The Strengths and Pitfalls of Large-Scale Text Mining for Literary Studies

Tahmasebi, Nina

2020-03-20

Tahmasebi , N & Hengchen , S 2020 , ' The Strengths and Pitfalls of Large-Scale Text Mining for Literary Studies ' , Samlaren , vol. 140 (2019) , pp. 198-227 . < http://svelitt.se/samlaren/2019 >

http://hdl.handle.net/10138/314258

Samlaren

Tidskrift för forskning om

svensk och annan nordisk litteratur

Årgång 140 2019

I distribution:
Eddy.se

Svenska Litteratursällskapet

Inlagans typografi: Anders Svedin

Utgiven med stöd av Vetenskapsrådet

Bidrag till *Samlaren* insändes digitalt i ordbehandlingsprogrammet Word till info@svelitt.se. Konsultera skribentinstruktionerna på sällskapets hemsida innan du skickar in. Sista inlämningsdatum för uppsatser till nästa årgång av *Samlaren* är 15 juni 2020 och för recensioner 1 september 2020. *Samlaren* publiceras även digitalt, varför den som sänder in material till *Samlaren* därmed anses medge digital publicering. Den digitala utgåvan nås på: http://www.svelitt.se/samlaren/index.html. Sällskapet avser att kontinuerligt tillgängliggöra även äldre årgångar av tidskriften.

Svenska Litteratursällskapet tackar de personer som under det senaste året ställt sig till förfogande som bedömare av inkomna manuskript.

Svenska Litteratursällskapet PG: 5367–8.

Svenska Litteratursällskapets hemsida kan nås via adressen www.svelitt.se.

ISBN 978–91–87666–39–1
ISSN 0348–6133

Printed in Lithuania by
Balto print, Vilnius 2019

The Strengths and Pitfalls of Large-Scale Text Mining for Literary Studies

By NINA TAHMASEBI & SIMON HENGCHEN

Introduction

Computational literary studies, an integral part of the text-based digital humanities, lie at the intersection of several research fields. Often, the methods come from computer science, language technology, and other fields, while the research questions stem from within the field of literature. In these disciplines many of the foundations are different: the view of data, accepted research methodologies, the understanding of results, the validity of results, and evaluation, all differ greatly.

A truly fruitful study within the scope of computational literary studies requires knowledge of these differences and the challenges that each brings with it. Often researchers who attempt to use digital data and methods for answering humanities research questions are less aware of the mathematical foundations of data science and the technicalities of the methods used. Researchers from the technical sciences, on the other hand, have less in-depth research questions, and typically target research questions that are close to the current technical capacity of the data and methods employed.

We strongly believe that an iterative process that allows information, methods, and research questions to be exchanged among all participating fields has the largest potential for significant contributions.

In this paper, we will take a data science and language technology viewpoint and try to outline the general processes required to extract information from large-scale digital literary text, in both a descriptive and prescriptive manner.[1] While doing so, we will try to highlight important aspects and pitfalls to be wary of. While this paper is not the first of its kind,[2] we particularly aim to contribute to a methodological discussion around the joining of digital methods and data to answer research questions outside of the technical fields. And, though the examples are mostly taken from literary studies, we believe that the recommendations and discussions are equally relevant to other computational text-based humanities.

Our view of a data-intensive research process starts with the digital text and moves through a *natural language processing* (NLP) pipeline toward results, the evaluation of results, and their relation to research questions. This paper will follow the same outline.

The data-intensive research process

Systematic data-intensive research typically has a clear process and several important components. There is the data, a text mining method, and results. Motivating this are research questions and hypotheses. In the process of data-intensive research, there are two main methods for making use of large-scale text. First, it can be used in an *exploratory fashion* to find and formulate interesting hypotheses; the work starts from a general research question without a priori hypotheses. Alternatively, the research can start with a well-defined hypothesis and employ large-scale text to find evidence to support or reject the hypothesis in a *validating fashion*.[3]

Figure 1 illustrates the process schematically. Both the exploratory and the validation paths follow the same process, but with different starting points. The exploratory path moves counterclockwise, from the research question via data and text mining methods, resulting in concrete hypotheses. The validation path starts with one or several clearly defined hypotheses; after choosing the path and the research questions, the data and most suitable methods can be chosen. The exploratory path primarily aims to discover patterns, while the validation path primarily aims at demonstrating or proving patterns.

A text mining method is employed to generate results from the text in both the exploratory and the validating paths.

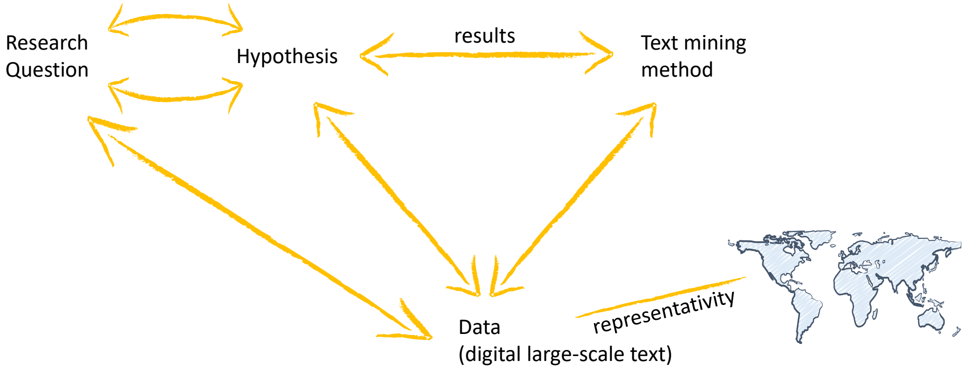

Figure 1: A schematic model of the research process in data-intensive humanities.

Digital text as a resource

As digital text we consider any part of text that contains running text, that is, sequences of words that contribute to the intent of the text, written by an author.[4] Data-intensive research can only extract information that is present in the data; therefore, what we include in the definition of text will form the basis for what questions we can answer. Some research questions investigate the properties of the text, like rhetorical styles, presence of a single or multiple protagonist(s), or comparisons between texts. Other research questions aim to study the world using the text as a proxy. For example, the questions may concern the effects of technology, new methods for communication, or standings of certain societal segments.[5] For such research questions, it is important to consider representativity. Many socioeconomic factors play a role in both modern and historical texts and influence the biases present in the text. If only middle-aged, religious white men are allowed to publish, these texts will inherently be biased by the beliefs, culture and content of this societal segment. When using text to study cultural or social factors, it is therefore important to remember not only who is present in the text (and who is not),[6] but also who generated the text. Selecting a particular text imposes the first reduction and we need to consider how the chosen texts relate to the "whole". Are there specific types of text, genres, or authors that are over- or underrepresented?

Digital text is based on existing, written text that was either born digital or has been digitized. Digitizing the text creates a model of the original text and imposes a second reduction of the text. Depending on the quality of the digitization, more or less of the textual information remains intact.[7]

Regardless of how we arrived at the digital text that we consider the basis of our investigations, the way we view this text will impose further limitations and open different possibilities. The following three main views are often encountered in the literature:

TEXT AS DATA

We begin with the first extreme, where text is seen as *data* no different from say, traffic data. Words in a sentence are seen as being a lane of cars, each word (or car) is observed, modeled and analyzed. The modeling can take one of many forms, but common models are vectors, characters, unique strings, or nodes in a graph.

The basic assumption is often that while the words, like cars in a lane, do affect each other, the order is not important.[8] The popular *bag-of-words model* is used; all words in a window, sentence, paragraph, or document are thrown in a bag and considered without regard to internal order. The model does not distinguish between "Am I happy" and "I am happy", even though the difference in meaning between the two is great.

This view neither needs nor desires heavy linguistic processing that requires knowledge of known properties of the language in the text. Often, some shallow processing like lower-casing, stopword removal,[9] and—sometimes—word filtering based on part of speech, are enough. A benefit of this kind of view is therefore that it does not suffer from the limitations of pre-existing tools and methods and is typically robust when applied to huge quantities of text.[10]

TEXT AS LANGUAGE

In this view, text is seen as a representation of (a certain use of) language. While this point may seem very trivial, it adds enormous strength as well as infinite complexity. If we treat the text as language, we can make use of the wealth of knowledge already available to us, through linguistics for example.

We know how words behave and interact and that the order of words is important. If we return to the traffic metaphor, this view is equivalent to observing not only the car but also the reason why the car is there, which corresponds to *meaning*. We can surmise that if you are traveling to work, you will likely return after eight to ten hours. The equivalent in language is *morphology*, regularities in our language that we can use. We might care about the mood of the driver, which corresponds to *sentiment*. We also know that you affect or are affected by other cars in your lane, which corresponds to *syntax*.

