FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Journalism 3.0: Multidimensional Cluster Visualization and Labelling on Twitter Data for Data Journalism

Bruno Miguel Alves Vieira

Master in Electrical and Computers Engineering

Supervisor: Prof. Carlos Soares

Co-supervisor: Jorge Teixeira

August 23, 2016

Abstract

Twitter is a prominent microblogging service that handles large amounts of diversified information. This information can be accessed by means of Twitter's API, allowing the data collection to be done by request or event handling. This possibility allows the analysis of large and current datasets for knowledge discovery.

TweeProfiles is a tool to analyze and visualize patterns of Twitter data on three dimensions, namely spacial, temporal and content. The clustering process is divided into two phases, the attainment of micro and a macro-clusters. The micro-clustering phase is a constant overlapping procedure of streamed tweet batches. The macro-clustering phase is a request based pattern discovery on a supplied dimension distribution.

Twitter's great amount of available information presents a strenuous task in identifying subjects of interest. The TweeProfiles tool identifies patterns within the gathered data as tweet clusters. Journalists explore social networks for news worthy information and could benefit from the TweeProfiles results.

The project described in this dissertation had two goals: evaluate the use of topic extraction techniques on clusters of tweets and adapt the TweeProfiles tool to assist the journalist's exploration of large amounts of Twitter data.

Cluster labelling is achieved by performing a text summarization Topic extraction task. We test three unsupervised algorithms, namely TF-IDF, Pagerank and LDA. A preliminary study of their behavior on Twitter data was performed, analyzing their ranking methods and top@N agreement. This task was assigned to the micro-clustering instance, therefore continuously labelling new formed clusters.

TweeProfiles was adapted based on requirements identified together with the JornalismoPortoNet(JPN) team. The evaluation was based on a usability test and an interview with the editor of JPN.

A cluster labelling step was successfully added to the TweeProfiles back-end and its front-end was adapted according to the feedback provided by JPN media journalists.

Resumo

O Twitter é um proeminente serviço de microblogging que lida com grandes quantidades diversificadas de informação. Esta informação pode ser obtida pela API fornecida pelo Twitter, permitindo que a recolha de dados seja feita por pedidos ou por tratamento de eventos. Esta possibilidade permite a análise de conjuntos de dados grandes e atuais na descoberta de conhecimento.

O TweeProfiles é uma ferramenta de análise e visualização de padrões de dados do Twitter em três dimensões, nomeadamente espacial, temporal e de conteúdo. O processo de clustering divide-se em duas fases, a obtenção de micro e macro-clusters. A fase de micro-clustering é um procedimento de constante sobreposição de lotes de tweets obtidos em stream. A fase de macro-clustering é baseada em pedidos realizados, com uma distribuição de dimensões, para obtenção de padrões.

A grande quantidade de informação disponível pelo Twitter apresenta uma tarefa árdua na identificação de temas de interesse. A ferramenta TweeProfiles identifica padrões dentro dos dados recolhidos como clusters de tweets. Os jornalistas exploram as redes sociais por informações dignas de notícias, podendo beneficiar dos resultados do TweeProfiles.

O projeto descrito nesta dissertação teve dois objetivos: avaliar o uso técnicas de extração de tópicos em clusters de tweets e adaptar a ferramenta TweeProfiles para ajudar os jornalistas na exploração de grandes quantidades de informação do Twitter.

A etiquetagem de clusters é conseguida através da realização de uma tarefa de Topic extraction para sumarização de texto. Testamos três algoritmos não supervisionados, nomeadamente o TF-IDF, Pagerank e LDA. Um estudo preliminar do seu comportamento em dados do Twitter foi realizada, analisando os seus métodos de classificação e acordo nos top@N. Esta tarefa foi associada à instância de micro-clusters, realizando continuamente a etiquetagem de novos clusters formados.

O TweeProfiles foi adaptado com base nos requisitos identificados juntamente com a equipa do JornalismoPortoNet(JPN). A avaliação foi realizada com um teste de usabilidade e uma entrevista com a editora do JPN.

Uma etapa de etiquetagem de clusters foi adicionada ao back-end da ferramenta TweeProfiles e o seu front-end foi adaptado de acordo com o feedback recebido pelos jornalistas do JPN.

Agradecimentos

I would like to thank Prof.Carlos Soares and Jorge Teixeira for their constant feedback, availability and overall good humor. The JPN team, particularly Filipa, Isabel and Sérgio and also the students from the Journalism M.Sc. programme

I would like to thank my friends, in and out of FEUP, for always being there with a stupid joke.

Finally, and most importantly, i would like to thank my parents for all their patience and support throughout.

Bruno Vieira

"In God we trust.
All others must bring data"

W. Edwards Deming

Contents

List of Figures

List of Tables

Symbols and Abbreviations

API	Application Programming Interface
PoS	part-of-speech
TF-IDF	Term frequency - Inverse document frequency
SVM	Support vector machine
CRF	Conditional random field
LDA	Latent Dirichlet allocation
VSM	Vector space model

Chapter 1

Introduction

Analytics conducted on Twitter data have provided valuable insight on a wide variety of topics, be it politics [3] or marketing [4], as it allows to follow certain subjects and discern patterns, therefore supporting decision making based on statistically significant data.

Twitter is a microblogging service comprised of several million people [5] that interact through the use of small texts, images, videos and hyperlinks. Its interaction is achieved through quick messages, or "tweets", whose likeliness to a side-by-side conversation [6] is what gives Twitter its dynamic nature, resulting on a large flow of information being shared.

Twitter's large amount of diversified data provides informative and therefore useful information, having brought forth several studies that focus on both analyzing and visualizing said information. To ease information gathering, Twitter's API allows users to request access to its data, which can be done by request, using Rest API, or by event handling, where each tweet is considered an event, using the Streaming API. The access to these APIs motivated the creation of data gathering platforms, one being SocialBus [7, 8] which focused on aiding researchers.

TweeProfiles [1] began as a project to both analyze and visualize patterns of diversified data retrieved from Twitter, through means of the SocialBus [7, 8] platform, at that time known as TwitterEcho, and Twitter's RESTful API within the Portuguese twittosphere. It explored the gathered data on four dimensions, spatial, temporal, social and content, presenting different methods to approach the clustering process. The result was a platform that can display patterns on all domains, be it one-dimensional or multidimensional, by applying its clustering process offline. Further iterations sough to improve interpretability, in TweeProfiles3 [9, 10], as well as the ability to handle streaming data, in TweeProfiles4 [2].

1.1 Motivation

As the TweeProfiles project matures, new opportunities for further improvements become apparent. Previous iterations expressed the hardship, from media journalists, in interpreting the cluster's subject. The subject would allow the exploration to be oriented towards certain categories, filtering non-target data. However, each Twitter post is considered small and noisy, due to the maximum

size of tweets coupled with the use slang words, abbreviations and grammatical errors, which negatively impact interpretability. As of TweeProfiles4, data gathering is done in streaming, therefore providing large amounts of said information to be explored.

1.2 Research Goals

This dissertation aims to adapt the TweeProfiles tool to support media journalists. The goal is to assign labels to the TweeProfiles attained clusters, by applying Topic Extraction techniques, therefore providing media journalists additional means to conduct data exploration. The TweeProfiles tool is also to be adapted to the journalists method of news discovery.

1.2.1 Research questions

The research questions to be handled focus on two stages.

- Can the expansion of Twitter clusters' labels, using Topic extraction approaches, improve journalists experience on creating news articles from TweeProfiles?

- Could the Tweeprofiles' tool be adapted to assist the exploration methods of media journalists?

1.3 Document Structure

The state of art review is conducted on Chapter 2, presenting an overview of Twitter, the Tweeprofiles project and a general overview of Topic Extraction. In Chapter 3, Topic extraction methods are applied on an benchmark dataset and on a TweeProfiles gathered dataset. The tool and the applied methods are described in Chapter 4. Chapter 5 presents a discussion on the achieved results and future work.

Chapter 2

State-of-the-Art

This chapter aims to contextualize on the various aspects that a Topic Extraction task on Twitter would require. A brief overview of Twitter, its APIs and the TweeProfiles project is presented, followed by state-of-the-Art research conducted on cluster data visualization. On a later section, a general overview of Topic Extraction is presented, referring general principles and conducted work on Twitter.

2.1 Twitter

Twitter's social interaction is accomplished through the use of short messages that are shared to the user's current followers. These messages contain specific key parameters, shown in Table 2.1.

Table 2.1: Twitter parameters

Concept	**Description**
Retweet (RT)	Share another user's tweet
Mention (@ + username)	Identify a user in a tweet
Reply (@ + username)	Answer to a previous user tweet
Hashtag (# + topic name)	Association of a keyword to a tweet
Localization	User's geo-coordinates when sending a tweet

Although the parameters shown above, along with text and other multimedia fields, define a tweet, the scope of available information is far greater, as can be seen in Twitter's documentation [11]. Information relative to the user, such as the identification number, can be obtained while gathering tweets. This together with other attainable information, through further querying of Twitter's API, can be used to build relations, such as social graphs that intertwine users.

2.1.1 Twitter's APIs

Twitter holds vast amounts of diversified data, an aspect that incites an analytic conduct to discern patterns or trends on a given subset. For this purpose, Twitter allows its data to be accessed via means of APIs, namely Twitter's REST and Streaming API.

The REST API allows a request-base user-centered research, providing structured public information, such as timeline and followers. It requires an oAuth authentication to be accessed and enforces a rate limit policy. This policy dictates that requests must not exceed 15, or 180 depending on the method invoked, on a 15 minute window or 120 requests per hour. Failure to comply with the API's rate limit policy will result on an HTTP 429 "Too Many Requests" response code. It's abuse will blacklist the requesting account or app.

The Streaming API differs from REST by not allowing singular searches, but instead providing real time data, where each tweet is flagged as an event. An oAuth authentication together with a persistent open HTTP connection is required to incrementally parse the response, which is only rate limited by not implementing backoff strategies, such as reducing the rate of reconnect attempts given an unexpected connection lost. To accommodate different use cases for real time data, Twitter offers three streaming endpoints, which are the Public, User and Site streams and whose description can be viewed in Table 2.2.

Table 2.2: Twitter's Streaming API endpoints, taken from Twitter's Documentation

Public streams	Streams of the public data flowing through Twitter. Suitable for following specific users or topics, and data mining.
User streams	Single-user streams, containing roughly all of the data corresponding with a single user's view of Twitter.
Site streams	The multi-user version of user streams. Site streams are intended for servers which must connect to Twitter on behalf of many users.

An example of a platform that utilizes the APIs, to provide a considerable data set of Twitter data, is SocialBus.

2.1.2 SocialBus

SocialBus [7, 8] is a platform that continuously gathers social network messages, currently supporting both Twitter and Facebook, to aid researchers in today's need of vast quantities of data for knowledge inference. Messages are obtained from an established connection to the appropriate API, such as Twitter's Streaming API, which are then sent to a message broker for data format translation. The following step handles message processing in two phases, stream processing, for operations such as language detection and tokenization, and batch processing, to extract different kinds of knowledge. The results of the stream processing phase are stored in MongoDB for posterior analysis. The current architecture can be viewed in Figure 2.1.

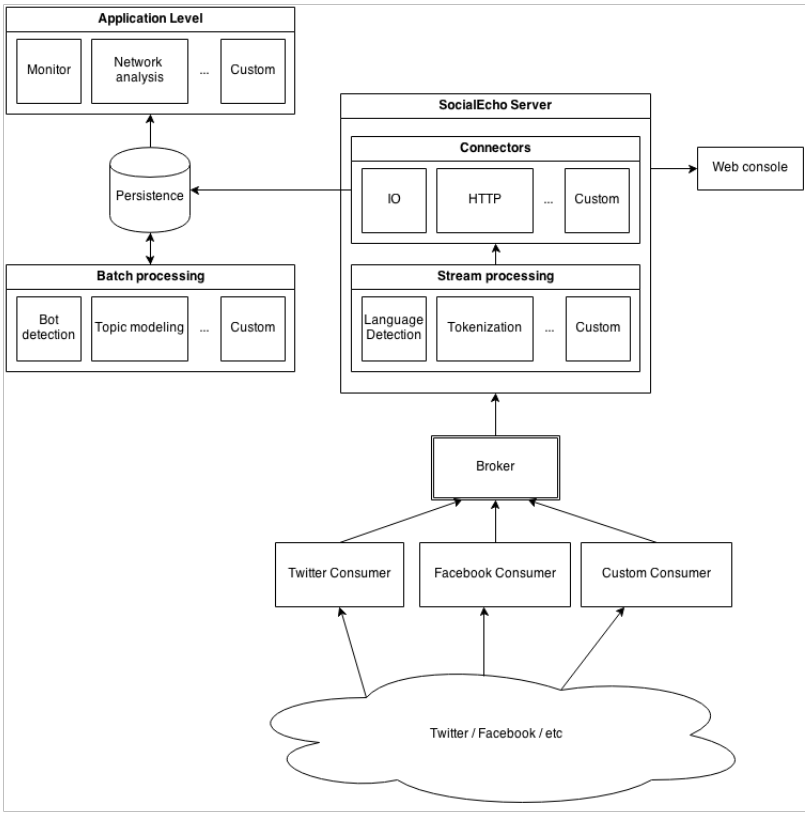


Figure 2.1: SocialBus system architecture reproduced from http://reaction.fe.up.pt/socialbus/

SocialBus's data gathering and pre-processing capabilities have proven to be reliable, due to its inclusion on the TweeProfiles project. Hence, the desire for its use to populate the database in the course of this thesis.

2.2 TweeProfiles

TweeProfiles [1] began as a platform to handle homogeneous and heterogeneous data, obtained from Twitter messages, while presenting a comprehensive representation of extracted information in aid of journalistic research. It focuses on tweet fields that are representative of the spatial, temporal, content and social domain. It obtains patterns on homogeneous data or a defined weighted combination, heterogeneous data, producing clusters with similar tweets. The density-based clustering algorithm DBSCAN(Appendix A) was chosen due to fulfilling requisites imposed by the system's domain and purpose, as in being able detect arbitrarily shaped clusters, not requiring the number of cluster beforehand, noise resistance in microblog messages and allowance of input of a dissimilarity matrix, generated by any distance function [1]. Each domain is analyzed separately by applying a suitable distance function, shown in Table 2.3 [1, 2], in order to attain dissimilarity

matrices, which the clustering algorithm requires and a combination of these matrices is used to obtain the multidimensional equivalent, e.g. spatio-temporal.

Table 2.3: Distance function by attribute type [1, 2]

Domain	Tweet Attributes	Distance function
Spatial	Latitude and Longitude	Haversine
Temporal	Timestamp	Timestamp difference
Content	Text	Cosine Dissimilarity
Social	Graph	Geodesic

The platform's visual interface presents the resulting clusters, displaying information relative to said cluster as well as other relevant data.

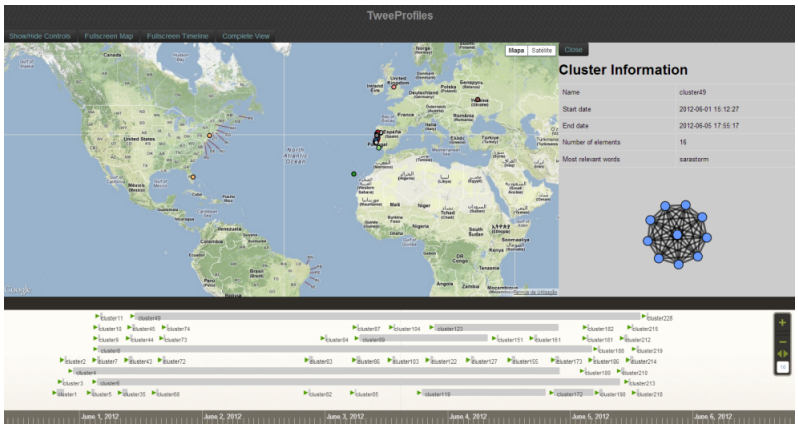


Figure 2.2: TweeProfiles interface showcasing temporal domain results and cluster information

Although successful, how it conveyed the resulting information was not deemed suitable, usability wise, hence TweeProfiles3.

