database and Expert Systems Applications*, pages 206–221. Springer.
- [Gaulton et al., 2011] Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., et al. (2011). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107.
- [Goeriot et al., 2015] Goeriot, L., Kelly, L., Suominen, H., Hanlen, L., Névél, A., Grouin, C., Palotti, J., and Zuccon, G. (2015). Overview of the clef ehealth evaluation lab 2015. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 429–443. Springer.

- [Golub and Reinsch, 1970] Golub, G. H. and Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numerische mathematik*, 14(5):403–420.
- [Goodwin and Harabagiu, 2016] Goodwin, T. R. and Harabagiu, S. M. (2016). Medical question answering for clinical decision support. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 297–306. ACM.
- [Goodwin and Harabagiu, 2017] Goodwin, T. R. and Harabagiu, S. M. (2017). Knowledge representations and inference techniques for medical question answering. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(2):14.
- [Gower, 1971] Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871.
- [Grover and Leskovec, 2016] Grover, A. and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM.
- [Hauser et al., 2017] Hauser, A. S., Attwood, M. M., Rask-Andersen, M., Schiöth, H. B., and Gloriam, D. E. (2017). Trends in gpcr drug discovery: new agents, targets and indications. *Nature Reviews Drug Discovery*, 16(12):829.
- [Head-ingB, 1965] Head-ingB, M.-i. S. (1965). Medical subject headings. *Nature*, 20.
- [Ho et al., 2018] Ho, P. H., Nguyen, N. A. T., and Vo, T. H. (2018). Dna sequences representation derived from discrete wavelet transformation for text similarity recognition. In *Modern Approaches for Intelligent Information and Database Systems*, pages 75–85. Springer.
- [Johnson and Maggiora, 1990] Johnson, M. A. and Maggiora, G. M. (1990). *Concepts and applications of molecular similarity*. Wiley.
- [Kashyap et al., 2016] Kashyap, A., Han, L., Yus, R., Sleeman, J., Satyapanich, T., Gandhi, S., and Finin, T. (2016). Robust semantic text similarity using lsa, machine learning, and linguistic resources. *Language Resources and Evaluation*, 50(1):125–161.
- [Kenter and De Rijke, 2015] Kenter, T. and De Rijke, M. (2015). Short text similarity with word embeddings. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 1411–1420. ACM.
- [Keys, 2011] Keys, M. S. (2011). Accelrys: San diego. *There is no corresponding record for this reference*.
- [Knox et al., 2010] Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., et al. (2010). Drugbank 3.0: a comprehensive resource for ‘omics’ research on drugs. *Nucleic acids research*, 39(suppl_1):D1035–D1041.
- [Krallinger et al., 2015] Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., Leaman, R., Lu, Y., Ji, D., Lowe, D. M., et al. (2015). The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(S1):S2.

- [Kuhn et al., 2007] Kuhn, M., von Mering, C., Campillos, M., Jensen, L. J., and Bork, P. (2007). Stitch: interaction networks of chemicals and proteins. *Nucleic acids research*, 36(suppl_1):D684–D688.
- [Lakhani et al., 2018] Lakhani, P., Prater, A. B., Hutson, R. K., Andriole, K. P., Dreyer, K. J., Morey, J., Prevedello, L. M., Clark, T. J., Geis, J. R., Itri, J. N., et al. (2018). Machine learning in radiology: applications beyond image interpretation. *Journal of the American College of Radiology*, 15(2):350–359.
- [Landrum et al., 2006] Landrum, G. et al. (2006). Rdkit: Open-source cheminformatics.
- [Leacock and Chodorow, 1998] Leacock, C. and Chodorow, M. (1998). Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283.
- [Lin et al., 1998] Lin, D. et al. (1998). An information-theoretic definition of similarity. In *Icml*, volume 98, pages 296–304. Citeseer.
- [Lipscomb, 2000] Lipscomb, C. E. (2000). Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.
- [Liu and Zhao, 2016] Liu, Y. and Zhao, H. (2016). Predicting synergistic effects between compounds through their structural similarity and effects on transcriptomes. *Bioinformatics*, 32(24):3782–3789.
- [MacQueen et al., 1967] MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.
- [McNaught, 2006] McNaught, A. (2006). The iupac international chemical identifier. *Chemistry international*, pages 12–14.
- [Melnikov and Vorobkalov, 2014] Melnikov, M. P. and Vorobkalov, P. N. (2014). Retrieval of drug-drug interactions information from biomedical texts: use of tf-idf for classification. In *Joint Conference on Knowledge-Based Software Engineering*, pages 593–602. Springer.
- [Milian et al., 2013] Milian, K., Bucur, A., van Harmelen, F., ten Teije, A., et al. (2013). Identifying most relevant concepts to describe clinical trial eligibility criteria.
- [Moffat et al., 2017] Moffat, J. G., Vincent, F., Lee, J. A., Eder, J., and Prunotto, M. (2017). Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nature Reviews Drug Discovery*, 16(8):531.
- [Moon, 1996] Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60.
- [Muegge and Mukherjee, 2016] Muegge, I. and Mukherjee, P. (2016). An overview of molecular fingerprint similarity search in virtual screening. *Expert opinion on drug discovery*, 11(2):137–148.