Having all this information gives us great benefits when interpreting and extracting information. In a sentence like *The view is nice but the quality is terrible; we won't be coming back to this hotel!* we can draw conclusions about the writer having a positive sentiment about the view but a negative sentiment about the quality, and thus the hotel. We know who is not coming back and that the hotel is what they are avoiding, not the view.

In general, natural language can be seen as infinitely complex. Even for humans, interpretation can be hard. Present any text to multiple readers and they will interpret the information in the text differently. Ask three people to define the meaning of a word, and the answer will be different for all three. Likely, the answer will be different if the same people are asked again a couple of years later.

For computers, the task is even more difficult. A computer does not have the background knowledge that can be expected in a normal conversation. Without having access to this extra information, interpretation of text becomes extremely hard.[11] Despite this, current machine learning (ML) and artificial intelligence (AI) systems are quite good at handling common scenarios in limited domains. For example, IBM's Watson competed and won against human players in Jeopardy![12] Any search engine can search and match information that does not require a deep understanding of the text. So, while computers can solve many text-based tasks fairly well, they rarely see and interpret text beyond its word-by-word, sentence-by-sentence meaning. They cannot interpret the suggestive, erotic, or provocative. They are unlikely to know how controversial a viewpoint is, or what its implications are. And extremely few can cross-reference previous ideas and knowledge packaged in different ways. They cannot, currently, do what is at the core of humanities research.

TEXT IN THE HUMANITIES

The third view, often taken in the humanities, is where the text is seen as a carrier of linked information, representing our culture, our identity, and our society. In addition to what information each sentence carries within it, we are interested in how this information is linked across sources, authors and times, and how it relates to the world and our knowledge of it.

The complexity inherent in the humanities perspective can best be described as an uncountable infinity. There are so many research questions to ask and so much information buried in each text (if we see beyond the words), that infinity is present in each gap, no matter how small. Therefore, combining large-scale text mining with humanities research questions—i.e. *distant reading*, to use Franco Moretti's well-known concept[13]—has enormous potential.

The processing pipeline

In text mining, we generally do not process an individual document but rather a collection of documents.[14] Our starting point is thus a group of individual documents that are either thematically chosen (all books by a certain author, or books related by topic), opportunistically chosen (anything we can get from a period of interest), or otherwise compiled. This group of documents is referred to as a *data set*, or sometimes as a *corpus*.[15]

Large volumes of text cannot be studied by taking all aspects and words into account. In text mining, the generalization needed is done by focusing on certain aspects of a text, certain parts, or both. In order to find the parts that are of interest, a NLP pipeline is often used.

An NLP pipeline for large amounts of text can include an arbitrary number of processing steps depending on the view of the text. Typically a selection of steps is used.[16] This overview serves to give a general understanding of how masses of text are transformed into information on which we can base our conclusions.

1. *Cleaning*. Firstly, documents are cleaned and tokenized. The cleaning may involve multiple steps. Cleaning typically makes all characters lowercase, adds spaces between words and punctuation marks, and removes additional non-digit or -letter characters. Once cleaning has been done, the text is tokenized to recognize individual tokens and sentences.[17]
2. *Stopword removal*. Very frequent but information-low words like *a*, *an*, *the*, *but* and *in* are removed, both to reduce the size of the data set and to remove noise.
3. *Normalization* of words.
   a) *Lemmatization*. Each word is transformed to its base form (i.e., dictionary form), plurals are turned into singulars, and any other inflections are removed. This step is typically performed to merge all information about a word instead of keeping all inflectional forms (*run*, *running*, *ran* etc. are replaced by *run*).
   b) *Stemming.* For some applications, like information retrieval (search), stemming is preferred to lemmatization. In stemming, only the stem of the word is kept. With stemming, for example, *runner* and *runners* are replaced by *run*—it becomes impossible to distinguish the verb *run* from the noun *runner*.
4. *Part-of-speech tagging*. A word's part of speech is determined for subsequent steps or filtering. Typically only nouns and verbs are kept.
5. *Dependency parsing*. The syntactic information in each sentence is parsed so that the relation between words is recognized. In the sentence *I like the view but not the room*, we can determine that the negation of *like* relates only to the *room*.
6. *Role labeling.* Semantic roles, that is, *who* does *what* in a sentence, are determined. Here, part of speech information is needed to specify, for example, that a person cannot live in another person, but only in a location.
7. *Co-reference resolution*. Co-references and anaphora are resolved so that we know that *she* and *her* refer to *Lyra* in the sentences: "Please be kind to Lyra for as long as she lives. I love her more than anyone has ever been loved."[18]
8. *Target words*. Few studies in data-intensive research make use of all words. Typically, infrequent words are filtered and only the *K* most frequent words are kept (stopwords excluded). *K* typically ranges from 10,000 to 100,000 (unique) words. In addition, parts of speech can be filtered.
9. *Context*. Most text mining relies on the distributional hypothesis; words that occur in the same contexts tend to have similar meanings.[19] This means that words in close proximity contribute to the understanding of a target word. How proximity is defined differs, but typically a context window is used. This window can be of arbitrary size and be defined as full sentence, paragraph or docu-

ment. More commonly, it is defined as *N* words around a target word. For example, for Google N-gram data, five-grams are the maximum, which means five words in a row; typically the target word is chosen as the word in the middle.[20]

10. *Representation of words*. For some methods, words are used as they are. Topic models, for example, work on words in their plain text form. For other applications, words are first represented in different ways such as vectors—one-row tables—or graphs. Sometimes, these representations are learned as a part of the text mining method; other times they are learned in a separate step. In some cases they can be pre-trained on other corpora and reused.[21]
11. *Text mining method.* Chosen depending on the type of analysis.
12. *Comparison*. Often the outcome from step 10 is compared *over time*.

Steps 5–7 are more commonly used in computational linguistics studies because they rely on heavier processing, take more time to execute, and have varying quality on kinds of texts for which they have not been trained, like historical text, social media text, etc. Without steps 5–7, we get what is more commonly known as *shallow NLP*.

The last two steps involve the text mining and data science components, and while the first steps are very generic, the last two steps determine what tasks can be targeted. They also typically determine which choices are made for the previous steps. For example, if a text mining method for word sense induction has been tested on nouns, the method is not guaranteed to work on verbs.

An important aspect of a data-intensive research methodology is the choice of text mining method.[22] It is as important to choose the right text mining method as it is to choose, for example, a means of transportation. In the end, we need to choose the most suitable means of transportation based on how many people are traveling, how far they are going, and in what terrain. The choice of method for text mining should be made with the same considerations in mind: the data at hand and our research question. It is, for example, very problematic to use current word embedding techniques on small data sets, or to answer research questions where we need to know exactly which sentences contributed to the results.[23] Hence, in such situations, we need to choose different methods.

Results—the output of text mining

In the data-intensive research process, the output of text mining is what we consider a *result*. This output can be anything from a part of speech, a sentiment, a number, a topic, or a vector to a true or false statement. It can also be an extract of text cleverly collected to match an information need of some kind. Regardless of form, results convey different kinds of information that we need to interpret and put in context.

DIFFERENT KINDS OF INFORMATION

The term text mining comes from its resemblance to mineral mining; using a data set of text, we refine the data until we arrive at the information we are looking for.[24] When we are done, only part of the original remains, sometimes in the form of direct extracts of text and at other times in other forms of information. In general, we can divide the different kinds of information into *primary* and *aggregated* information.

Primary information is the kind that is written out in clear text in the original document collection, and once found, can easily be confirmed by a human. The output of search engines is of the primary kind: given a search query, the engine returns a set of resources that provide the information needed.[25] The results are unaltered and the user can easily confirm their correctness by going through each resource.[26]

Semantic roles can also convey primary information. Here is a small example, borrowed from Alexandre Dumas's 1844 *The Count of Monte Cristo*. As those familiar with the novel will know, the main character Edmond Dantès is born in 1796, the son of Louis Dantès. He is also engaged to Mercédès Herrera, who later on will marry Fernand Mondego. Dantès is introduced to the readers as the first mate of the ship *Le Pharaon*.