TweeProfiles3 [9, 10] main contribution was a revamped user interface, designed with user feedback while employing appropriate frameworks to accomplish the desired user experience. The user feedback was obtained through inquiries and interviews to media professionals, with regards on previously used social media information gathering techniques and platforms, as well as expectations on what could the TweeProfiles platform provide, given its focus on Twitter data analysis. The resulting insight, of the aforementioned process, greatly benefited the project's use case defining. The user interface was accomplished by means of a php framework, Codeigniter, the Leaflet JavaScript API for user interaction and the Google Maps API for cluster visualization. The requested clusters could be handled on three dimensions, excluding the social domain from the original TweeProfiles as it was disruptive of its streaming capabilities. The outcome can be viewed in Figure 2.3.

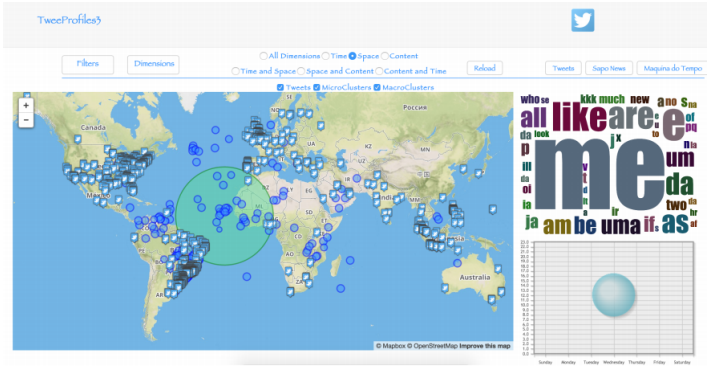


Figure 2.3: TweeProfiles3 interface

In addition to the aforementioned improvements, TweeProfiles3 integrated the SocialBus platform into the project's architecture, as it has been the means to which data is obtained from Twitter, therefore improving the project's data retrieval capability, granting greater control and insight on said operation, an important aspect for an upcoming iteration of the project, TweeProfiles4.

TweeProfiles4 [2] dwelt on the project's streaming capability, an endeavor approached on TweeProfiles second iteration, TweeProfiles2 [12], as well as providing a means to evaluate the clustering process. The challenges associated with user interaction coupled with the computational requirements, with handling homogeneous and heterogeneous data in real time, on an unpredictable user defined weight scheme led to an extension of the clustering process by splitting it into two phases, online and offline, as can be seen in its system architecture in Figure 2.4.

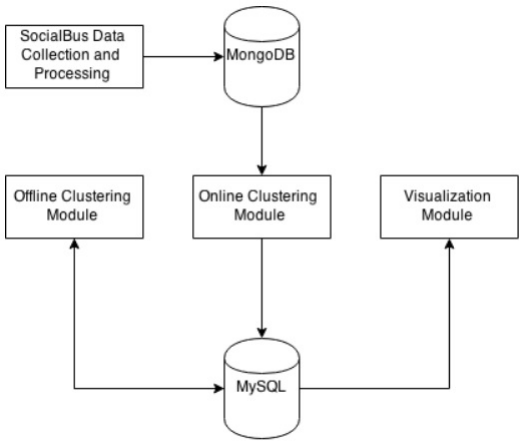
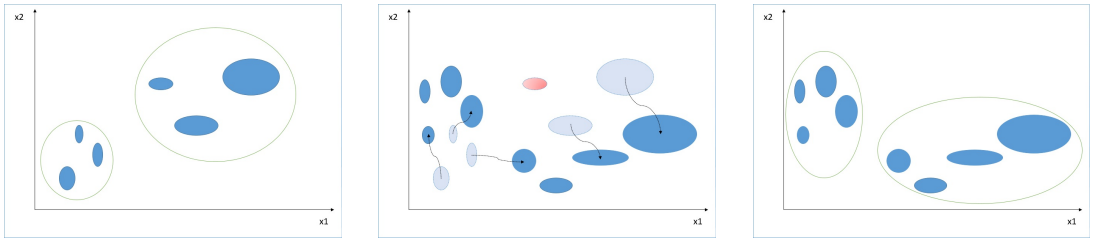


Figure 2.4: TweeProfiles4 system architecture

This approach allows handling the problem in different steps. The online phase is tasked to maintain an updated summary of each dimension, while receiving a continuous stream of data. The online phase achieves its goal by supplying the continuous stream of data to multiple clusterers, one for each spacial, temporal and content dimension. These apply the HybridDenStream

algorithm(Appendix B) thus providing micro clusters for each explored dimension. These microclusters are handled by an *OverlapManager* which is either tasked with maintaining updated microclusters, or, when prompted by an user request arrival, sent to the offline phase of the clustering process, which uses the DBSCAN algorithm to obtain macroclusters. An illustration of this procedure can be viewed in Figure 2.5.



(a) (b) (c)

Figure 2.5: Illustration of the micro and macroclustering evolution process (a) Instanced microclusters, in blue, with resulting macroclusters, outlined in green; (b) Microclusters being updated, showcasing new entries, cluster changes and removal; (c) New instanced micro and macro clusters after the update procedure

The evaluation procedure scrutinizes both microclusters and macroclusters to obtain internal and external measures. These are attained using the implemented distance functions, a requirement due to the absence of ground truth, together with a sample of tweets associated to each microcluster and then aggregated on resulting macroclusters. The visualization of the evaluation results was addressed with the development of an interface, using JavaScript chart libraries, to better discern the behavior of the evaluation measures on a time line base.

TweeProfiles4 delivered satisfactory results and greatly improved the overall project, with its performance optimization and especially the inclusion of the evaluation procedure.

2.3 Topic extraction

The analysis and pattern discovery in data sets holds a variety of sources and types of data which compelled the assimilation of different fields of study into data mining procedures. One such procedure is Text mining [13], the discovery of interesting knowledge in text documents. Commonly performed tasks involve document classification, clustering, summarization and concept identification to analyze textual information. These allow extracting a book's information for the search automation of libraries, analyzing a doctor's diagnosis transcripts and disease patterns. To this end, it employs a vast number of techniques, such as data mining, machine learning, natural language processing and information retrieval.

Topic extraction is a task that seeks to assign topics to documents, or document collections, that best describe them. The resulting topics can either be extracted from the content, a common practice, or automatically constructed, using additional information such as external sources. The topics also prove to be useful in accomplishing Text mining task [14].

The following sections aim to contextualize on corpora, data pre-processing, methods and evaluation performed on Topic extraction tasks, presenting definitions, approaches and conducted work.

A topic is characterized by [15] as being a sequence of one or more words, also known as keyword and keyphrase respectively, of high relevance in a document or collection. From a linguistic standpoint, a topic is usually defined as a noun, therefore a reference to an entity, or as a compound of nouns, verbs and adjectives, conveying an action or characteristic. Topic keywords and keyphrases also differ on how they depict the overall context, as keywords generally convey lower insight. While defining what is allowed to be recognized as a topic, rules, governed by the concern of keeping candidates to a minimum, should be enforced, as it might otherwise result on a strenuous process for long documents. As stated by [14], these rules are based on heuristics.

2.3.1 Corpora

A Corpora refers to multiple text corpus, a set of structured documents. Its study focus on document structure knowledge, specific to each source be it ranging web pages or scientific articles, providing generally fitting assumptions. There are four factors, described by [14], that detail a corpora, them being length, structural consistency, topic change and correlation.

Length influences topic extraction, as is the hypothesis that longer documents yield more candidate topics.

Structural consistency refers to the likely position of a topic on a document. Its premise requires the document to follow a standard structure, as is the case with scientific papers in which topics are likely to appear in the abstract or introduction. The lack of structure reduces the information usefulness, being most apparent in documents composed of web pages or forums.

Topic change also takes advantage of a documents structure by inferring where a topic should appear. A common observation from scientific and news articles is that the topic not only appears in the beginning, but also at the end. However, such observations do not hold true for documents of conversations. This is due to possible topic changes with time, thus topic change detection approaches in [16].

Topic correlation refers the possibility of topic relations, which again, holds more ground on news and scientific articles, which are normally structured documents.

Research for different types of corpora have many options in obtaining a dataset, be it by either querying Google or through repository websites. Some examples of websites that allow such gathering, including news, scientific articles and text, are DMOZ - the Open Directory Project[1], CiteSeer[2], Reuters[3], JAIR[4] and ACM Digital Library[5].

The term Corpus may also refer to a set of documents containing language information, a collection of words, expressions or stop words deemed representative of that language, e.g. the

[1] http://www.dmoz.org
[2] http://citeseerx.ist.psu.edu/index
[3] http://www.daviddlewis.com/resources/testcollections/reuters21578/
[4] http://www.jair.org/
[5] http://dl.acm.org/

Brown Corpus[6] for American-English. These documents can be used by topic extraction applications to provide additional sources of managing topic candidates, to validate or remove based on presence in the language corpus. Stop words alone lack descriptive qualities but might heighten insight on keyphrases, i.e. the difference between "University" and "Porto" versus "University of Porto", and utilized method, as the algorithm might be expected to correctly handle them [17]. The language's corpora coverage may also affect the resulting topics, as seen in [18] comparing the impact of a full, general and general with domain-specific on the results.

The task of evaluating relies on Ground truth, a term attributed to information gathered from direct observation. On the subject of Topic extraction, it portraits the topic that best describes a document and, therefore, the one to which extracted topics are evaluated against. Common methods of attaining a corpus ground truth involve either, the employment of human annotators, extracting the title from a hierarchical document structure and the query used on a search engine.

The use of human annotators was conducted in [19], employing both author and reader topics. This is a necessary conduct for unannotated or new documents, although requiring individuals, with the necessary background, to perform the task.

On already annotated corpora, the extraction of a title is a common practice given an hierarchical organization. This is due to information being presented with different levels of detail, with a general area on top and a more specific one bellow. In [17], ground truth for web pages are obtained from the assigned labels of each category in the DMOZ repository. Usage of a query on a search engine, e.g. Google, as a ground truth topic has also been performed in [15], requiring that the topic be in content due to the method in which Google provides search results.

In [20], three text classification APIs where used to obtain the corpora ground truth, namely the Alchemy API[7], OpenCalais API[8] and Textwise SemanticHacker API[9].

2.3.2 Data pre-processing

Data pre-processing is a required preliminary step in a data mining process. A procedure conducted on raw data to handle its inconsistency in quality, due to noise originated from data-gathering methods, as well as the need to normalize said data followed by feature extraction. In a text mining application, pre-processing tasks are used to handle text information, consisting of word tokenization followed by normalization techniques, such as lemmatization and stemming, stop-word handling and Part-of-speech(PoS) tagging.

Word tokenization is a lexical analysis process of identifying words and phrases with subsequent attribution of a token. This is accomplished by following heuristic rules on a word level, by acknowledging commas and punctuation marks as word or phrase terminators. The resulting tokens are then used for further processing.

[6] http://clu.uni.no/icame/brown/bcm.html
[7] http://www.alchemyapi.com/api/
[8] http://www.opencalais.com/documentation
[9] http://textwise.com/api/categorization

Normalization tackles the need for a unified form, which in text translates to a single canonical form. Lemmatization is a technique that reduces variant forms to a base form, the lemma of a given word. Stemming differs from the latter by operating on each word disregarding context, therefore producing word stems that might not correctly transcribe meaning. Although lemmas convey more accurate word representations, stemming is usually a faster and easier technique to implement, such as the Porter stemmer.

Stop words can, as was previously discussed, either be left or removed from a corpus. The most common solution, removal, can be performed using a language appropriate corpus or stop words can be inferred from a document's properties. In [21], methods that explore the document take into account term frequency, based on Zipf's Law, mutual information between a term and a document class, and measures of divergence, KullBack-Leibler, on random samples. The impact of stopword removal in clusters was evaluated in [22], stating that a custom stopword list provides better clustering results.

PoS tagging aims to assign each word in a corpus its designated part of speech, such as noun or verb, using both the word's definition as well as its context. This process allows subsequent steps to extract PoS patterns to characterize candidate topics.

The process of feature selection, in which words and phrases are dimmed candidate topics based on the features that characterize them. Its scope extends to two main categories, within-collection and external resource-based. Within-collection features further branches into subsections that tackle corpus knowledge, namely, statistical, structural and syntactic features.

Statistical features are obtained through calculations performed on corpus. This information has been studied trough means of frequency and occurrence both within as well as outside the corpus. The tf*idf measure is achieved by computing term frequency together with inverse document frequency, thus elevating candidate topics that have both a higher frequency in a corpus and lower on other. The first occurrence, as well as the distance between the words, provides information on the word position, an important aspect if the assumption that a topic should appear early holds. In [23], not only are the frequency of topics obtained, but also their pairwise co-occurrence.

Structural features refer to the location of a topic candidate in a corpus. As topics on structured document sources, such as scientific articles, are likely to appear in a given location, the frequency of candidates on said location may assist in topic appraisal.

Syntactic features provide grammatical information for each topic candidate. This is accomplished by inferring the PoS tag or suffix sequence assigned to the candidate. However, in [14] it is stated that, in web pages and scientific articles, such information may prove to be unhelpful in the presence of other feature types.

Knowledge acquired from sources besides the corpora is designated as an external resource-base feature. These approaches perform queries on search engines and repositories, such as Wikipedia, exploring the results to ascertain the topic's salience.

2.3.3 Methods

Topic extraction methods have two branches, manual assignment and automatic extraction. Manual assignment is conducted by human annotators, a practical approach to the issue, however, only viable on single or low corpus count. Automatic methods, on the other hand, are cost effective with higher exploration capabilities. In [24], a compilation of automatic topic extraction methods are divided into four categories, namely, Statistical, Linguistic, Machine Learning and other approaches.

The advantages of a Statistical approach comes from the methods not requiring training data and, simultaneously, being language and domain independent. Candidate topics statistics, from a corpus, that are gathered involve, TF-IDF measure, χ^2 test and word co-occurrence. The TF-IDF measure, as is presented in [25], is computed in the following fashion, tf in Function2.1; idf in Function2.2 and the resulting tfidf in Function2.3.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2.1}$$

Where:

$n_{i,j}$: the number of occurrences of term i in document d_j

$\sum_k n_{k,j}$: the number of occurrences of all terms in document d_j

$$idf_i = \log \frac{|D|}{|d_j : t_j \in d_j|} \tag{2.2}$$

Where:

$|D|$: total number of documents in the corpus

$|d_j : t_j \in d_j|$: number of documents where the term t_i appears

$$tfidf_{i,j} = tf_{i,j} \times idf_i \tag{2.3}$$

Term frequency denotes the term's presence in a document and the inverse-document frequency the presence between documents.

A Linguistics approach utilizes language driven features, such as lexical, syntactic and semantic, to characterize candidate topics. In [26], a lexical analysis is conducted by scoring lexical chains, constructed trough means of WordNet[10].

Machine Learning models can be inferred by algorithms using training data, supervised, while generally also being domain dependent. Methods used in this approach include, SVM [27], CRF [18], LDA [28]. As is common from algorithms that build models, as domain changes so must the model be re-learned.

[10] https://wordnet.princeton.edu/wordnet/

The LDA method presents mixtures of terms, where multiple terms constitute a topic. It assigns a document probability to the number of desired topics and then randomly assigns the document's words to these topics.