- [Nikfarjam et al., 2015] Nikfarjam, A., Sarker, A., O'Connor, K., Ginn, R., and Gonzalez, G. (2015). Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681.
- [Nikolova and Jaworska, 2003] Nikolova, N. and Jaworska, J. (2003). Approaches to measure chemical similarity—a review. *Molecular Informatics*, 22(9-10):1006–1026.
- [Nowotka et al., 2014] Nowotka, M., Davies, M., Papadatos, G., and Overington, J. P. (2014). ChEMBL beaker: a lightweight web framework providing robust and extensible cheminformatics services. *Challenges*, 5(2):444–449.
- [Ochoa et al., 2013] Ochoa, R., Davies, M., Papadatos, G., Atkinson, F., and Overington, J. P. (2013). mychembl: a virtual machine implementation of open data and cheminformatics tools. *Bioinformatics*, 30(2):298–300.
- [Overington et al., 2006] Overington, J. P., Al-Lazikani, B., and Hopkins, A. L. (2006). How many drug targets are there? *Nature reviews Drug discovery*, 5(12):993.
- [Palma et al., 2014] Palma, G., Vidal, M.-E., and Raschid, L. (2014). Drug-target interaction prediction using semantic similarity and edge partitioning. In *International Semantic Web Conference*, pages 131–146. Springer.
- [Pekar and Staab, 2002] Pekar, V. and Staab, S. (2002). Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.
- [Pierce et al., 2017] Pierce, C. E., Bouri, K., Pamer, C., Proestel, S., Rodriguez, H. W., Van Le, H., Freifeld, C. C., Brownstein, J. S., Walderhaug, M., Edwards, I. R., et al. (2017). Evaluation of facebook and twitter monitoring to detect safety signals for medical products: an analysis of recent fda safety alerts. *Drug safety*, 40(4):317–331.
- [Pons et al., 2016] Pons, E., Braun, L. M., Hunink, M. M., and Kors, J. A. (2016). Natural language processing in radiology: a systematic review. *Radiology*, 279(2):329–343.
- [Ran et al., 2017] Ran, Y., He, B., Hui, K., Xu, J., and Sun, L. (2017). A document-based neural relevance model for effective clinical decision support. In *Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on*, pages 798–804. IEEE.
- [Resnik, 1995] Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.
- [Roberts et al., 2016] Roberts, K., Simpson, M., Demner-Fushman, D., Voorhees, E., and Hersh, W. (2016). State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the trec 2014 cds track. *Information Retrieval Journal*, 19(1-2):113–148.
- [Roberts et al., 2015] Roberts, K., Simpson, M. S., Voorhees, E. M., and Hersh, W. R. (2015). Overview of the trec 2015 clinical decision support track. In *TREC*.