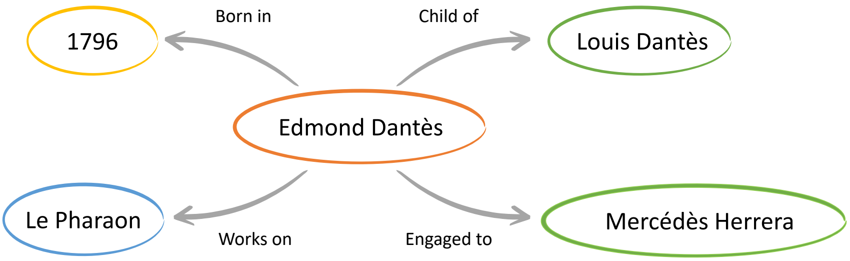

Figure 2: A graphical representation of some of the semantic roles found in Alexandre Dumas's The Count of Monte Cristo.

This information can be represented relationally in a graph, as in Figure 2. The graph is a mere alteration of the form in which the information was presented in the original text: the information is presented differently, but the information does not change. Obviously, more complex versions of such graphs are also possible—the attentive reader will remark that Edmond Dantès is also referred to (sometimes by himself, so as to hide his true identity) in the novel as Sinbad the Sailor, the Count of Monte Cristo, Lord Wilmore, Abbé Busoni, Monsieur Zaccone, Number 34, and The Maltese Sailor. All these aliases have different relationships with different characters in different parts of the story and result in more complex graphs.

If one uses the same relationships between nodes (e.g. "child of"; "engaged to") it becomes possible to aggregate all these graphs together—across novels, or authors, for example—so as to try to reveal more insights. These more complex graphs can then be used to compare the described social structures around characters, for example, to compare protagonists and antagonists, or the interrelationship between characters of different genres and authors, or over time. We can answer questions like: "Are there specific authors that depict more inter-relational characters?" or "Do some authors focus more on factual descriptions of characters rather than their relationships?".

The second kind of information is aggregated information. This is information that is not contained in the individual pieces of text but in the combination of several pieces, and with such alteration that no individual piece of text necessarily corresponds to the aggregated information.

The simplest example of aggregated information is a frequency count. If we count the number of times *Edmond Dantès* was mentioned in the text, this number is not present in any individual line of the text but lives in a new space.[27] The frequency count cannot be verified by looking at any individual piece of the text.

A cluster of words is another example. A cluster is a grouping of elements (here words or sentences) such that the elements in the cluster are more similar to each other than to elements in other clusters.[28] Not all elements have to fit in a cluster. In addition, clustering can result in hard or soft clusters. In soft clusters, the elements are allowed into multiple clusters where they fit. For example, *Edmond Dantès* can fit in a cluster related to *sailing* and one related to *prison*—the character is introduced to the reader as a sailor who quickly ends up in jail. If the members of a cluster have a first-order similarity, it means that they are directly similar to each other, for example, because they co-occur in the same sentences. A *sailor* and a *ship* are similar because they co-occur in sentences like *The sailor sleeps on their ship*. Second-order similarity means that elements are similar because their "friends" (the words they co-occur with) are similar. For example, *sailor* and *captain* might be clustered together, because both *sailor* and *captain* are words used to describe people who work and live on a ship, sail the sea, and hence often co-occur with these words. However, *sailor* and *captain* do not need to co-occur directly themselves. A cluster containing words that have second-order similarity belongs in the category of aggregated information; the individual relationships need not be verifiable via readings of individual pieces of text.

Topics derived using *topic modeling* are yet another example of information that is aggregated. A topic model is a statistical model, taking the text-as-data perspective, that uses the insight that if a document (or a part of a document) deals with a certain abstract topic, then certain words are more likely to be used than others. A topic can be seen as the likelihood for each word in the vocabulary to belong to the topic. Topics are typically presented as *vectors* like the first line below, where each position corresponds to a word (shown in the second line). The number is a degree of membership (a probability between 0 and 1, where 1 corresponds to complete membership). A word can be highly likely or less likely to belong in a topic. For example, in the topic of *sports*, a *ball* is highly likely while a *cable* is highly unlikely. Topic modeling can be seen as a cluster of words, if we assume that only the most likely words in each topic belong to the cluster.

For the cluster that corresponds to *sailing*, we would have in Dumas's *The Count of Monte Cristo* words like { ship, sea, sailor, Pharaon, ...}.[29] The clustering is soft, as each word can belong to multiple topics. While topics represent aggregated information because no one document is responsible for a topic (for example, not all sailing is done using Le Pharaon, Dantès' vessel), we can easily go back to all pieces of text that contribute to the topic to verify the true information conveyed by the topic.

$$\begin{array}{lllllllll} \text{sailing} \ = & \{0.5, & 0.45, & \ldots, & 0.05, & \ldots, & 0.02, & 0.01, & \ldots\} \\ & \{\text{ship}, & \text{sea}, & \ldots, & \text{sailor}, & \ldots, & \text{Pharaon}, & \text{storm}, & \ldots\} \end{array}$$

Topic models are excellent ways of statistically depicting important themes in a body of texts, and comparing them across different data sets. It provides, for example, a way to find how a literary canon relates to a genre, or to all published fiction in a country.

Vector spaces fall in this category as well. Here, each word is represented as a vector, and the vector is learned using statistical properties of the words. Typically, the final vector space represents each word's semantic relationship (as opposed to its topical or thematic relationship). The values of a vector do not necessarily translate to information that can be interpreted on its own (like the probability of a word belonging to a topic); instead the information lives in the relationships among the words. Whether this information is of high quality is extremely difficult to evaluate on its own.[30]

The final form of aggregation that we will discuss represents the degree of change in any kind of information that can be compared over time. If the information at each time point is already aggregated, this can be seen as aggregating already aggregated information. Regardless, the end result is a value of some kind that only vaguely relates to the original text; a million documents over twenty years can be summarized by a single numerical value, for example, 0.734.[31]

EVALUATION OF RESULTS

There are many ways of evaluating results and the outcome of text mining.[32] In this section, we outline overall strategies to evaluate results in relation to a hypothesis.

Suppose, for example, that we want to find how sentiments toward modern technologies (like electricity or computers) change over time. Our text mining method is thus sentiment analysis. First, we need to verify that what comes out of the sentiment analysis for an individual sentence is correct. We should also verify that the method does not capture sentiments with respect to neutral concepts.[33]

Second, we need to verify that the output of multiple sentences (from the same time period) is correct, that is, that it corresponds to the expected output. If three sentences express a positive sentiment with respect to computers and two express a negative sentiment, what output do we expect? Perhaps we want to have 3/5 = 0.6 positive sentiment; perhaps we want to have absolute values, 3 and 2, or percentage values, 60% and 40% positive and negative respectively. We might want to weight the results on the basis of the strength of the expressed sentiment; for example, *horrible* is stronger than *bad*, *extremely* stronger than *very*. This might lead us to report stronger negative sentiment, even though there are more positive sentiments expressed in terms of absolute numbers. Observe that already at this stage we are introducing bias and subjectivity into the process. We should make our choices explicit for reproducibility and further evaluation. We also need to use expert interpretation to evaluate the results.

Once we have validated the results for a small set of sentences, we need to verify that the large-scale results correspond to what we expect. It is naive to think that if each individual piece is primary information and thus easy to verify, the same must hold for the information in a large data set. If we want to perform manual evaluation, we need to sample the data in such a way that the results in the sample mimic the results of the whole corpus, and then verify the results, preferably using random sampling of a statistically significant portion of the corpus. Here it is important to verify that the choice of data set is reasonable: do the results change if we add or remove a set of documents, for example, by removing a book or journal from the collection? If so, the results that we see are not stable and should be evaluated more thoroughly for a proper explanation, or discarded.

Probabilistic models, such as most topic models, produce results that are different for each run, and therefore, the outcome of multiple runs should be evaluated (c.f. section 5.5).[34] If the method is not appropriate for the amount of data we have, the results might be wrong; too little data yields results that are incorrect or can differ significantly over different runs, and too much data can lead to portions that are not taken into consideration.

If we want to perform (semi-)automatic evaluation, there are some different strategies. We could evaluate against a test set of pre-chosen examples.[35] We could test the output of the method.[36] And we could use control conditions.[37] The more aggregated our information, the more of these strategies should be chosen to obtain reliable results.

REPRESENTATIVITY OF RESULTS

The results of text mining resemble the fable of the blind men and the elephant. A group of blind men who had never encountered an elephant were given the chance to acquaint themselves with one by touching it. Each of them described the elephant in terms of the part that they had touched: an elephant's tail feels like a rope and the man who felt the tail described the elephants as like a rope; the belly feels like a wall, and the man who felt the belly described the elephant as being like a wall.