Other approaches are usually a combination of previous methods and features, performed to include knowledge such as heuristics. In [19], a limitation of the TF-IDF measure is presented, stating that it produces better results in clustering and classification tasks then on topic extraction. The proposed solution to this predicament was the inclusion of a boosting factor, portrayed in Function2.4.

$$B_d = \frac{|N_d|}{|P_d \times \propto|}, \quad \text{if } B_d > \sigma \quad \text{then } B_d = \sigma \tag{2.4}$$

Where:

$|N_d|$: number of all candidate terms in document **d**

$|P_d|$: number of candidate terms whose length exceeds one in document **d**

$\propto$ and σ: weight adjustment constants

This factor is used with the TF-IDF measure as is shown in Function2.5, noting that the term position feature is also used, although not mandatory.

$$w_{ij} = tf_{ij} \times idf \times B_i \times P_f \tag{2.5}$$

Where:

w_{ij} : weight of term t_j in Document D_i

tf_{ij} : frequency of term t_j in Document D_i

idf : $\log_2$ N/n where N is the number of documents in the collection and n is the number of documents where term t_j occurs atleast once. If the term is compound, n is set to 1.

B_i : the boosting factor associated with document D_i

P_f : the term position associated factor. If position rules are not used, this is set to 1.

Algorithms such as VSM and Graph-based also present models appropriate for text representation. The VSM represents documents as vectors of identifiers and uses a term's weigth, such as the TF-IDF measure, to compute document-query similarity. Where as VSM is most suitable for capturing single word frequency, Graph-based models explore relationships and structural information.

Pagerank is a graph based solution and it uses the terms frequency in documents and co-occurrence with other terms on the collection to attain their relevance. Their relevance is elevated by the terms frequency in a document and its appearance on documents of the same collection.

As was previously stated, methods can be categorized as either supervised or unsupervised. In supervised approaches, training data is supplied to tune the model, thus improving performance. Disadvantages of this approach come from the requirement of topics being manually annotated on

the training data. This is then coupled with bias towards the domain which they have been trained. Unsupervised approaches, on the other hand, do not require labeled data.

2.3.4 Evaluation

Evaluation of the results through metrics is a common practice in data mining. The metrics used, however, must be proper to the target data and task. In topic extraction, the score of systems can be computed by employing either human or automatic evaluation. Human evaluation poses the same dilemmas of Human annotators, as in availability, cost and ambiguity. The typical approach, as is stated in [14] and performed by the SemEval-2010, is to create a mapping between the ground truth topics and the resulting topics using exact match, then scoring based on precision, recall and F-score. In [18], topic extraction is defined as a classification task and evaluation calculations are presented for precision in Function2.6, recall in Function2.7 and F1-Measure in Function2.8.

$$Precision = \frac{truepositive}{truepositive + falsepositive} \tag{2.6}$$

$$Recall = \frac{truepositive}{truepositive + falsenegative} \tag{2.7}$$

$$F1\text{-}Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2.8}$$

Another evaluation measure is presented in [15], overlap in Function2.9 and precision in Function2.10. Overlap is a measure of similarity between each extracted topic to the predefined topic of the cluster. Precision indicates how the extracted topic that best fits the predefined topic is ranked.

$$overlap(p_i, p_t) = \frac{|p_i \cap p_t|}{|p_i \cup p_t|} \tag{2.9}$$

Where:

p_i : extracted topic

p_t : predefined topic

$$precision(p^k, p_t) = overlap(p_{max}, p_t) \times \left[1 - \frac{rank(p_{max}) - 1}{k}\right] \tag{2.10}$$

Where:

$p_{max} \in p^k$: is the first topic with maximum overlap in the top k topic list

$rank(p_{max})$: is the rank of the top k

Other measures, Exact and Partial match, have been computed in [17].

2.3.5 Topic extraction conducted on Twitter

A summarizing approach to Twitter's content has been presented in [29]. It obtains it's dataset using Twitter's Rest API from Singapore users. For the data pre-processing step, the removal of stopwords was conducted. To extract topics, a LDA model was applied, which assumes a single topic assignment for each tweet. On each candidate topic, a topical PageRank algorithm is used to rank keywords and then use the top ranked to generate keyphrases. Ranking keyphrases is conducted by applying a probabilistic model. This uses a hypothesis based on *Relevance* and *Interestingness*. To evaluate results, an adjustment was performed on the *nDCG* metric from information retrieval.

Chapter 3

Topic extraction on twitter data

This chapter provides an analysis of the applied Topic extraction approaches on a Twitter dataset. Before testing the methods on the Twitter dataset, some experiments were made on a simple, benchmark dataset, created to illustrate the differences between the methods. The Topic extraction approaches are then performed on both datasets, whose results are interpreted in regards to each algorithm and agreement.

3.1 Experimental setup

This section presents the benchmark dataset and scrutinizes a set of micro-clusters, with affiliated tweets and language, gathered from the TweeProfiles platform. The experiment's intent is to provide an understanding of the data, such as cluster size distribution and number of candidate topics encountered.

The benchmark dataset analysis attempts to concisely illustrate the differences between each Topic extraction approach. The actual data set, gathered from the TweeProfiles tool, provides a reference to the overall handled size and topic distribution to be handled.

3.1.1 Benchmark dataset

The benchmark dataset was made to have specific characteristics to better perceive the applied Topic extraction approaches. The desired characteristics focused heavily on word frequency, in each tweet and between them. As the intent is to summarize tweet groups, these were coupled with the absence of stopwords and all tweets were attributed to a single group. The designed dataset can be seen in Table 3.1.

Table 3.1: Corpus example

Tweet 1	primeiro primeiro teste
Tweet 2	segundo teste criado
Tweet 3	teste terceiro criado

The dataset presents a low word count, enough to have interesting term frequency properties yet reduced number of extracted n-grams to ease evaluation. It also has same word occurrence in all tweets and co-occurrence in a single tweet. The n-grams, unigram to trigram, were then extracted and its relation observed in Figure 3.1.

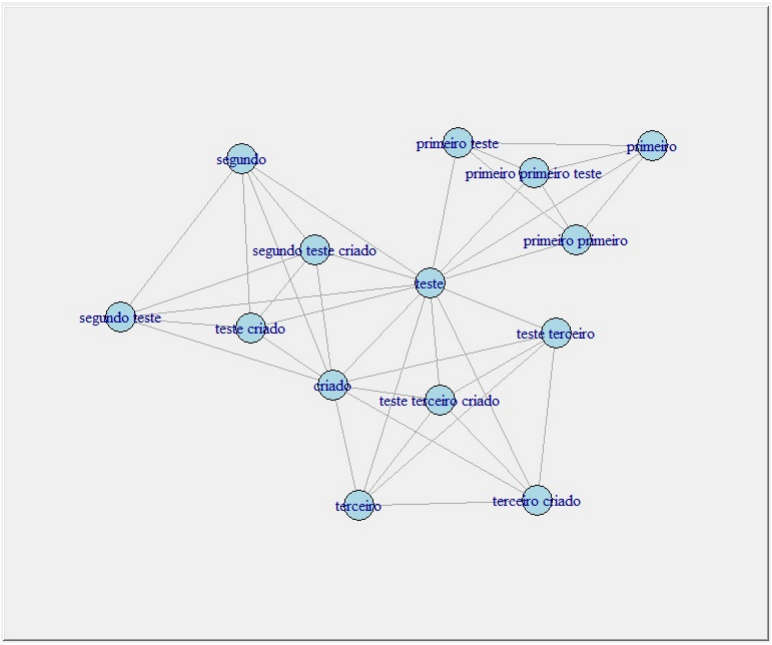


Figure 3.1: N-gram relation graph

In the previous Figure, each node denotes a n-gram in a document and its term connectivity is shown by the connected lines. Its inspection demonstrates the high connectivity of the word "teste", a common word, as opposed to a less connect word such as "primeiro". This graph, however, does not represent word frequency.

3.1.2 SocialBus dataset

The dataset is comprised of tweets, with corresponding micro-cluster identifiers and language, obtained from the TweeProfiles platform on June 18 2016. This instance was gathered in approximately 1 hour, restricted to tweets that possess both geolocation and language attributes and persisted on a MySQL database. These entries were stored for micro-cluster processing and evaluation purposes, and therefore have been limited to a maximum size of 100 tweets per micro-cluster. The number of tweets reaches 208 and are attributed to 46 unique micro-clusters, a distribution that can be seen in Figure 3.2.

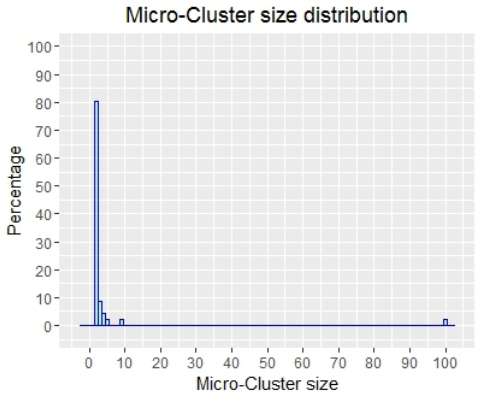


Figure 3.2: Micro-Cluster size distribution

The inspection of the previous distribution indicates the predominance of micro clusters with two tweets, 80%. The remaining micro clusters vary in size, of which one holds the maximum allowed size.

The dataset's language distribution consists of 90% English and 10% Portuguese tweets, with the overall text holding 712 unique words, after stopword and singleton removal.

3.2 Exploratory data analysis

The applied Topic extraction algorithms were the unsupervised TF-IDF, Pagerank and LDA methods. These have been thought to explore and weight the different types of data, ranking n-grams according to different interpretations of a topic. The n-grams are extracted by breaking down the words in each tweet, set to vary from one to three words. This is performed without PoS stipulations, due to varying nature of tweet sizes.

The TF-IDF method uses the calculated measure to order the result. This ranks the n-grams by elevating those that are frequent in a tweet, yet sparse on the collection, cluster. These ranks do, however, show undesirable results on tweets that share the same nature, i.e. if two tweets are about jobs offerings, the topic "job" would rank lower.

The Pagerank method regards topic importance differently then the TF-IDF, ranking n-grams regarding their relatedness in the scope of the collection. It ranks by exploring their frequency in tweets and co-occurrence with other n-grams on the collection. Each n-gram is ranked higher the greater its frequency in a tweet and its appearance on tweets of the same collection. Although capable of suitably handling the aforementioned TF-IDF problem, unrelated tweets would be ranked accordingly, resulting in what could be considered common and uninteresting n-grams being elevated.

Finally, the LDA method differs from the previous algorithms by not solely ranking individual n-grams but also presenting mixtures of these as collection summarizing topics. It requires an imposed number of topics for the algorithm to discover and the number of n-grams that would compose each of them. These impact the task's performance, since requesting a configuration

too small for the collection would result on topics lacking the desired insight. Topic distribution discovery can be achieved in two ways, using a Variational Expectation Maximization (VEM) algorithm and Gibbs sampling. The prior uses a maximum-likelihood estimate while the latter a Markov-chain Monte Carlo method to perform inference. On performance, in [30] an empirical study on both algorithms was conducted were the Gibbs sampling was said to converge more rapidly to a known ground-truth, thus it was chosen for the current task. The ranking of its topics has been performed by use of each topic's probability assigned to each tweet. As a mixture of n-grams from the collection, each topic has a degree of relatedness to each tweet, a topic-document distribution value, thus those with higher overall score where chosen to be ranked higher. This ranking method is shown in equation 3.1.

$$\underset{i}{\operatorname{argmax}} \sum_{j} {w_j}^i \tag{3.1}$$

Where:

${w_j}^i$: n-gram with index j in Topic i

A more in-depth analysis for ranking topic significance of LDA topics can be viewed in [31].

3.3 Results

This section portraits the results obtained on the benchmark dataset and details further the results from the SocialBus dataset. The latter results are also used to perform an agreement analysis on the top@N topics.

3.3.1 Benchmark dataset

The following Table 3.2 displays the ordered topics obtained from the benchmark dataset.

Table 3.2: Sorted resulting topics

Order	**TF-IDF**	**Pagerank**	**LDA**
1	primeiro	teste	primeiro
2	primeiro primeiro	criado	teste terceiro criado
3	primeiro primeiro teste	primeiro	criado
4	primeiro teste	teste criado	teste
5	segundo	segundo	primeiro primeiro
6	segundo teste	segundo teste criado	terceiro
7	segundo teste criado	segundo teste	segundo teste
8	terceiro	terceiro criado	primeiro teste
9	terceiro criado	terceiro	segundo
10	teste criado	teste terceiro	primeiro primeiro teste
11	teste terceiro	teste terceiro criado	terceiro criado
12	teste terceiro criado	primeiro teste	teste criado
13	criado	primeiro primeiro	segundo teste criado
14	teste	primeiro primeiro teste	teste terceiro

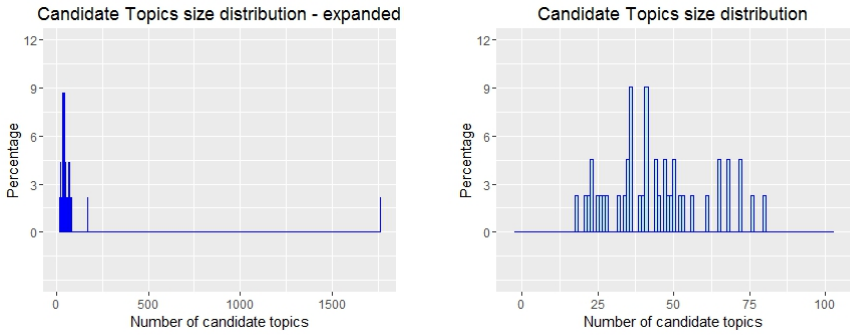
The n-gram extraction method attains a higher number of topics in comparison with the number of words. This makes the difference between two consecutive topics in the ranking to not be significant. On a closer inspection of the TF-IDF and Pagerank ordering, their unique ways of exploring become apparent. The TF-IDF significantly lowered the rank of the topics "criado" and "teste", due to their presence on multiple tweets, and elevated the co-occurring topic "primeiro". This was not the case for the Pagerank results, displaying the opposite for the first two mentioned topics, only then displaying the latter. Although a human annotator would likely rank "teste" as an important cluster topic, this is due to the relatedness of the tweets.

The LDA was performed with three attributed n-grams to each topic. The number of topics chosen was 35, due to previous topic distribution analysis performed on actual datasets. On inspection of the top three ranked, both "primeiro" and "criado" are frequent terms on the collection, contrary to "teste terceiro criado". The top three topic results presented have different traits. These are representative of the collection, by elevating frequent terms and exploratory, by elevating a term not recognized as relevant by the other methods.

It was concluded that the three methods applied provide collection representative topics, ranked accordingly. The LDA would, however, require the display of multiple terms.

3.3.2 SocialBus dataset

The dataset was inspected in regards to its candidate topics size on each micro-cluster. This distribution can be seen in Figure 3.3.



(a) (b)

Figure 3.3: (a) Candidate Topic distribution; (b) Closer inspection without outliers

The micro-clusters topic distribution in Figure 3.3a shows an outlier, far from the standard distribution, with 1764 candidate topics. This is attributed to the micro-cluster with 100 tweets, which is not representative of the gathered collection, as is seen in Figure 3.2 and 3.3b. The latter distribution presents a median value of 42.5 topics for each micro-cluster. This value differs from the previously stated value of 35, that would be requested by the LDA, however it is the lower quartile value.

The applied methods rely heavily on the statistical features of each topic for weighting. This led to the analysis of sparsity values on the dataset, whose calculation can be seen in equation 3.2.