- [Rogers and Hahn, 2010] Rogers, D. and Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754.
- [Rosse and Mejino, 2008] Rosse, C. and Mejino, J. L. (2008). The foundational model of anatomy ontology. In *Anatomy Ontologies for Bioinformatics*, pages 59–117. Springer.
- [Saha et al., 2010] Saha, B., Hoch, A., Khuller, S., Raschid, L., and Zhang, X.-N. (2010). Dense subgraphs with restrictions and applications to gene annotation graphs. In *Annual International Conference on Research in Computational Molecular Biology*, pages 456–472. Springer.
- [Salvadores et al., 2012] Salvadores, M., Horridge, M., Alexander, P. R., Fergerson, R. W., Musen, M. A., and Noy, N. F. (2012). Using sparql to query bioportal ontologies and metadata. In *International Semantic Web Conference*, pages 180–195. Springer.
- [Santos et al., 2017] Santos, R., Ursu, O., Gaulton, A., Bento, A. P., Donadi, R. S., Bologa, C. G., Karlsson, A., Al-Lazikani, B., Hersey, A., Oprea, T. I., et al. (2017). A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery*, 16(1):19.
- [Segura-Bedmar et al., 2014] Segura-Bedmar, I., Martínez, P., and Herrero-Zazo, M. (2014). Lessons learnt from the ddiextraction-2013 shared task. *Journal of biomedical informatics*, 51:152–164.
- [Singh et al., 2017] Singh, K., Mulang, I. O., Lytra, I., Jaradeh, M. Y., Sakor, A., Vidal, M.-E., Lange, C., and Auer, S. (2017). Capturing knowledge in semantically-typed relational patterns to enhance relation linking. In *Proceedings of the Knowledge Capture Conference*, page 31. ACM.
- [Takeda et al., 2017] Takeda, T., Hao, M., Cheng, T., Bryant, S. H., and Wang, Y. (2017). Predicting drug–drug interactions through drug structural similarities and interaction networks incorporating pharmacokinetics and pharmacodynamics knowledge. *Journal of Cheminformatics*, 9(1):16.
- [Vasiljeva and Arandelovic, 2016] Vasiljeva, I. and Arandelovic, O. (2016). Automatic knowledge extraction from ehers. In *IJCAI 2016-Workshop on Knowledge Discovery in Healthcare Data*.
- [Vasiljeva and Arandjelović, 2017] Vasiljeva, I. and Arandjelović, O. (2017). Diagnosis prediction from electronic health records using the binary diagnosis history vector representation. *Journal of Computational Biology*, 24(8):767–786.
- [Vilar and Hripcsak, 2016] Vilar, S. and Hripcsak, G. (2016). The role of drug profiles as similarity metrics: applications to repurposing, adverse effects detection and drug–drug interactions. *Briefings in bioinformatics*, 18(4):670–681.
- [Vincze et al., 2008] Vincze, V., Szarvas, G., Farkas, R., Móra, G., and Csirik, J. (2008). The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, 9(11):S9.

- [Vivaldi and Rodríguez, 2015] Vivaldi, J. and Rodríguez, H. (2015). Medical entities tagging using distant learning. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 631–642. Springer.
- [Wang et al., 2016] Wang, B., Yu, X., Wei, R., Yuan, C., Li, X., and Zheng, C.-H. (2016). System prediction of drug-drug interactions through the integration of drug phenotypic, therapeutic, structural, and genomic similarities. In *International Conference on Intelligent Computing*, pages 377–385. Springer.
- [Weininger, 1988] Weininger, D. (1988). Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36.
- [Whetzel et al., 2011] Whetzel, P. L., Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C., Tudorache, T., and Musen, M. A. (2011). Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic acids research*, 39(suppl_2):W541–W545.
- [Willett, 2014] Willett, P. (2014). The calculation of molecular structural similarity: principles and practice. *Molecular informatics*, 33(6-7):403–413.
- [Wishart et al., 2017] Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2017). Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082.
- [Wishart et al., 2006] Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006). Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 34(suppl_1):D668–D672.
- [Yeganova et al., 2012] Yeganova, L., Kim, W., Comeau, D. C., and Wilbur, W. J. (2012). Finding biomedical categories in medline®. *Journal of biomedical semantics*, 3(3):S3.
- [Yi et al., 2017] Yi, Z., Li, S., Yu, J., Tan, Y., Wu, Q., Yuan, H., and Wang, T. (2017). Drug-drug interaction extraction via recurrent neural network with multiple attention layers. In *International Conference on Advanced Data Mining and Applications*, pages 554–566. Springer.
- [Yu et al., 2015] Yu, X., Geer, L. Y., Han, L., and Bryant, S. H. (2015). Target enhanced 2d similarity search by using explicit biological activity annotations and profiles. *Journal of cheminformatics*, 7(1):55.
- [Zesch and Gurevych, 2010] Zesch, T. and Gurevych, I. (2010). Wisdom of crowds versus wisdom of linguists—measuring the semantic relatedness of words. *Natural Language Engineering*, 16(1):25–59.
- [Zhang et al., 2017] Zhang, S., Zhang, X., Wang, H., Cheng, J., Li, P., and Ding, Z. (2017). Chinese medical question answer matching using end-to-end character-level multi-scale cnns. *Applied Sciences*, 7(8):767.