All of the men were correct in their description of part of an elephant. They are also all wrong, because none of them described the full elephant. If we take the analogy further, the men could put their pieces together and still not be able to describe a full elephant. At best, their description would serve as an approximation of a complex creature.

The results from text mining behave in much the same way—after having applied the full processing pipeline to our large-scale texts, we end up with a partial viewpoint. From this viewpoint we can only see a part of the full scene, as illustrated in Figure 3. We could try to evaluate and quantify the correctness of the viewpoint; however, as with the descriptions of the elephant, we also need to consider how complete or representative our viewpoint is. How much of the original texts is represented using the information that we have derived? How much of the original text remains (after the filtering steps) and was fed into the text mining method?

To evaluate completeness, we need to take the full processing pipeline into account. Assume that we start with a set of documents that contain running text. We run our processing pipeline from the list described earlier. Each choice that we make (including individual steps in the list taken or omitted) or choices within a step, will keep different portions of the original text. Here is an example of a single sentence:[38]

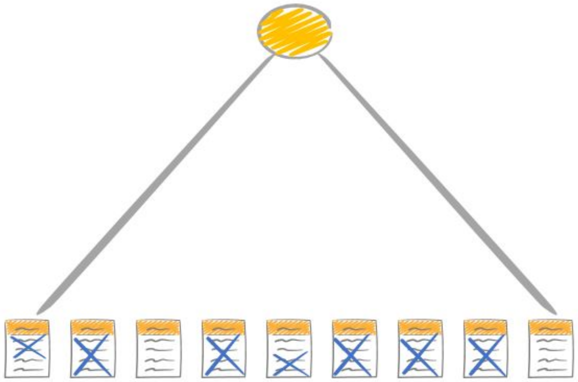

Figure 3: After the processing pipeline we end up with results that represent one viewpoint of the data set we started with. A single viewpoint might not be representative of the full data set, because it can have been derived using a small portion of the texts.

```
I like the room but not the sheets. (original sentence)
I like     room             sheets  (cleaning & removing stopwords)
I like     room             sheet   (lemmatizing)
           room             sheet   (keeping only nouns)
           room                     (frequency filtering)
  like                              (keeping only verbs)
```

As a larger real-world example we use *Pride and Prejudice* (1813) by Jane Austen that contains a total of 117,657 words (excluding punctuation marks and single letter words).[39] After filtering stopwords 54,970 words remain (53.3% of the words are removed). If we consider only nouns, 19% of the original tokens remain; and if we consider only verbs, only 13% are kept. Table 1 shows the remaining words after filtering for different parts of speech.[40] Please note that after filtering out stopwords, there are additional part of speech remaining than those represented in the table.

Table 1: This table shows the percentage different parts of speech constitute in Jane Austen's Pride and Prejudice.

Part-of-Speech	Percent of non-stopwords	Percent of all tokens	Total number of tokens
Nouns	41	19	22,304
Verbs	28	13	15,630
Adjectives	14	7	7,842
Adverbs	9	4	5,218
All of the above	93	43	50,994

While we can get an exact count of how much of the text we are filtering out using the different choices, we cannot know how much of the information stored in the original text is lost after filtering. The remaining text (and the information still contained therein) is processed by the text mining method that aims to find general patterns from large texts. This means further reduction of the text as well as the information contained in it. The result of the complete pipeline is a viewpoint, and we end up with different viewpoints depending on the choices we make (see Figure 4). The difference between using only nouns and only verbs is large and will result in viewpoints far from each other. In the same way, using different text mining methods will lead to different viewpoints that are likely far apart.

When interpreting the information in each viewpoint, the situation can be much worse than that represented in Figure 3. Not only do we have an incomplete picture of the whole data set at each individual viewpoint, but we also often only make use of a partial piece of the information available from our viewpoint. To give an example, a topic that is derived using topic modeling is a probability distribution over all words in our vocabulary (those words that remain after the processing steps). That means that each word belongs to the topic with different *strengths*.

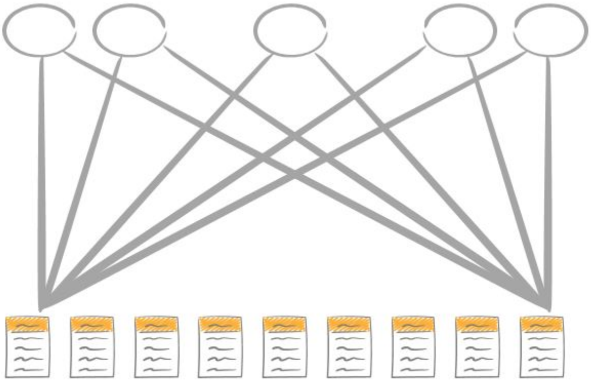

Figure 4: Different choices in the NLP pipeline result in different viewpoint from which we interpret our texts.

However, often only the strongest words are used to interpret one topic, typically only ten to twenty-five words, and without considering a term's likelihood of belonging to a topic.[41] So, instead of interpreting a topic with all the available information, many scholars interpret them by only using a fraction of the available information.

REASONING ABOUT HYPOTHESES ON THE BASIS OF RESULTS

Finally, the results of text mining are used in reasoning about one or several hypotheses. Regardless of whether the results are aggregated or primary, they should be interpreted with respect to our hypothesis and our document collection. There are some considerations to keep in mind during the process.

Rejecting a hypothesis If the results do not support our hypothesis, we need to determine the cause. A negative result is not necessarily a confirmation that the hypothesis is incorrect (or if our hypothesis is negative, we cannot automatically assume that a negative result means that we are correct). Instead, any of the alternatives below can be valid, and all should be considered.

1. The texts that we have chosen do not contain the information we seek. No text mining method can find what is not present in the data. Manual analysis of the texts is one way to determine that the information is indeed present. However, most text mining methods require multiple instances or mentions to be able to find the information. Thus we must determine that a sufficient number of examples of the information is contained in the data set. (This is very hard to do if one takes the exploratory path and hence does not know in advance which information is of interest!)
2. The chosen method is unable to capture the information that we seek. Even if the texts contain the phenomena that we are looking for, our method might not be suitable to capture the signals. Testing other methods is one way to verify this.
3. Our interpretation of the results is incorrect. We might think that our results support rejecting a hypothesis but that can be due to incorrect granularity of the results, incorrect normalization, or badly formatted results. This point may seem trivial, but it is often overlooked.[42]
4. The hypothesis we had is incorrect and should be rejected. Only after having excluded the other alternatives (1–3 in the list) should the hypothesis be rejected.

Accepting a hypothesis If the results support our hypothesis, there are still some questions to be answered before we can determine the validity of the results.

1. Are we correctly interpreting our results? In the same way as when our results support rejection, we need to verify that we have a correct interpretation of the results, taking into account, for example, different normalizations or formulas used for summing.[43]
2. How representative are the results? Verify the results by going back to the original documents and finding out how many documents, or paragraphs, contributed to this result. If it turns out that out of hundreds, thousands or more relevant documents, only a small fraction participate, then we must proceed cautiously when accepting a hypothesis and making use of the information for generalization.[44]
3. What are the results valid for? Do our results hold only for this specific data set, or is it likely that they are also true for other data sets? Can we use them for making inferences about the world outside? For example, does the result hold if we use different portions of the data set or different collections of text?

Point 2 relates to the *explanatory power* of the output. Our results are a small window through which we view our large-scale and possibly long-term text. Our window gives us access to an incomplete picture of the view, and different positioning of the window will result in different views. The different positioning of windows corresponds to the method and the preprocessing that we have chosen. If we draw conclusions about the life work of an author on the basis of a few written pages, or only very frequent nouns, the results that we derive may not be valid. It is highly likely that our results will be overturned when others scrutinize them in detail.

TOPIC MODELING EXAMPLE ON LITERATURE

We further illustrate evaluation by using an example from Matthew Jockers and David Mimno, and their work on significant themes in 19th-century literature.[45] We start by going through their paper and then reproducing their procedure on a small example to showcase some of the difficulties involved in data-intensive research.

A corpus of 3,279 works of fiction from the United States and Great Britain, spanning 1752 to 1899, is used as the basis for their topic model. The authors intend to investigate differences between female and male authors with respect to topics. The corpus is preprocessed by removing stopwords as well as character and personal names identified by named entity recognition software. Further along in the paper (3.2.3), the authors state that "the thematic information in this corpus could best be captured by modeling only nouns". Hence only common nouns were kept. The texts were segmented into passages of approximately a thousand words with breaks at the nearest sentence boundary, and the authors state that after a process of trial and error this resulted in "a set of highly interpretable and focused topics".