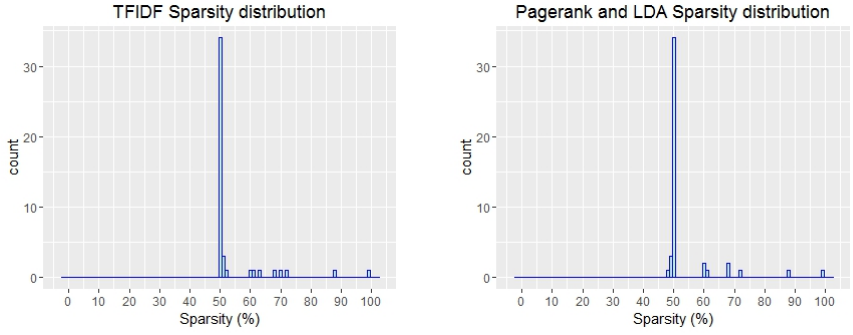
$$1 - \frac{\text{length of } wv_i}{\prod \text{TDMdim}} \times 100 \tag{3.2}$$

Where:

wv_i : weighting value i vector

$TDMdim$: Term-Document, or inverse, matrix dimensions

This distribution aims to illustrate how the methods perceive the data and can be seen in Figure 3.4.

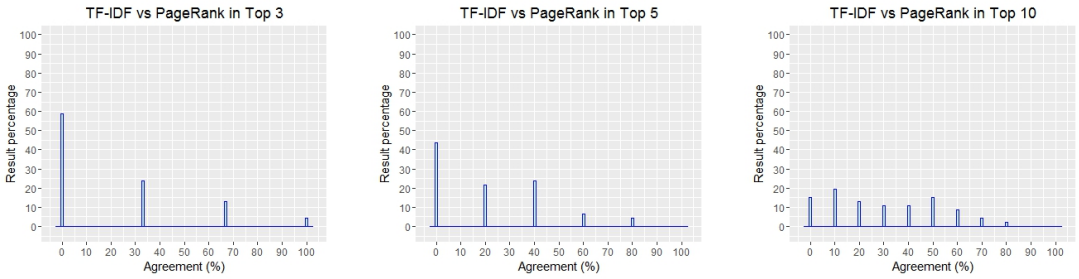


(a) (b)

Figure 3.4: Corpus sparsity distribution (a) TF-IDF sparsity distribution; (b) Pagerank and LDA sparsity distribution

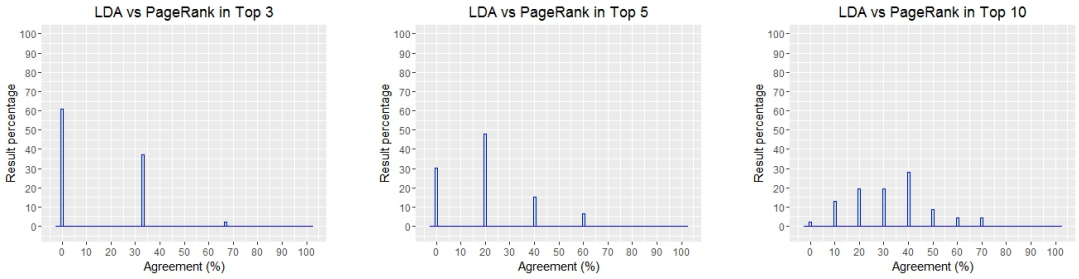
The separation between TF-IDF with Pagerank and LDA is due to the required weighting of these methods, before method specific operations. The result is, however, similar and displays the predominance of micro-clusters whose corpus have 50% sparsity. This indicates that, generally, the tweets in micro-clusters do not have overlapping terms, that is a word is not repeated.

The results of each method were further evaluated in regards to their agreement on the top ranking topics. To this end, portions of the results, top@N, were compared. The existence of a topic on the top@N of two methods was considered a positive agreement. The following Figures 3.5, 3.6 and 3.7 display the results on top@3, top@5 and top@10.



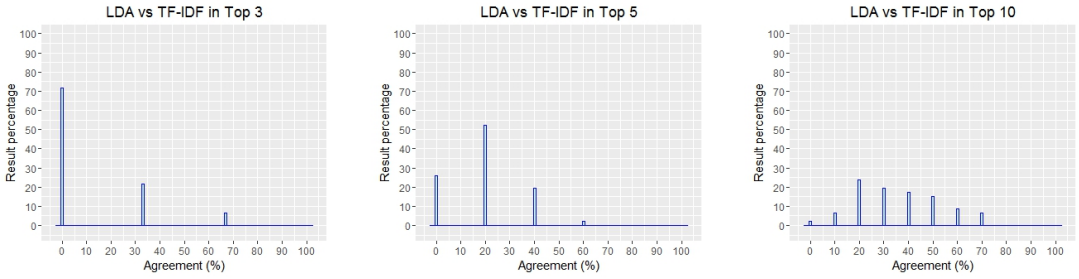
(a) (b) (c)

Figure 3.5: TF-IDF vs Pagerank (a) top@3; (b) top@5; (c) top@10



(a) (b) (c)

Figure 3.6: LDA vs Pagerank (a) top@3; (b) top@5; (c) top@10



(a) (b) (c)

Figure 3.7: LDA vs TF-IDF (a) top@3; (b) top@5; (c) top@10

The top@3 analysis displays an expected agreement distribution, roughly 60 to 70% disagreement in the first three topics. This result is consistent with the understanding of the applied algorithms, which also takes into account the corpus sparsity distribution. On inspection of the TF-IDF - PageRank distribution, it is noted a complete agreement on approximately 4%, 2 in 46, of the micro-clusters. One of these holds two tweets, that after pre-processing operations are presented in Table 3.3.

Table 3.3: Corpus example

Tweet 1	kansascity mo hospitality job housekeeper marriott kansas city airport veterans jobs hiring careerarc
Tweet 2	first cup coffee three weeks double espresso go slow slow coffee bar

The result of both TF-IDF and Pagerank were "coffee", "slow" and "airport". The first two were expected, due to their frequency in Tweet 2. The third topic was attained simply due to the sorting algorithm, as it holds the same weight as all other n-grams. Same weight n-grams are ordered alphabetically. The remaining analysis, top@5 and top@10, displays the expected tendency of requests with a higher size, higher values of agreement are achieved.

Chapter 4

Tweeprofiles for journalism

This chapter describes the tools used and the implemented procedures to achieve the proposed goals. The choice of software is the one in the current implementation, however the integration uses standard mechanisms(APIs). The sections cover both the back-end and front-end development as well as the results.

4.1 Development

The development stage describes the TweeProfiles tool capabilities and the implemented changes. The tool's capabilities are used to understand the data collection step, the analytic process and application interface. The implemented changes adapt the procedures to accommodate the Topic extraction tasks and desired usability.

4.1.1 Data collection and management

Data collection and management is performed by the SocialBus tool located on the TweeProfile's server. This is a standalone java application crawler of Twitter's tweet stream. The tweet stream is analyzed and persisted on a MongoDB database, should it have the required parameters. These parameters consist of user identification and geolocation, reflecting the multidimensional nature of the tool, of which the need for spatial information hinders the data gathering size. The main steps are illustrated in Figure 4.1.

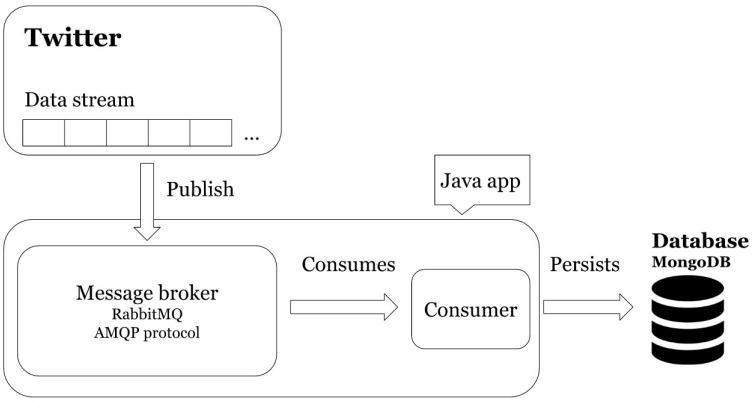


Figure 4.1: Twitter consumer

This procedure stores the gathered tweets on a MongoDB database, which is also used by the analytic server.

4.1.2 Analytics server

The analytics in the TweeProfiles tool is performed by a separate java application. This application handles both the clustering of the tweet stream, as well as the final clustering process, whose results are to be presented.

Stream clustering is an ongoing step, whose tweets are obtained from the MogoDB database. These are processed in batches, with a pre-defined minimum and maximum number of tweets, 2 and 100 respectively. The acquisition of a full tweet batch marks the start of the clustering process and also the removal of tweets gathered on the MongoDB. As was previously stated, this is a data subset, which is then further diminished due to language restrictions, as only Portuguese and English tweets are to be considered. The language is detected by a process in the consumer and is estimated to diminish the gathered dataset to 30 - 50%. The HybridDenStream algorithm operations are later applied to this data batch, thus creating micro-clusters. The resulting micro-clusters are the product of both the new tweet batch as well as their overlap with previous micro-clusters. An illustration of this process is shown in Figure 4.2, where the time to fill a batch is considered an instance.

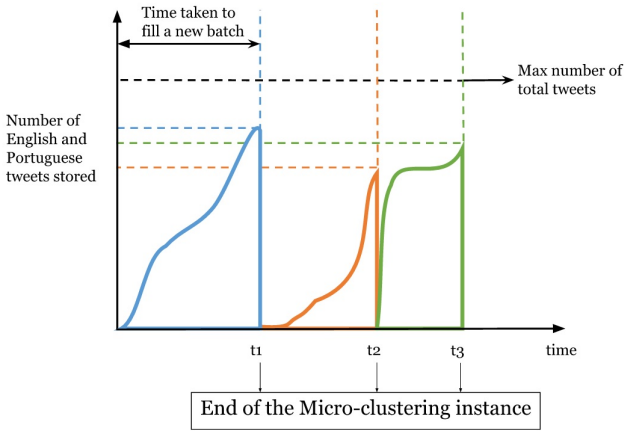


Figure 4.2: Micro-cluster size distribution

Variations to the maximum requirement of tweets were performed, noting the impact on micro-cluster update time. The results shown in Table 4.1 were noted to provide micro-clusters that maintained similar size distributions. The tool was then set to gather batches of 300 tweets, as opposed to the initial 5000.

Table 4.1: Micro-clustering instance elapsed time

Max data size	Average time (hours)
5000	12
2000	6
300	0.5

The details of the tweets, micro-clusters and their relation are persisted in a MySQL database. To those details, the tweet's language information and its unprocessed text was additionally stored, for Topic extraction procedures. The stream clustering step is illustrated in Figure 4.3.

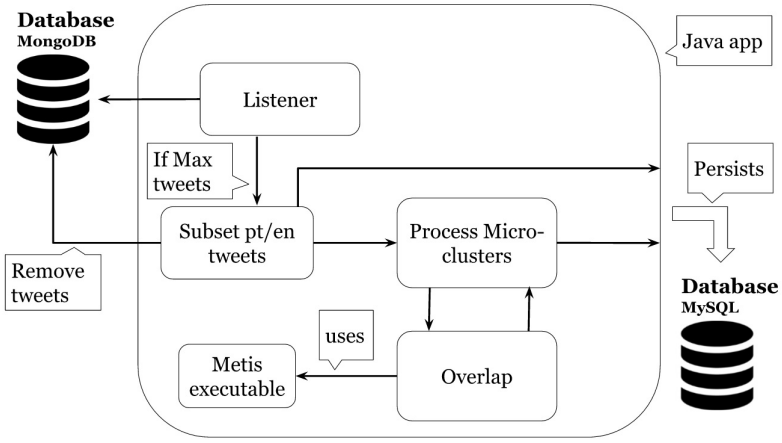


Figure 4.3: Micro Cluster creation procedure

The final clustering process refers to the result of the DBSCAN algorithm, macro-clusters. These are composed of the previously computed micro-clusters, whose similarity is weighted based on user requested dimensions. These requests are handled by a HTTP server application, functioning on parallel with stream clustering.

The need to relate micro and macro-clusters, later to be used by the App server, led to process changes which allowed them to be persisted in the database.

4.1.2.1 Topic extraction task

The Topic extraction task is performed by a script initiated by the Analytics server, after micro-clustering database persistence. The installed R version was the recent 3.2.4 that, although not supported for the Debian 7 version, allowed the used of updated packages. Its execution is called by the stream clustering application at the end of each instance. The task is performed in three main steps, namely the setup, the algorithm execution and a finalizing step, as shown in Figure 4.4.

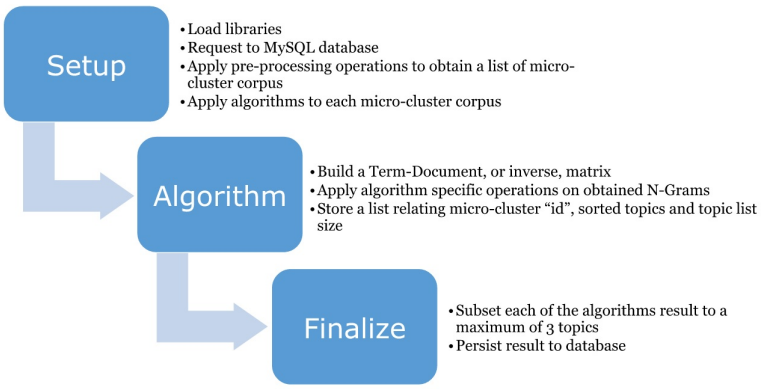


Figure 4.4: R script overview

The setup is comprised of operations that retrieve and prepare the dataset. The obtained dataset is put trough a two stage pre-processing operation, a text oriented revision and corpus processing.

The first pre-processing operation focuses on the text quality and building the corpus of each cluster. The retrieved dataset is filtered continuously for the text and language of a requested cluster. The resulting text is then analyzed for undesirable content to be removed. The common tweet contains a variety of special characters, mostly @ and #, hyperlinks, punctuation and numbers. The use of emoticons is also a common practice, however these were replaced by question marks, due to encoding. These special characters have an impact on topic discovery, as they would obscure other more desirable topic candidates. This process is not without flaws, as the candidate topic ***"S&P500"*** would be discarded due to both the ampersand and 500 being striped. The resulting texts, together with the clusters identifier and language, are then used to create a corpus object. These steps are shown in Figure 4.5.

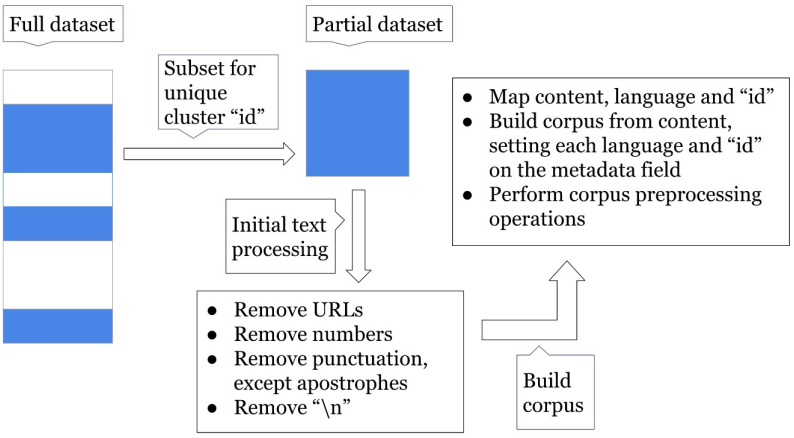


Figure 4.5: Cluster corpus building

The second pre-processing operation focuses on language oriented operations. The previously created corpus is iterated and its tweets pre-processed according to their specified language, Portuguese or English. Tweets of both languages have their stopwords, apostrophes and single words removed. However, an exception was made for the Portuguese tweets by also removing a few select English stopwords thought to be prominent, given the trend of the gathered tweets. This was due to a regular lack of success in estimating the tweets correct language, which could have been caused by lack of content. This behavior can be seen in Figure 4.6.

	id	text	lang
84	53	Frango na cerveja com creme de cebola!!!!!! @ Cohab…	pt
103	1	I'm at Praça de Eventos Fábio M. Paracat in Boa Vista, …	pt
104	74	I'm at Praça de Eventos Fábio M. Paracat in Boa Vista, …	pt
109	79	Animação pra subir pro ensaio zero ????????	pt
125	99	#cheers to #Finland #Finlandfunland #ladsontour #l…	pt
131	1	2º round ???? (@ Universidade Iguaçu (UNIG) in Nova I…	pt
132	70	2º round ???? (@ Universidade Iguaçu (UNIG) in Nova I…	pt
178	1	Lindo saaaaabado	pt
179	1	#PolishopRumo2020BH Mais um dia de evento ... @ E…	pt
180	97	#PolishopRumo2020BH Mais um dia de evento ... @ E…	pt
184	1	Recomendo que façam o mesmo	pt

Figure 4.6: Subset of Portuguese tweets gathered on June 18, 2016

An additional considered step was stemming and lemmatization to attain a word's common base form. Although stemming was applied, some of its results did not reach an accepted quality or were even harmful. For this reason, the application of the gathered packages, that performed lemmatization, was postponed and therefore not properly studied. For the previously stated reasons, these steps were left out.