Following this, the authors are interested in investigating the proportion of words written by female and male authors related to specific topics. The remainder of the paper presents methods for analyzing differences between genders, on the basis of sound statistical properties, including the use of control data by random shuffling of author genders with respect to works. They also investigated the effects of individual works with respect to a given theme using bootstrap sampling.[46]

While the paper presents an excellent example of going deeper and beyond the results of text mining, there are a few things that are taken for granted that could affect the outcome and the conclusions drawn. First and foremost, the topics themselves are not discussed in any detail: the quality of the topics and the viability of the topics are left out. We are required to accept the authors' statement that the topics are indeed highly interpretable and focused.[47] There is no discussion around what is gained or lost by excluding other parts of speech: what happens to our topics we include adjectives, verbs, and adverbs?

Additionally, the topics are based on only nouns that are not names. How much of the text remains once we have filtered out everything else? Table 1 above showed a small example of a single novel, but we do not know what the corresponding numbers would be in the data set used by Jockers and Mimno. How different would the topics be if we added verbs, adverbs, adjectives? The authors count the proportions of words in the novel assigned to a specific topic, after having removed all other words. But if a topic has a higher proportion among men, maybe it is because men use more nouns and fewer adverbs? This highlights the importance of normalization, and the need to explicitly state how normalization was performed: are we comparing the proportion of words assigned to a topic compared to all nouns, or compared to all words written by that author?[48]

Pride and Prejudice We illustrate some of the intuitions on Austen's *Pride and Prejudice* as our basis. Because topic modeling produces aggregated results we cannot start the process of evaluation of single instances. Instead we need to test and validate multiple instances, by measuring the quality of the topic models, and by evaluating the topics on their own. Do we find topics that correspond to what we expect?

Therefore, our first step is to evaluate the topics on their own, and the quality of the topic model as a whole. While there are multiple ways to evaluate the latter, we choose a topic-coherence measure that considers whether the words in a topic tend to co-occur together.[49] This procedure constitutes testing different number of topics to find the most coherent model.[50] We test with up to 40 topics with increments of 5 and different passage sizes. The results can be found in Figure 5 where it seems that 7 topics produce the best coherence.[51] Each topic is the dominant topic of between 10.5 to 18.6% of the passages (the corresponding number would have been $\frac{1}{7}$ = 14.3% if the passages were assigned randomly to a topic).[52]

The next step is to evaluate the correctness of the method on large-scale text. Here we can choose a pre-chosen strategy or evaluate the topic outcome of the method. The latter is the most common method. Consider the topics and evaluate them with respect to intuition; do they make sense? While this corresponds to *precision* (how good are the results), the first strategy corresponds to *recall* (how many of the expected themes do we find). Recall is important, as it tells us how much of the information in the book/s contribute to the themes.

We can test our interpretation of the topics by checking how many of the most likely passages reflect our interpretation of that specific theme. We can also apply different kinds of control conditions: what happens if we test our topics on passages that are completely off topic, for example taken from modern or scientific language?

For this evaluation, the first step is to recognize that our topic model is probabilistic and therefore produces different results each time it is run.[53] The second step is to look closely into the topics themselves. Do the topics make sense? And while parts of the answer lie in looking at the most likely words for each topic, this is not sufficient. In

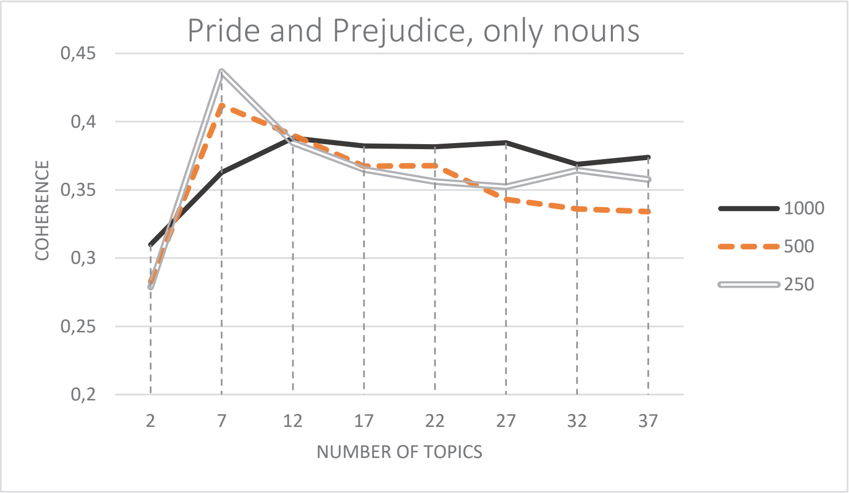

Figure 5: The coherence of the topic models when using only nouns and in passages of roughly 250, 500, and 1000 words.

Table 2 we show different topics, derived using the same passage size of 250 and 7 topics, with different preprocessing. In the first column we have nouns with names, and in the second we have nouns without names. The third column represents nouns, verbs, adjectives, and adverbs (NVAA) with names, and the last column represents NVAA without names. In the table the first three columns correspond to the "same" theme, while the fourth column is chosen at random as the topic model does not seem to result in a corresponding theme after we have removed the names.

How are we to determine which of these is a "better" topic? To understand these topics we need to look at the paragraphs that contributed to the topics and make a qualitative judgment.[54]

When it comes to humanities data, we encourage the researcher to thoroughly investigate the output of any text mining method, including topic modeling. Indeed, research indicates that automatic metrics for the quality of a topic do not always correlate with human judgments.[55]

Once our topics are evaluated properly through close examination and reading of passages, we can capture trends over time or relation to metadata like gender. Again, these results should also be evaluated, for example using methods presented by Matthew Jockers & David Mimno.[56]

nouns with names	nouns without names	NVAA	NVAA without names
darcy	sister	darcy	feel
miss	miss	bingley	give
bingley	room	sister	love
sister	brother	miss	happy
friend	hour	elizabeth	family
brother	conversation	jane	happiness
evening	party	friend	present
ball	attention	pleasure	hope
country	visit	attention	marriage
pleasure	ball	brother	mention
gentleman	door	evening	affection
room	table	netherfield	mind
dance	minute	half	object
conversation	rest	behaviour	general
consequence	book	leave	power
delight	opportunity	join	wife
partner	silence	scarcely	heart
dare-say	smile	engage	make
card	admiration	visit	promise
persuade	question	country	persuade

Table 2: The top twenty words chosen for one topic across different models using different words. NVAA stands for nouns, verbs, adjectives, and adverbs.

Hypotheses and research questions

One of the great challenges of computational literary studies—and of the digital humanities in general—is reasoning about how results from text mining can be used to corroborate or reject a hypothesis. This amounts to interpreting the results and "translating" them into conclusions about the original research question. Here is where the humanities' in-depth domain knowledge comes into play. However, let us compare three different starting points for a research process when it comes to the relationship between research question and hypothesis.

1. *One research question and one hypothesis:* A researcher is interested in how the general sentiment with regards to a concept, such as a trade or technology has

changed over time. The research question focuses on "how", and data and method are designed to follow the exploratory path. If this results in a more precise hypothesis about how notions have changed, then this hypothesis can be corroborated or refuted through the validation path with adjusted data and method.

2. *One research question and several hypotheses:* A researcher is interested in how gender-equality discussions have affected children's literature. This research question is very broad and needs to be broken down into several questions, and a number of them must be used to answer the question in full.[57] Suitable data and methods need to be devised. By following the exploratory path, these questions can be reformulated as propositions or hypotheses, which are tested using the validating path.[58]
3. *Data and text mining method but no research question:* We can envision a case where there is an interesting source of data but no clear research questions (for example, the digitized letters of an influential author). A text mining method can be used to find interesting patterns and signals to explore further. That is, we follow the exploratory path to find a rewarding hypothesis. The focus is on the data and the text mining method. Often, a method like topic modeling is used as a way of obtaining an overview of different themes around a concept of interest. These topics can be explored and good hypotheses formulated in a more informed fashion.

There are dangers with the exploratory path, and in particular with the last point in the list where there is no clear research question. If the results are very interesting, it can be hard to see beyond the results and properly reason about their value and correctness.[59] Using the evaluation strategies outlined in previous sections, the exploratory path can be useful for discovering new insight. To ensure the correctness of the results when using the exploratory path, results need to be verified using multiple runs of the same method with different parameters (and the same parameters if the method is probabilistic) as well as additional methods. It is also good practice to test using different parts of the data set, to ensure that certain parts do not affect the results significantly. In other words, we must ensure that we are not uncovering particularities of specific parts of the data set, but rather general trends.

Model of interpretation – Interpreting research questions using results

In the traditional humanities, the researcher is the bridge between results and interpretation. In data-intensive research, the situation is slightly different. The typical result of a text mining method is not necessarily directly interpretable in terms of the hypothesis, nor do the hypotheses need be directly interpretable with respect to the research question. The process of moving between results and the research question is in itself a result and that requires evaluation.

To exemplify a *model of interpretation*, consider the work on sub-corpus topic modeling (STM) by Tim Tangherlini and Peter Leonard on the impact of Darwin's theories on Danish literature. Two of Darwin's books (the sub-corpus) in the Danish translation, are concatenated and topics are derived from these books. These topics are then labeled, and some of them are used for "trawling" the Google Books versions of Danish literature. The authors state: "As hoped, the algorithm discovered a number of texts supporting the contention that Darwin's topics were influential outside of the natural sciences, including several excerpts from the intellectual press".[60] A few different passages are presented in the paper, and the end of the experiment concludes: "These examples are a small sampling of the 'catch' that the STM trawl-line produces—apart from discovering numerous examples from the literary realm (both canonical and non-canonical), the trawl-line vastly expands our understanding of the reach of Darwinian ideas in the Nordic region, penetrating not only into realms such as historiography, but also into realms such as public policy".[61]

Clearly, the model of interpretation is missing. It is vague and unclear how the authors go from the topics derived from the sub-corpus, and the few examples presented in the paper of the literature that would correspond to the said topics, to the conclusions they draw. Firstly, how many passages are there in total that are found by the topics? How strong is the connection between the topics and the passages? Were there any passages that fit the topics *before* Darwin's books were published? That is, are these really Darwin's topics or are they general topics that are also found in his books? There is a large gap between the ideas put forward by Darwin, which we know to be novel, and the information that is modeled in the topics, which might very well be general. If they had clearly stated their model of interpretation, how they moved from the output of a topic model and corresponding passages in literature, to answering the research question, others would have been able to repeat their experiments. Alternate methods could use different corpora (instead of Google Books), other parameters of the topic model, or other text mining algorithms. As it stands now, it is not possible to repeat the experiment in a comparable way.

We argue that all data-intensive projects that aim to answer broad research questions, like those in the humanities, should make their model of interpretation clear and preferably evaluate it with respect to alternative models.

Conclusions

Computational literary analyses (and digital humanities in general) have great potential to reform and contribute to both the humanities and the data sciences. Digital methods and material open the doors to confirming existing hypotheses using large-scale texts with more authors, longer time spans, and new kinds of analyses. They also open the possibility of asking new questions in venues previously unattainable.

None of these methods removes the need for in-depth knowledge; from formulating research questions and breaking them down into reasonable hypotheses to interpreting the results and reasoning about their implications, the humanities scholar is an integral part of the research. By combining a data-intensive research methodology with traditional and modern humanities, much can be gained, both in terms of new insights and in terms of new data science methods for tackling these complex issues.

In this meeting of the data sciences and the humanities, there are only gains to be had, and the meeting should be approached with respect from, and toward, both sides. The data scientists bring with them an understanding of digital methods, results, and large-scale, long-term analysis, and how challenges related to these can be overcome. The humanities scholars bring wide research questions, the interpretation, and the relation between a research question, hypotheses, and data. Together, both grow stronger.

NOTES

[1] This paper is built on a presentation given at the annual meeting of the Swedish Society of Literature. https://play.gu.se/media/0_h2399x0b, accessed: November 1st, 2019. Parts are based on a keynote address given at the Sixth Estonian Digital Humanities Conference in Tartu in September 2018. Finally, some parts of this paper have been published by Tahmasebi and colleagues and are part of collaborative work. (Nina Tahmasebi, Niclas Hagen, Daniel Brodén & Mats Malm, "A Convergence of Methodologies. Notes on Data-Intensive Humanities Research", in *Proceedings of the 4th Conference on Digital Humanities in the Nordic Countries (DHN 2019)*.) Nina Tahmasebi produced the first version of the paper. Simon Hengchen contributed examples and references, and re-wrote small sections of the paper. Both authors gave final approval for this version of the paper. The authors' thanks go to Niclas Hagen, Daniel Brodén, and Mats Malm for interesting and stimulating discussions during the writing of our joint paper. A particular thanks to Mats Malm for providing valuable insight and reading earlier versions of this paper. This paper has been funded in part by the project *Towards Computational Lexical Semantic Change Detection* supported by a project grant (2019–2022; dnr 2018-01184), and the Centre for Digital Humanities at the University of Gothenburg, to Nina Tahmasebi.

[2] See for example Dong Nguyen, Maria Liakata, Simon DeDeo, et al., "How we do things with words. Analyzing text as social and cultural data", 2019, arXiv preprint arXiv:1907.01468.

[3] A broader perspective on the data-intensive research process, including the use of research infrastructures and archives, can be found in David M. Berry & Anders Fagerjord, *Digital Humanities. Knowledge and Critique in a Digital Age*, John Wiley Sons, 2017. The authors remark that it is important to keep the focus on the research questions, instead of turning the field into a race for better and more effective algorithms (page 49). However, it is our strong belief that focus on the research questions will automatically lead to bolder steps into the unknown that will result in better and more effective algorithms, so the two objectives are not exclusive, but rather joint.

[4] That excludes HTML, XML, and other annotation frameworks, and includes titles, references, captions and so on, written by the author/s. While digital transcripts of spoken language, whether from plays, conversations or discussions, also constitute digital text, they are rarely considered in textual data sets because they often differ substantially in character. Often times, there is a lot of metadata involved in describing who said what, or directions to the actors, that interfere with what is being said. However, transcribed discussions often constitute a counter-example, and are included in many textual corpora, like Hansard (E. Odell. Hansard Speeches and Sentiment V2.5.0 [Data set], Zenodo, 2018. http://doi.org/10.5281/zenodo.1183893) or Swedish Parliament records (Riksdagens öppna data, https://data. riksdagen.se/data/anforanden/)

[5] Like the research question put forward by Timothy Tangherlini and Peter Leonard in "Trawling in the Sea of the Great Unread. Sub-corpus Topic Modeling and Humanities Research", *Poetics*, vol. 41, 2013:6, pp. 725–749. "Can we find traces of this shift to a natural-scientific understanding of society presaged by the translation of Darwin's works in the 1870's by Jacobsen in the larger corpus of Danish language works in Google Books?" (p. 735)

[6] In historical texts, like the Google Books corpus, men are almost ten times more likely to be mentioned than women, until the beginning of the 20th century, when the two concepts begin moving toward the middle and finally meet somewhere in the 1980s. See Google N-gram viewer, men and women, https://tiny.cc/5wus6y, accessed: 2019-05-16, 2019.

[7] This also applies to most originally digital text that is studied only as running text without information on layout, font color or size, or relation to figures or pictures. Studies in fan fiction, for example, use born-digital text. If such text is to be collected from the web pages directly, that is "scraped", this introduces additional sources of noise. Detecting the core parts of the text embedded in a web page structure encoded in HTML is far from trivial and can result in very noisy data.

[8] If the first car slows down, so must the following cars. But if the last car in the lane slows down, that has no effect on the preceding cars; however which car is first or last does not matter, unlike with words where it often is important.

[9] Stopwords are very frequent words that rarely carry information, like *and*, *or*, *it*, *a*.

[10] Often our tools are trained on "standard" language. Their performance can be significantly worse on historical texts or modern out-of-domain texts.

[11] Currently, computers do not have access to the additional sensory data available to us, which further imposes limitations. They do not see eye movement, facial expressions, hand gestures, or hear the tone of voice. They have access only to what has been said, not to how it was said. Think of an email or text message that was hard to interpret and that felt strange. It may be that you were not certain whether the content was meant as a joke or as a harsh reprimand. The situation would likely have been different if you had received the same message directly, face to face, and had been able to interpret additional clues such as a smile or a frown.

[12] Rob High, "The Era of Cognitive Systems. An Inside Look at IBM Watson and How It Works", IBM Corporation, Redbooks, 2012; D. A. Ferrucci Introduction to "This is Watson", *IBM Journal of Research and Development*, 56(3.4):1:1–1:15, May 2012. ISSN 0018-8646. doi: 10.1147/JRD.2012.2184356.

[13] Franco Moretti, *Graphs, Maps, Trees. Abstract Models for a Literary History*, Verso, 2015.