The modified content is returned in a list format and is then used for corpus content replacement, as is shown in Figure 4.7.

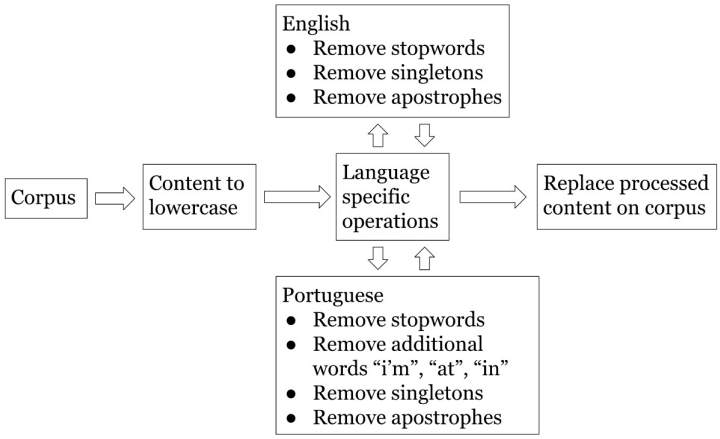


Figure 4.7: Corpus processing

The algorithm step performs the TF-IDF, Pagerank and LDA on the resulting corpus list. These start by relating extracted n-grams to tweets in a Term-Document matrix with a weight value, as seen in Table 4.2.

Table 4.2: Term-Document matrix

	Doc_x	Doc_y	Doc_z
$Term_a$	w_1	w_2	w_3
$Term_b$	w_4	w_5	w_6
$Term_c$	w_7	w_8	w_9

The n-grams discovery was set to range from one to three words, without PoS stipulations. This was due to the varying length of candidate topics in a tweet, possibly possessing one or even zero topics. The weight values type vary in each algorithm, as explained next:

- **TF-IDF**

The TF-IDF builds a Term-Document matrix with TF-IDF weight values. Its terms and values are extracted and related in a new matrix, seen in Table 4.3.

Table 4.3: Term-Weight matrix

$Term_a$	$Term_b$	$Term_c$	...	$Term_\omega$
w_1	w_2	w_3	...	w_ω

The terms are ordered with decreasing weight and the resulting terms vector is stored. The additional operation of storing the number of discovered terms is also performed, however, its value is not used.

The result structure is shared by all the algorithms, with the first row being the cluster's identifier, the second row the number of discovered topics and the third row the sorted terms. An example is shown in Figure 4.8.

	V1	V2	V3
1	1	3	6
2	1762	41	169
3	c("raspberry", "beauty", "calls", "dallas", "dbrewing", "...	c("current weather marlborough", "current weather n...	c("rabbitbrewery", "regrann", "bar", "bar kpa", "bar kp

Figure 4.8: Illustrative algorithm result

- **Pagerank**

Pagerank differentiates itself from the other algorithms by not using a requested term weight vector. It instead uses the terms tweet occurrence information to attain a term-term adjacency matrix. This matrix is used to build an undirected text graph, used by the Pagerank algorithm. The graph and algorithm use the implementation provided by the R package igraph [32].

The sorting of its terms is done similarly to TF-IDF's.

- **LDA**

The LDA was noted to be the most demanding in its application and algorithm requirements. The R implementation of LDA [33] dictates the use of a Document-Term matrix with the removal of documents without terms. It requires the number of topics, composing terms and fitting method to be supplied. To comply with the latter requisites, the number of topics to be discovered was set to 35 with a Gibbs sampling fitting method. This produces a topic model on which the request for the first three terms is made. The topic model contains Document-Topic distribution information, as can be seen in Figure 4.9, on which a Document probability is attributed to each Topic and its sum is one.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12
1	0.02421308	0.02421308	0.02421308	0.02421308	0.02421308	0.04116223	0.02421308	0.02421308	0.02421308	0.02421308	0.02421308	0.04116223
2	0.02132196	0.03624733	0.03624733	0.02132196	0.05117271	0.03624733	0.06609808	0.02132196	0.02132196	0.02132196	0.02132196	0.02132196
3	0.02040816	0.03469388	0.02040816	0.02040816	0.02040816	0.02040816	0.03469388	0.07755102	0.04897959	0.02040816	0.02040816	0.02040816
4	0.01930502	0.01930502	0.01930502	0.01930502	0.01930502	0.01930502	0.01930502	0.04633205	0.03281853	0.03281853	0.03281853	0.03281853
5	0.01984127	0.01984127	0.01984127	0.03373016	0.03373016	0.01984127	0.03373016	0.01984127	0.03373016	0.01984127	0.01984127	0.04761905
6	0.02012072	0.02012072	0.02012072	0.03420523	0.03420523	0.04828974	0.02012072	0.04828974	0.02012072	0.02012072	0.02012072	0.02012072
7	0.02012072	0.03420523	0.02012072	0.02012072	0.03420523	0.04828974	0.02012072	0.02012072	0.02012072	0.07645875	0.04828974	0.02012072
8	0.03281853	0.03281853	0.03281853	0.01930502	0.01930502	0.04633205	0.01930502	0.03281853	0.03281853	0.01930502	0.01930502	0.01930502
9	0.01855288	0.01855288	0.01855288	0.04452690	0.03153989	0.01855288	0.03153989	0.01855288	0.03153989	0.01855288	0.03153989	0.01855288

Figure 4.9: Document-Topic distribution example of 9 Documents(rows) on the first 12 Topics(columns)

The Document-Topic distribution matrix is reduced by summing the document values on each topic. The resulting vector is then sorted and its corresponding terms ordered accordingly. The

final result reduces the previous vector to its unique terms.

All the algorithms implemented possess a debug field, by default "off", to provide a step-by-step inspection of their operations.

The finalizing step prepares the algorithms sorted terms and persists them on the MySQL database. It iterates the results, joining the first three topics with a semicolon and performing an *insert* query. These topics are related to their corresponding micro-cluster.

To evaluate the script performance, its main procedures where timed, with the results shown in Table 4.4.

Table 4.4: R script elapsed time

	Elapsed time (s)
Library loading	2.65
Function loading	0.01
MySQL query	26.29
Corpus pre-processing	9.80
TF-IDF	7.43
Pagerank	2.07
LDA	11.88
Persist results	0.05

In Table 4.4, it's shown that the database connection and query procedure have a less desirable process time. This is due to one of the database's table, that stores tweet information, growing larger with each micro-cluster instance. The LDA elapsed time, considerably larger then the other algorithms, may be attributed to the static request of topics and terms to be discovered. Ideally, both of these LDA requests would vary with each corpus size.

The R version choice set the requirement to manually install packages, by downloading their binary files on the CRAN repository and using the command line prompt *R CMD INSTALL*. it allowed the use of up-to-date code, as opposed to the install packages command. An example of usage gain would be the tm Package update the corpus object structure, greatly improving its handling.

4.1.3 App server

The App server is the front-end website, from which the user interacts with the discovered patterns. It uses the PHP framework CodeIgniter with a Model–view–controller (MVC) architecture. It interacts with the back-end in two manners, a Web page request and a macro-clustering request with a chosen dimension distribution. These interactions can be seen in Figure 4.10.

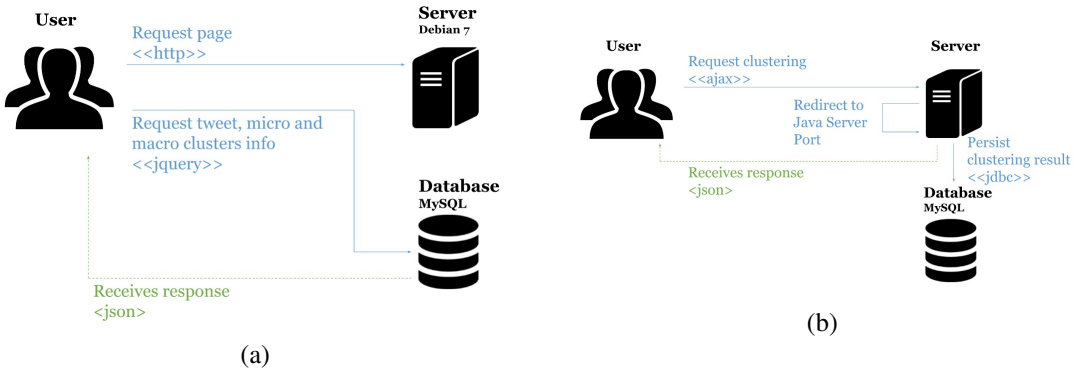


Figure 4.10: (a) Web page load event sequence; (b) Request clustering event sequence

The obtained micro and macro-cluster relation allowed the request for macro-cluster specific tweets and topics. This change allowed a more suitable cluster exploration experience and later changes to be user oriented.

The App server was adapted regarding the business understanding gathered from previous iterations and newly acquired information from inquiries. Business understanding refers to knowledge of the actors, their procedures and needs on the journalistic process. This step was divided into four interactions with the JPN(JornalismoPortoNet) team, namely, a first interview with its administration to showcase the proposed project, a first exploratory inquiry, a usability test and a final inquiry.

The starting proposal took into account the feedback provided by journalism students in TweeProfiles3 [9], coupled with the idea of what the Topic extraction task could provide. The state of the TweeProfiles tool and the mockups of the perceived usage were presented to the administration. These mockups were created without in depth knowledge of the business processes, thus, focused on technical aspects rather than the journalistic needs.

The mockups presented had three types of exploration, a cluster based, topic based and subject based exploration.

- **Cluster based exploration**

This exploration would focus on macro-clusters, the result of the DBSCAN algorithm. These are tweet groups formed from related micro-clusters on a requested set of dimension values, such as 0% spatial, 30% temporal and 70% content. Its steps are shown in Figure 4.11, with the UI presented in Portuguese.

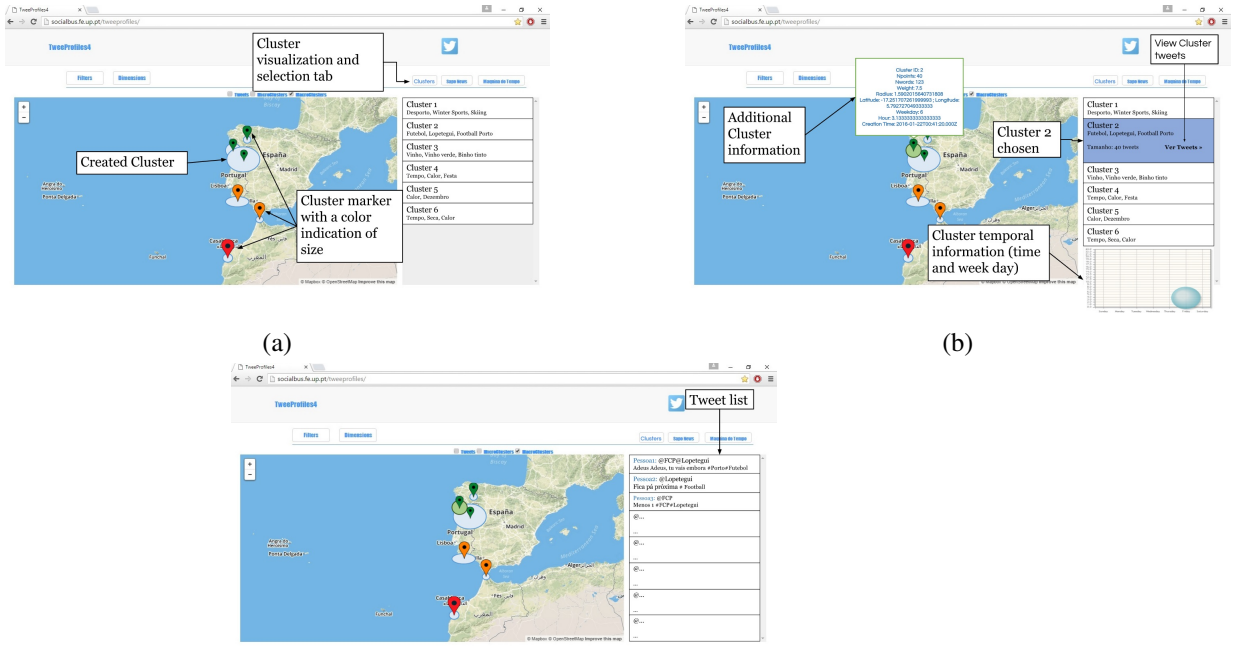


Figure 4.11: (a) General macro-cluster view in map and on a selection table; (b) Selected macro-cluster info; (c) View of related tweets on a selected macro-cluster

This proposal would allow the user to perceive their interest on a tweet cluster given its map position, size representation through color and associated topics. On selection, be it by either clicking on a marker or table row, an info window would display the cluster's information and also give a visual cue on the respective table row. The corresponding tweets could then be viewed by clicking the "See Tweets" link, which would alter the tab to display them. The table placement replaces the word cloud, since it was considered, in previous feedback, to be of low interest. From a technical standpoint, this proposal implied the addition of markers linked to a table row, where the cluster topics would be shown. The markers icon would vary with cluster size and its selection would have a visual impact on the table. Topic extraction would be used for text summarization.

- **Topic based exploration**

Topic based exploration would differ from the previous proposal by showcasing all attained topics. The topics would be displayed on the table, sorted by cluster occurrence, as can be seen in Figure 4.12.

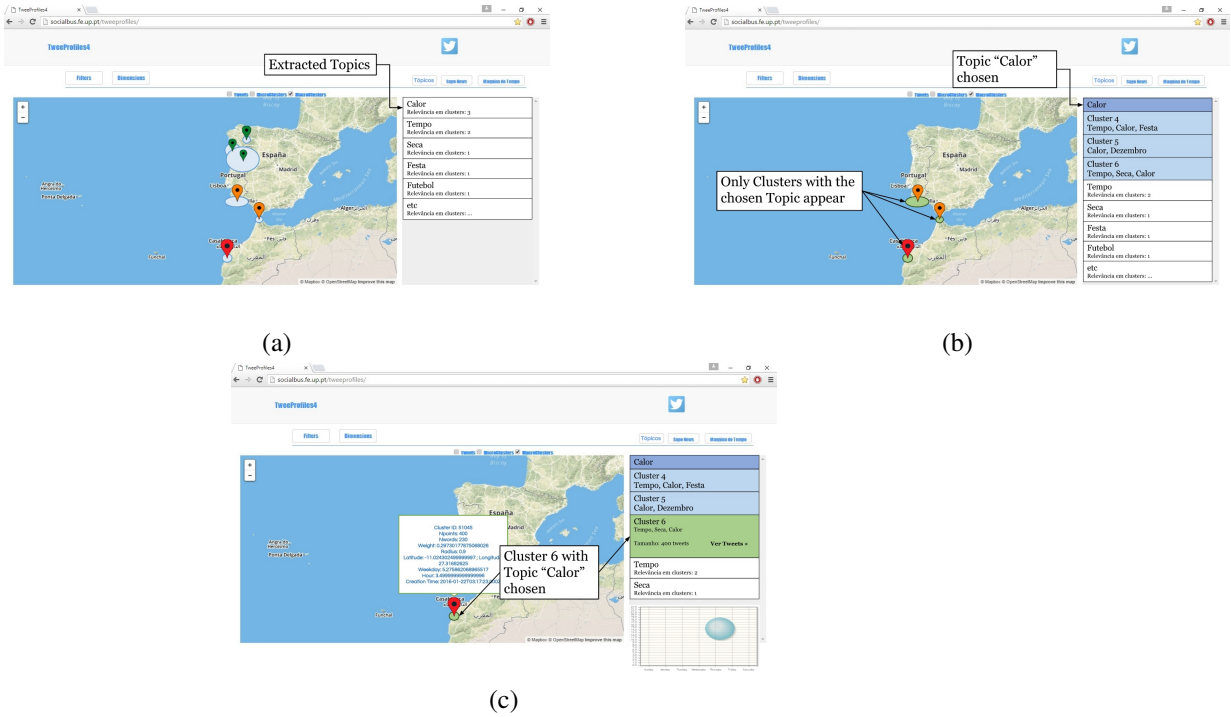


Figure 4.12: (a) Topic listing with cluster occurrence value; (b) Topic selection changes; (c) Cluster is selected

This proposal would attempt to guide the user to a cluster, given the user's interest on a topic. To provide a visual input in this multiple section process, selection of a topic would limit, based on occurrence, the number of clusters shown. Its technical difficulties would resemble that of the previous proposal.