[14] In text-based computational sciences, a document is any unit of text. Depending on the research question, this translates in DH as a paragraph, a chapter, a whole novel, etc.

[15] Originally, a corpus was a linguistically motivated collection of text aimed at representing language phenomena, but very often in the digital humanities the term *corpus* is used interchangeably with *data set*. See Sue Atkins, Jeremy Clear & Nicholas Ostler, "Corpus Design Criteria", *Literary and Linguistic Computing*, vol. 7, 1992:1, pp. 1–16 for a discussion on creating a corpus and sampling biases in corpora.

[16] The list is in no way comprehensive, and many more possibilities are available than those presented here. Similarly, while some steps must be done in a certain order (one cannot remove stopwords if the text is not tokenized, for example), the order presented below is purely for presentational purposes and does not reflect all NLP pipelines. In addition, certain steps can be parallelized.

[17] Some terminology: a *token* is a single occurrence of a linguistic unit (usually, a word), whereas a *type* is an abstract class representing all occurrences of the same token. To illustrate this point: *to be or not to be* contains 6 tokens (to; be; or; not; to; be), but 4 types (to; be; or; not). For the discussions in this paper, a token is a space-separated word (set of characters).

[18] Philip Pullman, *The Amber Spyglass*, London: Scholastic/David Fickling Books, 2000. (p. 517).

[19] The distributional hypothesis, first introduced by Harris (Zellig Harris, "Distributional structure", *Word*, vol. 23, 1954, pp. 146–162), can be characterized by the quote "You shall know a word by the company it keeps", John Rupert Firth, "A Synopsis of Linguistic Theory, 1930–1955" (p. 11), in *Studies in Linguistic Analysis*, J. R. Firth et al. (eds.), Oxford: Blackwell, 1957.

[20] Context can be defined differently, and can involve words in a certain grammatical relation, separated by certain patterns, for example, *A such as B*, *A including B*. Such patterns

can be captured in a context window, although a context window captures much more than only the words involved in these patterns.

[21] The HistWords vectors, a set of pre-trained vectors for historical texts often reused by others, is an example of reused representation of words. William L. Hamilton, Jure Leskovec & Dan Jurafsky, "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change", ACL 2016. Representations can also be created using count-based methods, where the dimensions of the vector are directly interpretable and correspond to other words, see for example, H. Schütze, "Automatic Word Sense Discrimination", *Computational Linguistics*, vol. 24, 1998:1, pp. 97–123.

[22] Typical text mining methods are, among others, topic modeling, sentiment analysis, clustering, and argument mining.

[23] In Maria Antoniak & David Mimno, "Evaluating the stability of embedding-based word similarities," *Transactions of the Association for Computational Linguistics*, vol. 6, 2018, pp. 107–119, the authors remark that some properties of word embedding models are highly sensitive to small changes in the training corpus, and especially so in smaller corpora.

[24] Unlike mineral mining though, where only one kind of mineral can be mined regardless of method, the information we mine from the very same texts can be very different, ranging from word statistics to arguments, sentiments, translations and so on.

[25] A resource can be a document, web page, or wiki entry, for example.

[26] Whether relevant information was missed is another matter. That is often referred to as *recall*: how much of the relevant information was captured. Extractive (multi-)document summarization also falls into this category. The task is to pick a set of sentences from the original document(s) that best describe the important information conveyed inside the document(s). Typically there is an additional constraint that the summary should be short and concise. When the summary has been constructed, each sentence is unaltered and a human evaluator can read the summary and conclude whether it captures the expected information.

[27] 44 times in the English-language edition available at http://www.gutenberg.org/ebooks/1184, 36 in the Finnish translation (http://www.gutenberg.org/ebooks/45448), and a combined 37 times in the four tomes in the original French, available at: http://www.gutenberg.org/ebooks/17989, http://www.gutenberg.org/ebooks/17990, http://www.gutenberg.org/ebooks/17991, http: //www.gutenberg.org/ebooks/17992.

[28] There are many ways of measuring similarity, for example semantic similarity, or the number of overlapping characters in each word.

[29] *Le Pharaon* is the name of Dantès' ship. As such, it is a likely word for "sailing" in the novel by Dumas but not especially in the larger context of all novels ever written, and even less so in works in French since *pharaon* is a noun that refers to an ancient ruler of Egypt.

[30] As a result, the quality is often evaluated by subsequent tasks, like machine translation or detection of analogy (for example, *man* is to *woman* as *king* is to *queen*). In the case of neural embeddings, it is not possible to return to the original documents or sentences that contributed to the representation of each word. Neural models correspond to a suite of algorithms that, among other things, can be used to learn vector representations of words

that capture different semantic properties. PPMI and SGNS embeddings are among the more common methods for producing vectors that capture the semantics of words. See further Yoav Goldberg, *Neural Network Methods for Natural Language Processing*, London 2017. for methods and applications of neural embeddings for language modeling.

[31] Recently, there has been much scholarly effort in detecting change over time, in particular when it comes to semantic changes. These works translate well to change in themes; how have different themes been discussed over time, but also change in meaning of individual terms; how are God, love or particular leaders represented over time. See: Qiaozhu Mei and ChengXiang Zhai. "Discovering evolutionary theme patterns from text: An exploration of temporal text mining" In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD '05)*, 2015, pp. 198–207; Lea Frermann and Mirella Lapata. "A Bayesian model of diachronic meaning change" *Transactions of the Association for Computational Linguistics*, vol. 4, 2016, pp. 31–45.

[32] For a larger overview, we refer the readers to Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Upper Saddle River 2000; and Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. Cambridge, 1999. For an overview specific to computational literary studies, see Nan Z. Da, "The computational case against computational literary studies", *Critical Inquiry*, vol. 45, 2019, pp. 601–639.

[33] If we aim at deep parsing, we should make sure that the method produces the correct sentiment with respect to complex sentences with multiple possible targets, for example, *I like the room but not the view*.

[34] A *run* is each time the data is processed using the NLP pipeline and results are obtained. As a topic model is a probabilistic model, it starts by making a random choice, and refines it with each iteration. Note that if one uses the same seed for the random number generator that starts the process, one can force different runs to produce the same output. Fixing a seed is a way of forcing the first guess.

[35] We can evaluate our results with respect to a set of words or examples that we know should behave in a certain way. After having run the text through our NLP pipeline, we compare our results for the pre-chosen set of examples to the expected results. The *expected result* is typically generated manually and requires in-depth knowledge of the field. For example, from previous studies coupled with results of general elections, we have an estimate of the general sentiment toward nuclear power, or presidential candidates, for a specific time period and place. If we use the outcome of elections as an estimate of the public opinion expressed in text, we consider such estimates to be *silver standards*. A silver standard is a less-than-perfect standard because there are multiple levels of uncertainty involved: those represented in the text need not be the same people who voted. We could have attained a gold standard had we asked the authors of the text for their opinion, which is almost impossible to do retroactively. Another way of attaining such silver standards is to use the ratings people give to a movie as a way of grading the opinions they express in their review of the movie. We compare the output of our method to what we expect and measure how often

the method is correct, and possibly to what degree it is correct. While some results may be binary, others can be a numerical value; being off by a few percent can still be considered partially correct. However, only using examples that we know should behave in one specific way is not sufficient. We should also chose examples that reflect the opposite view. We consider these as *counterexamples*. The utility of this test can best be explained by considering a method that is meant to detect change over time. The method might be finding that everything changes. If we only test the method on concepts that we expect to change, we would only be confirming our own bias. If we test the method also on concepts that should not change, we control for this issue and measure how well the method can separate between the two classes.

[36] We can evaluate a small portion of the outcome of the method using manual analysis. We can go about this in two different ways: (a) randomly sample a set of concepts and investigate whether the method is correct, or (b) test the top and bottom outputs of the method if the method produces a ranking. In the example of change over time, it would correspond to investigating the concepts that are found to change the most and those that change the least, and determine how correct the method is for each set. This kind of evaluation measures *precision* but not *recall*, that is, we cannot say anything about how many correctly changed concepts are not among the top results, but we can say how many of the ones that we do have that are correct. See Ricardo A. Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern Information Retrieval*, Boston, 1999, for a definition and further examples of precision and recall.