- **Subject based exploration**

This last proposal would guide the user to a cluster by grouping them in set subjects. This was thought to narrow the search for professionals of that area of interest. The interaction would be similar to the topic based exploration proposal and can be seen in Figure 4.13.

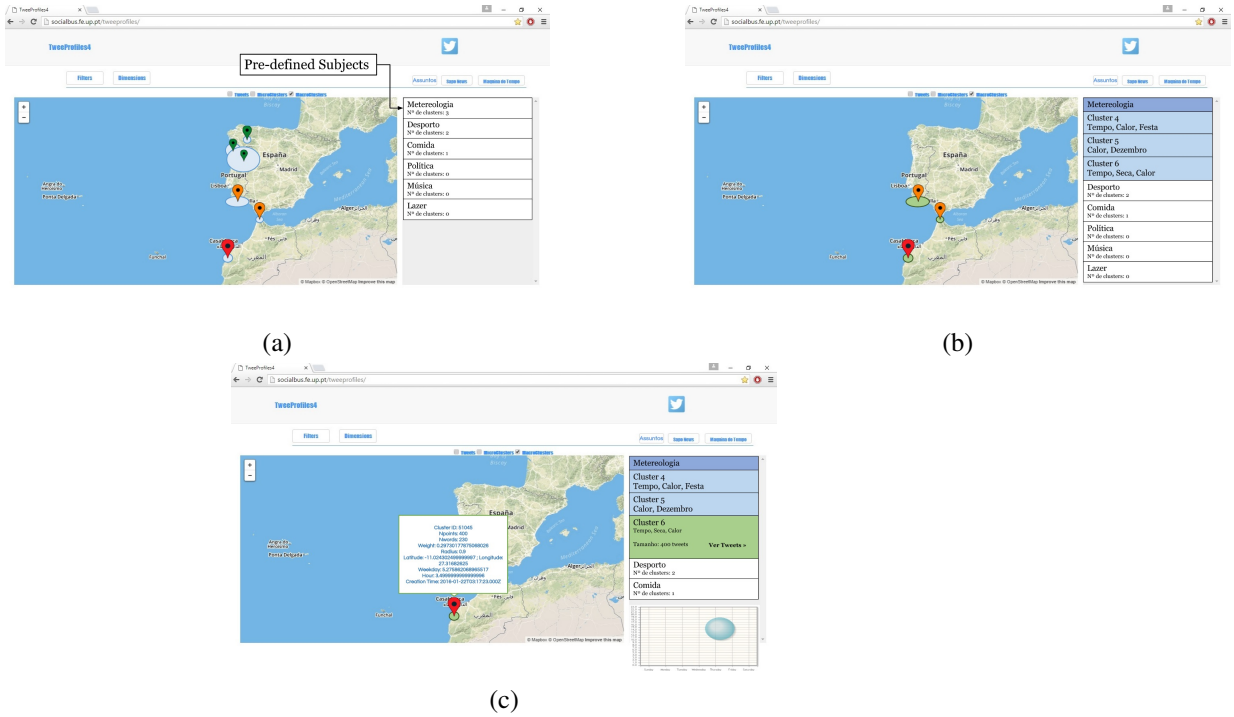


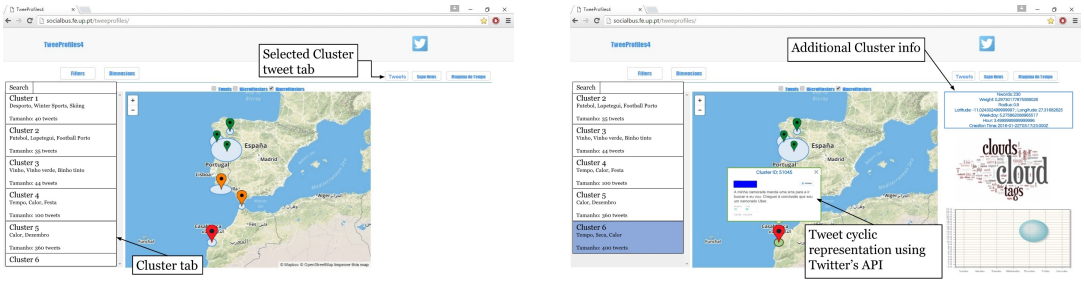
Figure 4.13: (a) Subject listing with cluster occurrence value; (b) Subject selection changes; (c) Cluster is selected

Its distinguishing aspect, regarding the previous implementation proposals, would be the use of Topic extraction for classification.

In addition to the different types of exploration, alternatives to how the information is displayed were also presented. These try to make use of the website's layout as well as utilize Javascript and Twitter's API capabilities to assist the conducted exploration.

- **Sectioned usage**

This proposal considered the size of each element on the website page. The map and word cloud size was deemed unappropriated, as they occupied a considerable large portion for their purpose. To better utilize said space, the aforementioned cluster table was added to the left of the map area, thus freeing the right side to hold additional information. The word cloud was also reduced in size, allowing for the cluster information to be presented above it. These changes split the exploration in two parts, the table and map would provide quick visual information to aid cluster selection and the panel on the right side would be useful for more detailed analysis. The marker's info window was then proposed to show a cyclic tweet representation of the selected cluster. This result is shown in Figure 4.14.

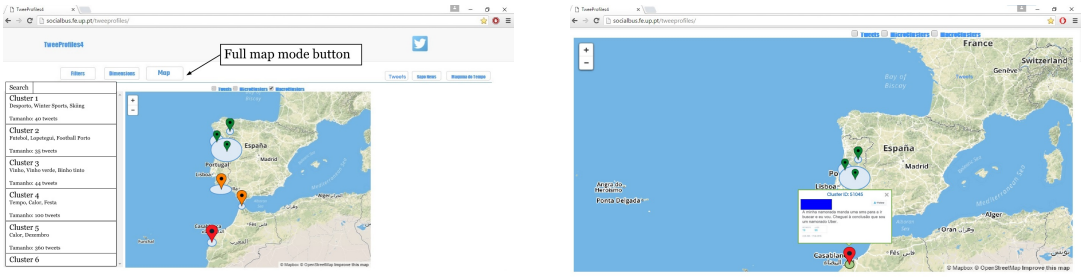


(a) (b)

Figure 4.14: (a) Selection step with focus on cluster table and map; (b) Additional information presented

- **Full map view**

The final proposal would use the previous change to the info window content in conjunction with the idea of having multiple clusters on the map. This would allow the inspection of multiple clusters at once, however, it would also require a larger surface, due to overlapping text. To this end, a button is presented, above the map, to toggle the display of both sides of the map, as is shown in Figure 4.15.



(a) (b)

Figure 4.15: (a) Placement of the button on the upper area, with the rest of exploration altering buttons; (b) Full map view

The previous exploration proposals and mockups where shown to the JPN administration, which allowed the interaction with their students through inquiry sessions and usability tests.

The first inquiry session took place in JPN on February 24 2016 with a total duration of 30 minutes. It was divided into a presentation of the TweeProfiles tool, in its current state and proposed ideas, followed by the inquiry. The inquiry was divided into three main sections, focusing on personal and professional work experience, the desired TweeProfiles exploration method, with most fitting office, and suggestions.

There were 16 respondents to this inquiry, with a similar distribution of students and interns. Their age varied between 20 to 32 years with a predominant few months experience in journalism related work. Their use of social networks in a professional setting had Facebook as a predominant answer, with Linkedin and Twitter as runner-ups.

The respondents preferred information focused on, by order of interest, organizational entities, events, discussed topics and people.

The answers to the most fitting journalistic role, for the proposed types of exploration, showed a predominant assignment of *Editor* to both cluster and subject based explorations, while *Pauteiro* was assigned to topic based. The preferred method of exploration was the cluster based proposal, with subject based a close second. The journalistic role most likely to benefit from the TweePro-files tool was assigned similarly to both *Editor* and *Pauteiro*. The additional use cases, sectioned usage and full map view, were deemed appropriate to assist the aforementioned methods.

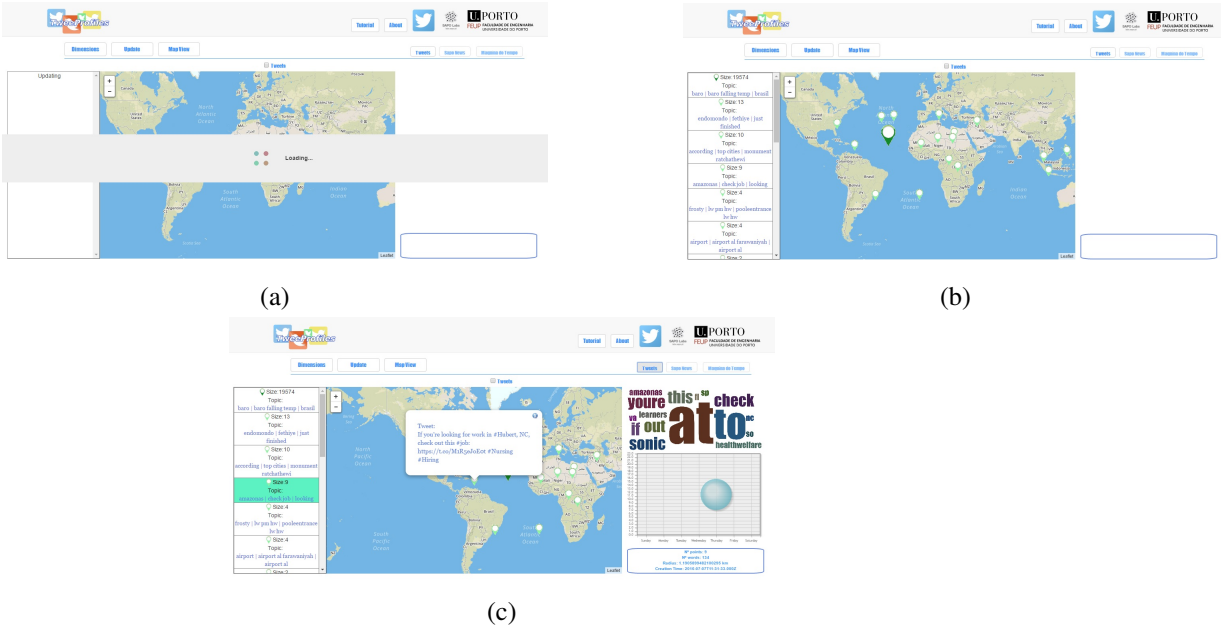
The suggestions emphasized the interest on more statistical information, on the users age and genre, hyperlink co-occurrence and Twitter accessibility through its tweets. It was also mentioned the need of a tutorial video, to hasten the users grasp of the tools capabilities.

The inquiry results led to the conclusion that an adaptation of the TweeProfiles current cluster based exploration, with a focus on the needs of the *Editor*, would achieve a greater acceptance among journalists. The analysis of the inquiry was further detailed in a document, written in Portuguese, that was sent to the JPN's administration, which can be seen in Appendix D.

4.2 Results

This section presents the results of the changes made and their evaluation. The results focus on the implemented changes to the App server and cluster assigned topics. The evaluation was performed regarding the usability test and final inquiry done by the JPN.

The App server was adapted accordingly to the inquiry results. The cluster selection table was added, in which the first three topic results of the associated micro-clusters were placed, along side the cluster's tweet size and marker representation. This limit to the provided topics was due to space restrictions. The cluster's tweet size alters it's marker in both size and color. A cluster's selection alters its table cell by highlighting it. An additional banner is shown until the database requests finish. The removal of the cluster radius representation was performed due to its incorrect representation when changing the map's zoom level. This interaction was thought to have a bigger use due to the possibility of multiple infowindows being shown. Even when fixed, it was left out to reduce the amount of information on screen, which led to some confusion by the journalists as to why where the markers in a given place. The loading banner, cluster visualization and selection interactions are shown in Figure 4.16.



(a) (b)

(c)

Figure 4.16: (a) Loading banner; (b) Cluster table and map view; (c) Select cluster info

A tutorial video is accessible by pressing the Tutorial button on the web page's header. It showcases and describes the available features. The software chosen to produce it were Camtasia Studio 8 and Open Broadcaster Software(OBS) Studio. It can be viewed through an embedded player, as is shown in Figure 4.17.

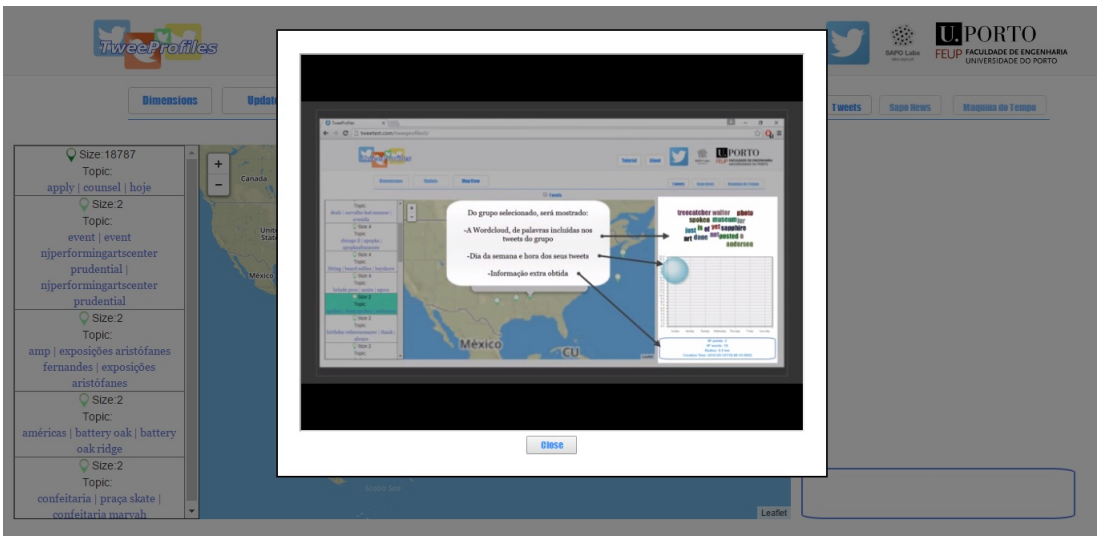


Figure 4.17: Tutorial video viewing

A cluster's selection triggers its marker infowindow, centering the map on it if done through the table. The infowindow was redesigned to display its tweet text, alternating every three seconds. The possibility to have multiple infowindows active was added, to allow simultaneous analysis of multiple clusters. The Tweets menu was altered to request an embedded tweet representation from

Twitter, providing quick access to its content. This action respects Twitter's privacy policy, since only available tweets can be shown. Both tweet representations allow the user to select a tweet, opening a new tab to their Twitter origin. These are portrayed in Figure 4.18.

The following figures have the user's information covered, to comply with Twitter's privacy policy.

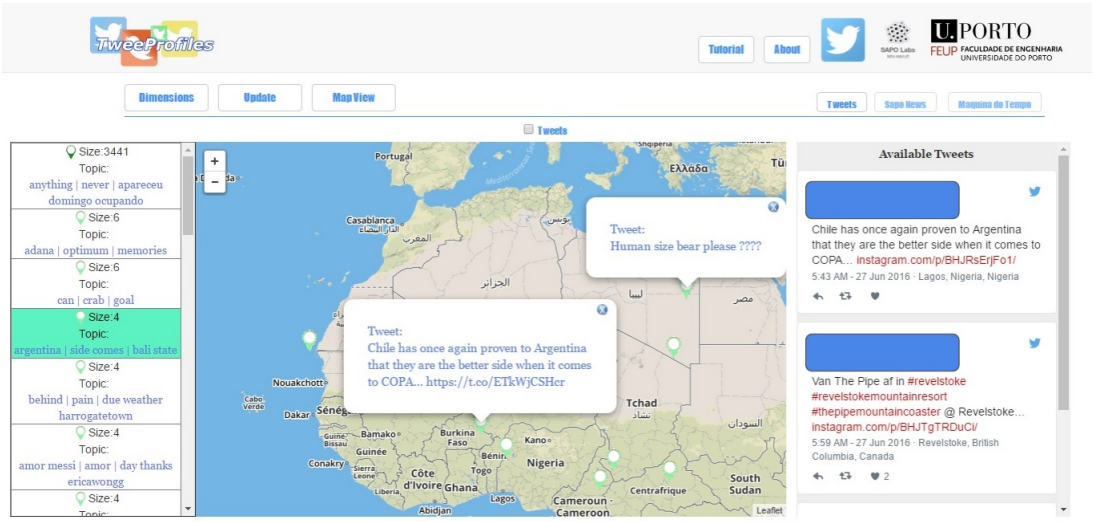
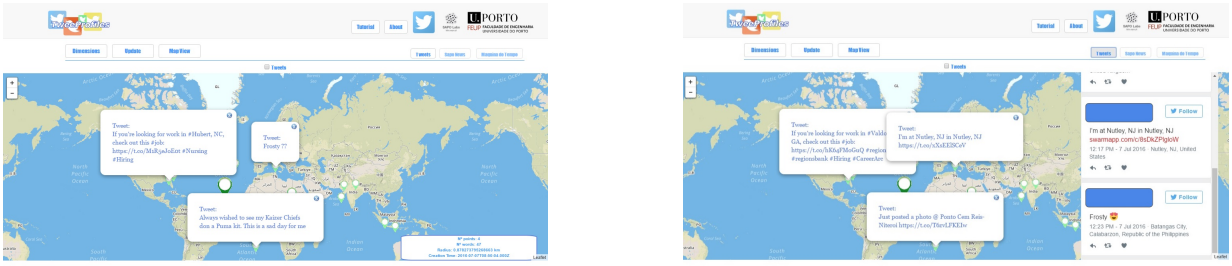


Figure 4.18: Cluster and tweet visualization options

The Full map view option, hiding the lateral bars, was included. It can be switched to, and from, by clicking the Map View button on the web page's header. To not hinder the exploration, additional information is kept and its tweets can still be accessed. The results can be seen in Figure 4.19.



(a) (b)

Figure 4.19: (a) Map view's multiple selected clusters and info; (b) Map view's tweet visibility

The cluster's topic display had varying results. The smaller cluster's content could either be similar or distinct and still possess adequate topics, as seen in Figure 4.20. The larger cluster's, however, would need to have similar content to not suffer information loss, as seen in Figure 4.21 and 4.22.

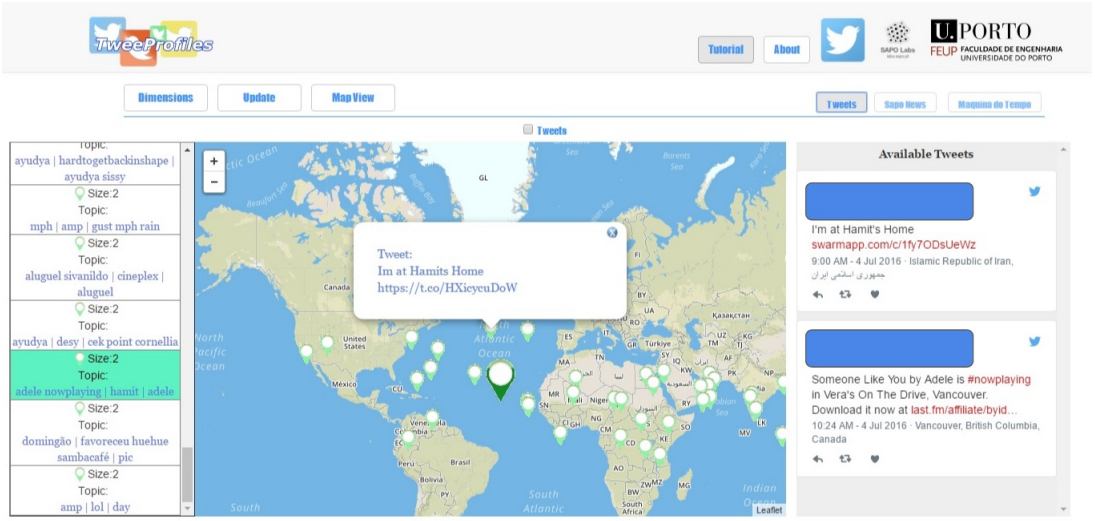


Figure 4.20: Small cluster with distinct content, topics: adele now playing; hamit; adele

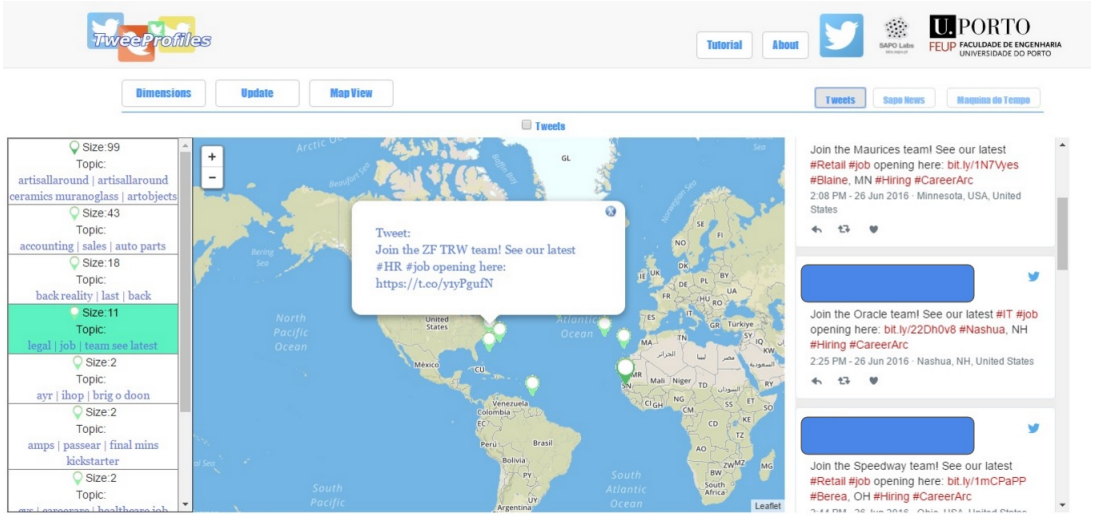


Figure 4.21: Large cluster with similar content, topics: legal; job; team see latest

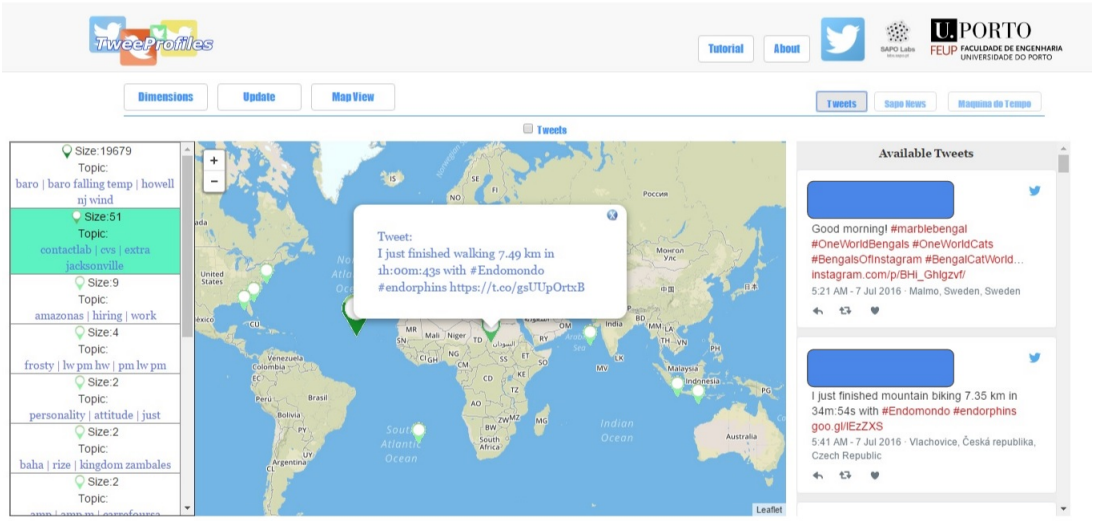


Figure 4.22: Information loss on a large cluster with distinct content, topics: contact lab; cvs; extra jacksonville

Small clusters can, however, have less desirable topics due to both their limit and how the applied methods explore. In Figure 4.23, a single tweet is represented by the cluster's topics due to the word frequencies. It also displays the aforementioned emoticon change to question marks, which in this case is undesirable.

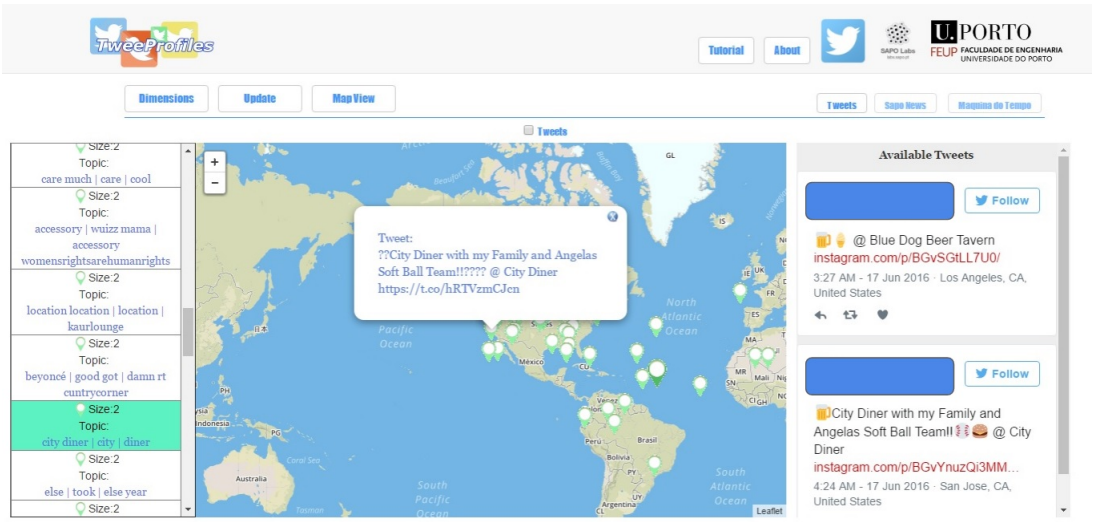


Figure 4.23: Cluster's topics target lower tweet, topics: city diner; city; diner

The removal of stopwords eases the search process of content descriptive topics. This can, however, lead to topics not correctly describing the message. In Figure 4.24, the lower tweet's message of wanting a "beer on a hot summer" was stripped of stopwords, thus "beer hot" was discovered. This example supports the findings in [22], as stopwords should not always be removed.

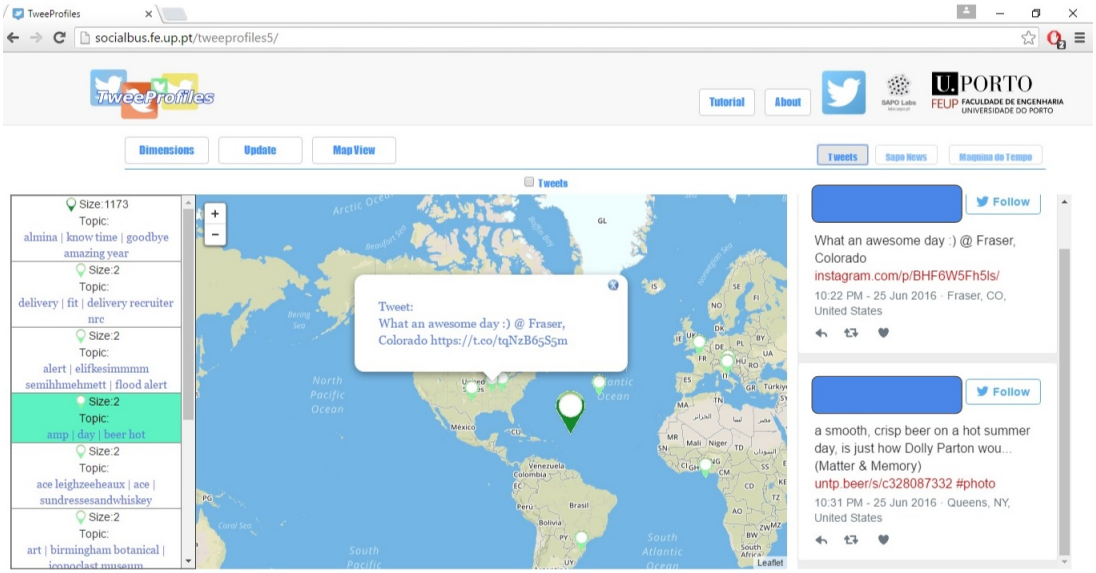


Figure 4.24: Stopword removal flaw, topics: amp; day; beer hot

4.2.1 Usability Test

A usability test was conducted in JPN on May 17 2016. The participants were tasked with using the TweeProfiles tool to discover news worthy material, with data gathered from the previous day. The results of their search would be stored on a provided Google form.

Although the majority did not find news worthy material, one student was able to. This was the tweet discussing *Sinead O'Connor*, in Figure 4.25, to be so.