[37] A third method of evaluation is to use controlled data (or experiments). In this case, we do not try to find what is really out there, but artificially change the conditions and test whether or not we can find the changes we have made. In the example of the sentiments, we can choose a set of randomly selected neutral words (or create new words like *chortle*, a word coined by Lewis Carroll in his 1855 poem *Jabberwocky*, and made famous in his 1871 novel *Through the Looking Glass*. By using a non-existing word, we make sure that there are no accidental signals that interfere.) and then generate sentences around the words with sentiments that we know are positive or negative. In this way, we determine how much positive and negative sentiment the method is expected to find and compare the output to the expected outcome. Another way of controlling data is to shuffle the data around. This is particularly important if we are interested in diachronic analysis. If we expect our method to find differences (for example, change over time), these differences should not be found if we have shuffled all our data between the years (or authors, or data sets, for example).

[38] For a larger discussion on the effects of preprocessing (lemmatization, stopword removal) for topic models, we refer to Alexandra Schofield & David Mimno. "Comparing apples to apple: The effects of stemmers on topic models", *Transactions of the Association for Computational Linguistics*, vol. 4, 2016, pp. 287–300; and to Alexandra Schofield, Måns Magnusson, & David Mimno. "Pulling out the stops: Rethinking stopword removal for topic models.", *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 2, 2017, pp. 432–436.

[39] Downloaded from Project Gutenberg, http://www.gutenberg.org/ebooks/42671

[40] The text was processed with spaCy (https://spacy.io/), a Python NLP library. Different part-of-speech taggers will give us different results and might change the values presented in Table 1.

[41] For example, the fifteen first words are equally likely to belong to a topic, still, only the ten first are used for interpretation.

[42] Andrew Goldstone and Ted Underwood. "The quiet transformations of literary studies: What thirteen thousand scholars could tell us", *New Literary History*, vol. 6, 2014, pp. 359–384. They write that "individual topics always need to be interpreted in the context of the larger model." (p. 367) They show this, for example by comparing individual topics to the token frequency of the data set as a whole, and to other topics that use the same interesting words. The need to interpret results in context is true for most results.

[43] In the case of diachronic text and temporal analysis, can the result be a property of having better quality or a higher quantity of texts over time? One possibility is to verify by manual inspection of a random sample of the texts.

[44] It is important to consider relevant documents, not all documents. If we are investigating attitudes toward nuclear power using topic models, we should consider how many documents about nuclear power each topic relating to nuclear power covers. There could be one topic about a specific opinion considering nuclear power which is expressed in only a few documents. This scarcity of opinion becomes particularly relevant when we consider collections with extensive text reuse. In parliamentary data where one report leads to a motion, that leads to a debate, and finally a proposition, the same text can be reused multiple times and form a coherent topic, but may not be very representative of the collection as a whole.

[45] Matthew L. Jockers & David Mimno, "Significant themes in 19th-century literature", *Poetics*, vol. 6, 2013, pp. 750–769. Part of special issue on Topic Models and the Cultural Sciences.

[46] "What became clear was that the high value of the corpus mean for [the theme] 'Convents and Abbeys' was largely the result of few outlier texts that were pulling the mean in an artificially high direction." (p. 763–764). By using statistical testing and a sound methodology, some intuitions were overturned, making a strong argument for not trusting the result of text mining, or the conclusions drawn on the basis of the results, without further investigation.

[47] The situation is eased by the fact that the topics are available for further investigation. However, the topics are in the form of word clouds, which makes investigation more difficult.

[48] Let us assume that Author A has 1000 words assigned to topic T, and Author B has 700 words assigned to the same topic T. If we only look at the absolute numbers, Author A has a higher assignment of topic T than Author B has. Now, let us assume both authors have books that are in total 100,000 words long. The *global proportion* of words assigned to topic T by author A is $\frac{1000}{100000} = 0.01$ and by Author B is $\frac{700}{100000} = 0.007$. Again, Author A has a higher assignment of topic T than Author B has. If however, Author A has more nouns than Author B, for example, Author A has 15,000 nouns, and Author B has 9,000 nouns,

then the *noun proportion* of topic T for Author A is $\frac{1000}{15000} = 0.067$ and the corresponding for Author B is $\frac{700}{9000} = 0.078$, which means Author B has a higher assignment of topic T than Author A has. So, depending on which normalization we choose, we get two opposing results.

[49] See http://mallet.cs.umass.edu/diagnostics.php for a description on how coherence is calculated.

[50] In topic modeling, the number of topics is referred to as the k parameter.

[51] The maximum value is reached at $k = 7$ for a passage size of roughly 250 words, where the coherence is 0.44, a value that we can consider relatively small. We tested with different chunk sizes of 500 and 1000 words each, and this resulted in coherence decreasing to 0.412 and 0.388 respectively.

[52] In their paper, Jockers and Mimno choose only nouns excluding named entities in passages roughly equivalent to 1000 words, however, their data set was significantly larger than our small example. Furthermore, Jockers and Mimno opted to remove names, as these are very prominent among the top words. Removing names in our case results in different optimal passage sizes and number of topics, where passages of 150 words and 52 topics provides the highest coherence value of 0.416. However, while the coherence score is only slightly lower than the version with names, the topic distribution over chunks is skewed. Here the topic assignment is between 0.15% to 6.8% (the random equivalence here would be $\frac{1}{52} = 1.9\%$), and only half of the topics cover more than ten passages, indicating that coherence scores alone are not sufficient to judge the quality of the topics.

[53] For example, five different runs with the optimal model (passage size 250 and 7 topics) resulted in an average coherence score of 0.43 with the highest value at 0.45 and the lowest 0.42. A varying coherence score is indicative of varying topics.

[54] Tangherlini & Leonard 2013 provide such functionality in their sub-corpus topic model: "[…] the researcher can check the ranked list of the top *n* documents contributing to any given topic and can adjust proposed labels, as well as the initial values of document length, number of topics T,. […]" (p. 732)

[55] Jonathan Chang, Sean Gerrish, Chong Wang, et al., "Reading tea leaves: How humans interpret topic models", *Advances in Neural Information Processing Systems*, 2009, pp. 288–296.

[56] Usually, young women write differently than old men, so how can we make sure that the algorithm picks up themes and not style? To remove special bias from the topic modeling, we refer to other flavors of topic modeling, for example as proposed by Laure Thompson & David Mimno, "Authorless topic models: Biasing models away from known structure", *International Conference on Computational Linguistics*, 2018, pp. 3903–3914.

[57] Note that instead of breaking down the research question into several separate questions, we can keep one large question and formulate a large number of hypotheses. However, for simplification, we can also separate the RQ into multiple RQs. If we think of the research question, RQ, as a field and the hypotheses as smaller fields (circles), we should choose the hyphotheses to cover as much of the RQ circle as possible. In some cases, we need to for-

mulate hypotheses that overlap to capture as much of the RQ as possible. Sometimes the hypotheses can cover all of the research question; sometimes that is not possible.

[58] It is also possible to directly use the validating path by defining hypotheses. For example, a clear hypothesis could be that the number of female main characters in children's literature has increased over time; however, this hypothesis does not cover the research question in full but requires multiple additional hypotheses. The number of female characters must be normalized by the total number of main characters to account for the increasing number of children's books over time.

[59] For an interesting case, we refer to the Torah codes, a method for finding predictions about the future as written in the Torah. The predictions were so fascinating that the inventors did not recognize that the method (Equidistant Letter Sequence) had an inherent property: given a sufficiently large data set and a large set of possible things to look for, the likelihood of finding interesting results by chance is large (for example, a detailed account of the death of JFK can be extracted from Moby Dick). See: Doron Witztum, Eliyahu Rips, & Yoav Rosenberg. "Equidistant letter sequences in the book of Genesis." *Statistical Science*, vol. 9, 1994, pp. 429–438.; Brendan McKay. "Assassinations Foretold in Moby Dick!" https://users.cecs.anu.edu.au/~bdm/codes/ moby.html, accessed 2019-04-25, 2019.

[60] Tangherlini & Leonard 2013, p. 736.

[61] Tangherlini & Leonard 2013, p. 737.

ABSTRACT

Nina Tahmasebi, Centre for Digital Humanities, Språkbanken, University of Gothenburg
Simon Hengchen, COMHIS, Department of Digital Humanities, University of Helsinki

The strengths and pitfalls of large-scale text mining for literary studies

This paper is an overview of the opportunities and challenges of using large-scale text mining to answer research questions that stem from the humanities in general and literature specifically. In this paper, we will discuss a data-intensive research methodology and how different views of digital text affect answers to research questions. We will discuss results derived from text mining, how these results can be evaluated, and their relation to hypotheses and research questions. Finally, we will discuss some pitfalls of computational literary analysis and give some pointers as to how these can be avoided.

Keywords: text mining, data-intensive research methodology, computational literary analysis