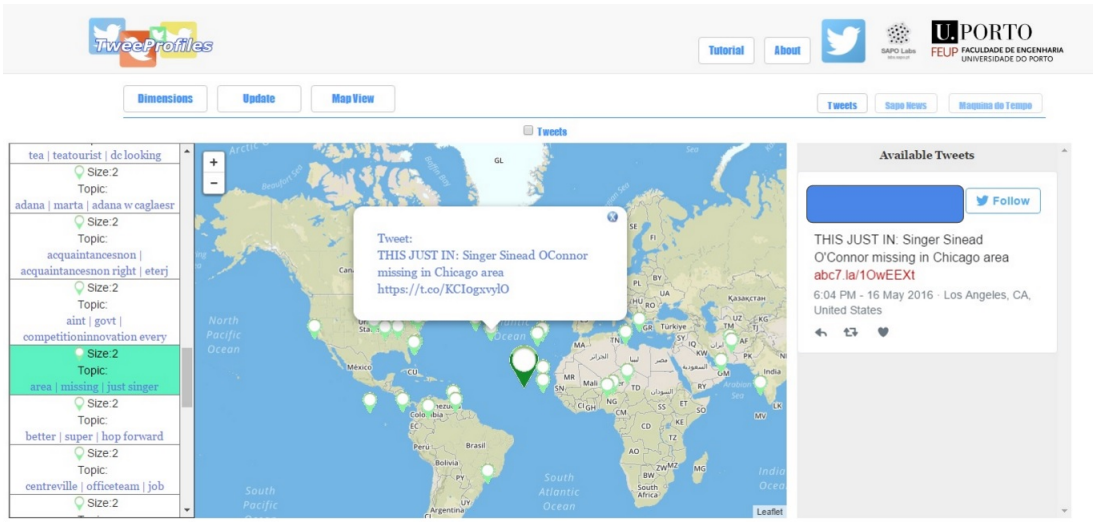


Figure 4.25: Tweet discussing Sinead O'Connor, topics: area; missing; just singer

This tweet was gathered by the TweeProfiles tool and stored at 18:04 on the previous day, as is seen in the database result Figure 4.26. The date field shows the API result, presented in GMT+0.

tweetid	text	date
732255345448357890	this just in singer sinead oconnor missing in chicago area	2016-05-16T17:04:13.000Z

Figure 4.26: Database tweet insertion information

TVI24, a Portuguese news station, produced a news article 1 hour and 19 minutes before the tweet's occurrence by citing a different tweet on the same subject, as seen in their post [1]. Their interest in the occurrence thus validates the choice.

4.2.2 Final inquiry

The final inquiry requested a scaled, one to five, evaluation of the tools features. The respondent was a JPN Editor, rating the tools overall capabilities, to assist Editors in finding news worthy information, a three.

The most valued features, with a four rating, were the cluster's week and hour information and the created tutorial.

The least valued features, with a two rating, were the cluster table, the cluster's wordcloud, the additional cluster information and the marker's infowindow displaying the cluster's tweet text. The cluster table had an additional commentary, stating that the information was not intuitive.

[1] http://www.tvi24.iol.pt/musica/chicago/sinead-o-connor-esta-desaparecida

Chapter 5

Conclusions and Future work

This chapter contains the summary of the performed tasks, a discussion of the results and future work suggestions.

5.1 Summary

The dissertation's main goals were to assign labels to tweet clusters and to adapt the tool to the journalists method of news discovery on Twitter data.

The cluster's label assignment was performed by a Topic extraction approach. The task involved the preliminary study of algorithms to explore different representations of the data, of which methods using the TF-IDF weight, Pagerank and LDA were chosen. These were assumed to rank the tweet's candidate topics in a different manner. They were implemented using the R language. Evaluation was performed with an experimental analysis on a small, artificially generated dataset and on real data obtained using the SocialBus platform. A top@N agreement analysis on the latter dataset results was used to verify the initial assumption. The results corroborated the assumption on the requested number of topics N. The task was added to the micro-clustering process on the Analytics server.

The Analytics server was adapted to provide micro to macro-cluster relations as well as additional info required by the Topic extraction step. The required number of gathered tweets for a micro-cluster update was also changed, to provide more up-to-date information.

The App server's cluster display was altered to make use of the micro to macro-cluster relation, relating tweets and topics. The adaptation to the needs of the journalists was performed in two phases, with four iterations. These iterations were conducted with the assistance of JPN. These interactions consisted of a project presentation, a first inquiry and a usability test. The project presentation was performed to both the administration and students. The first inquiry presented the mockups for three methods of exploration and additional use cases. It requested the most appropriate method, corresponding journalistic role and additional suggestions. The inquiry results were used to shape the App server's user interaction. The user interaction was further iterated

with additional use cases, such as cyclic tweet representation and embedded tweet display. The implemented feature were evaluated with a usability test and a final inquiry.

The acceptance of the tool was moderate due to the way labels are display and it missing the journalists most valued aspect, quick access to newly gathered data.

5.2 Discussion

In this section, some debatable aspects are presented and discussed.

• "Can the expansion of Twitter clusters' labels, using Topic extraction approaches, improve journalists experience on creating news articles from TweeProfiles?". To answer this question, topics were assigned to the attained clusters and this expansion was evaluated in the usability test. This addition assisted a journalist to identify an interesting subject.

• "Could the Tweeprofiles' tool be adapted to assist the exploration methods of media journalists?". This question was answered in the final inquiry, by asserting their interest in breaking news. The TweeProfiles is a pattern discovery tool, therefore an adaption of this method was implemented.

• The use of only the first three topics to be displayed was initially thought to be sufficient information for cluster selection. It was ultimately considered unfitting, due to information loss, requiring a new form of display such as a wordcloud.

• The reduction of the required number of tweets in a micro-clustering instance, although maintaining similar size distributions, would need its patterns to be re-evaluated. This task would determine if the discovered cluster's content still possessed the desired similarities.

5.3 Future work

This section presents some future work suggestions:

• **Application-related issues**

• The journalists require potential news worthy information the moment it is gathered. A possible change to TweeProfiles would be the addition of a Tweet stream display. This stream would update with the arrival of newly gathered tweets.

• The possibility of gathering other social media information would possibly appeal to journalists, as the inquiry results would indicate.

• **Technical issues**

• A restructure of the MySQL database would improve query time, especially targeting the tweet table. This would greatly benefit all queries that require tweet information.

• The server cannot handle multiple macro-clustering requests at the same time. This forced the Analytics server to be disabled during the usability tests. Its change would allow further test to display current information.

• An improvement to the R script's process time would be to nullify the libraries and function loading time by accessing an already loaded version.

• **UI issues**

• Cluster table information should be altered to supply more intuitive information. A possibility would be to change its content to the journalist's requirement of statistical information.

Appendix A

DBSCAN

A.1 DBSCAN algorithm

Algorithm 1 DBSCAN

```
1:  procedure DBSCAN(MinPts : neighborhood_threshold, D : dataset, ε : radius_parameter)
2:      Mark all objects as unvisited;
3:      while no object is unvisited do
4:          Randomly select an unvisited object p;
5:          Mark p as visited;
6:          if the ε-neighborhood of p has at least MinPts objects then
7:              Create a new cluster C and add p to C;
8:              Let N be the set of objects in the ε-neighborhood of p;
9:              for each point p' in N do
10:                 if p' is unvisited then
11:                     Mark p' as visited;
12:                     if the ε-neighborhood of p' has at least MinPts points; then
13:                         Add those points to N;
14:                     end if
15:                 end if
16:                 if p' is not yet a member of any cluster then
17:                     add p' to C;
18:                 end if
19:                 Output C;
20:             end for
21:             else mark p as noise;
22:         end if
23:     end while
24: end procedure
```

Appendix B

HybridDenStream extension

B.1 HybridDenStream extension algorithm

Algorithm 2 HybridDenStream extension

procedure HYBRIDDENSTREAM EXTENSION($D, \varepsilon, \beta, \mu, \lambda$)
  $T_p = \frac{1}{\lambda}\log(\frac{\mu\beta}{\mu\beta-1})$
  Get the next point X at current time t from data stream D;
  Try to merge X into its nearest p-micro-cluster c_p;
  if $r_p \leqslant \varepsilon$ **then**
    Merge X into c_p;
    Send assignment $A(X, c_p)$ to OverlapManager agent;
  else
    Try to merge X into its nearest o-micro-cluster c_o;
    if $r_o \leqslant \varepsilon$ **then**
      Merge X into c_o;
      Send assignment $A(X, c_o)$ to OverlapManager agent;
      if $w_o > \beta\mu$ **then**
        Remove c_o from outlier-buffer and create a new p-micro-cluster c_{pn} by c_o;
      end if
    else
      Create a new o-micro-cluster c_{on} by X and insert into the outlier-buffer;
      Send assignment $A(X, c_{on})$ to OverlapManager agent;
    end if
  end if
  if $(tmodT_p) = 0$ **then**
    for each p-micro-cluster c_p **do**
      if $w_p < \beta\mu$ **then**
        Delete c_p;
      end if
    end for
    for each o-micro-cluster c_o **do**
      $\xi = \frac{2^{-\lambda(t-t_o+T_p)}-1}{2^{-\lambda T_p}-1}$
      if $w_o < \xi$ **then**
        Delete c_o;
      end if
    end for
  end if
  if clustering request arrives **then**
    Generate clusters;
  end if
end procedure

Appendix C

First JPN Inquiry

Journalism 3.0: Multidimensional Cluster Visualization and Labelling on Twitter Data for Data Journalism

Este inquérito enquadra-se no projeto de dissertação intitulado “Journalism 3.0: Multidimensional Cluster Visualization and Labelling on Twitter Data for Data Journalism”, no âmbito do Mestrado Integrado em Engenharia Electrotécnica e de Computadores, da Faculdade de Engenharia da Universidade do Porto com a colaboração do JPN - JornalismoPortoNet.
O projeto pretende ponderar a exploração realizada no TweeProfiles, uma ferramenta de recolha e visualização de tweets, com o intuito de a adequar ao processo jornalístico acarretado por profissionais. O projeto será desenvolvido por Bruno Vieira e supervisionado pelos Engenheiros Carlos Soares e Jorge Teixeira.
O objetivo do inquérito é aferir a sensibilidade das pessoas para ferramentas de apoio ao jornalismo, com foco na investigação dos dados disponíveis e partilhados nas redes sociais, em particular do Twitter. O conhecimento adquirido será usado para moldar a ferramenta, sendo disponibilizado e discutido com os colaboradores do JPN.

*Obrigatório

Dados pessoais

1. **1- Qual é a sua idade?**
   (em anos)

   .....

2. **2- Quais as suas habilitações académicas?**
   Marcar apenas uma oval.

   - ◯ Licenciado
   - ◯ Mestre
   - ◯ Doutorado
   - ◯ Outra: .....

3. **3- Qual é a sua função profissional atualmente?**

   .....

4. **4- Alguma vez trabalhou numa área relacionada com o jornalismo/comunicação social? ***
   Marcar apenas uma oval.

   - ◯ Sim
   - ◯ Não *Passe para a pergunta 7.*

Passe para "TweeProfiles."

5. **4.1- Que funções executou?**

.....

.....

.....

.....

.....

6. **4.2- À quanto tempo trabalha/trabalhou na área?**
   (aproximadamente em anos)

.....

Dados pessoais

7. **5- Usa alguma rede social profissionalmente?**
   (poderá escolher várias opções)
   Marcar tudo o que for aplicável.

   - [ ] Facebook
   - [ ] Twitter
   - [ ] Linkedin
   - [ ] Outra: .....

8. **6- Que tipos de informação geográfica associada a redes sociais considera relevante para uma ferramenta de jornalismo/comunicação social?**
   (poderá escolher várias opções)
   Marcar tudo o que for aplicável.

   - [ ] Localização do utilizador que está a gerar a informação
   - [ ] Localizações que são mencionadas ou relevantes para a notícia (locais mencionados no Tweet sobre um evento, por exemplo)
   - [ ] Outra: .....

9. **7- Que tipo de informação espera ver numa aplicação web interativa desenhada para incorporar redes sociais como fonte de informação para apoio ao jornalismo/comunicação social?**
(poderá escolher várias opções)
Marcar tudo o que for aplicável.

- [ ] Pessoas (idade, sexo)
- [ ] Organizações (localização, contacto)
- [ ] Análise de Sentimento (se, dado o seu contexto, a publicação tem teor positivo ou negativo)
- [ ] Eventos (informações, localização)
- [ ] Sequência de Eventos (eventos em ordem cronológica)
- [ ] Tópicos discutidos
- [ ] Outra: .....

TweeProfiles

O TweeProfiles é uma ferramenta de exploração e visualização de informação obtida na rede social Twitter. A informação, contida nos tweets, é analisada e consequentemente usada para a identificação de grupos. Um grupo de tweets, criado pelas suas semelhanças, é referido como cluster.

Atendendo ao seguinte caso ilustrativo da criação de clusters:

Tendo os seguintes tweets:

-"Os portugueses escolheram a #TVI para saberem quem ganhou as #Presidenciais. Foi também a vossa escolha?"
-"Não me digam que este tempo é o São Pedro a chorar a saída do nosso PR do Palácio de Belém... #Presidenciais"
-"Perfil de Marcelo Rebelo de Sousa acabou d ser actualizado! #presidenciais #MarceloRebelodeSousa #presidenciais2016"
-"Precisas de boleia para assistir ao #FCPorto-@BVB? Sabe mais em #FCPBVB "
-"Fevereiro / February / Febrero @ Dragão. FC Porto - @BVB, 25/02, 20h05 (GMT) "

Assumindo a criação de clusters com base no seu conteúdo, seriam obtidos dois clusters, agrupando os tweets sobre as presidenciais e outro sobre o FCPorto.

Um dos objetivos do projeto é a atribuição de Tópicos a clusters, sendo um Tópico definido como um conjunto de palavras representativas dos tweets associados.
Atendendo à necessidade de confrontar diversos modos de exploração, o conceito de assunto é apresentado como a generalização de diversos Tópicos, ou seja, o Assunto Meteorologia abrangeria Tópicos como calor e frio.

De seguida são apresentados esboços para três formas de exploração usando o TweeProfiles.

Caso 1: Exploração baseada em Clusters

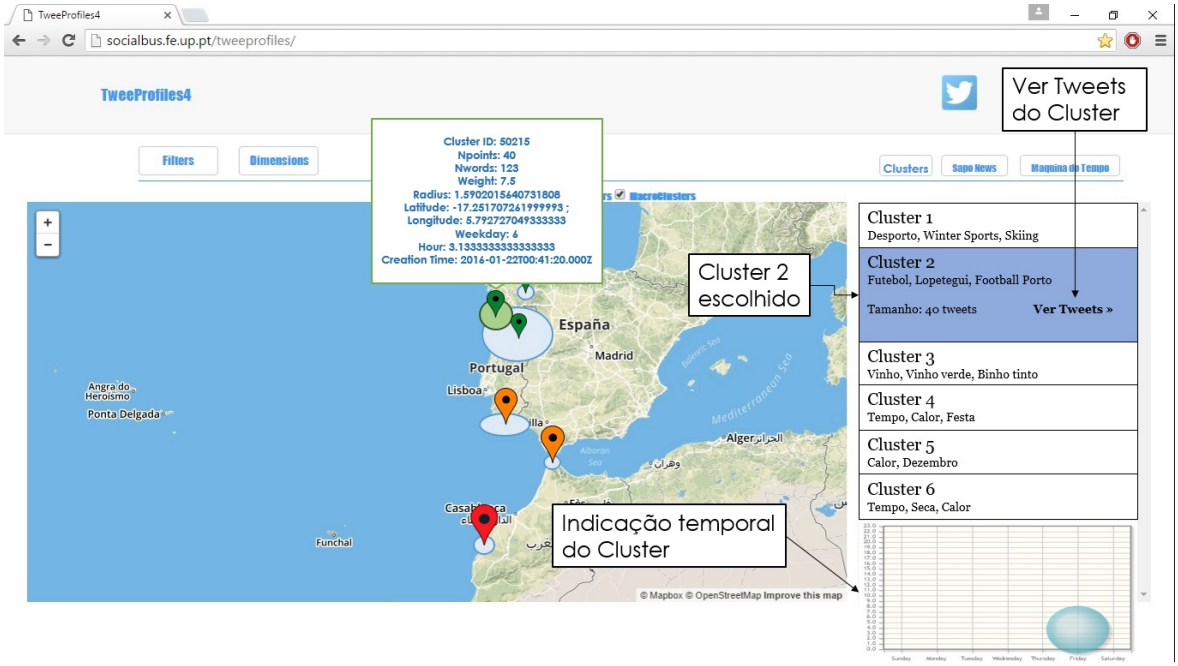
Caso 2: Exploração baseada em Tópicos

Caso 3: Exploração baseada em Assuntos

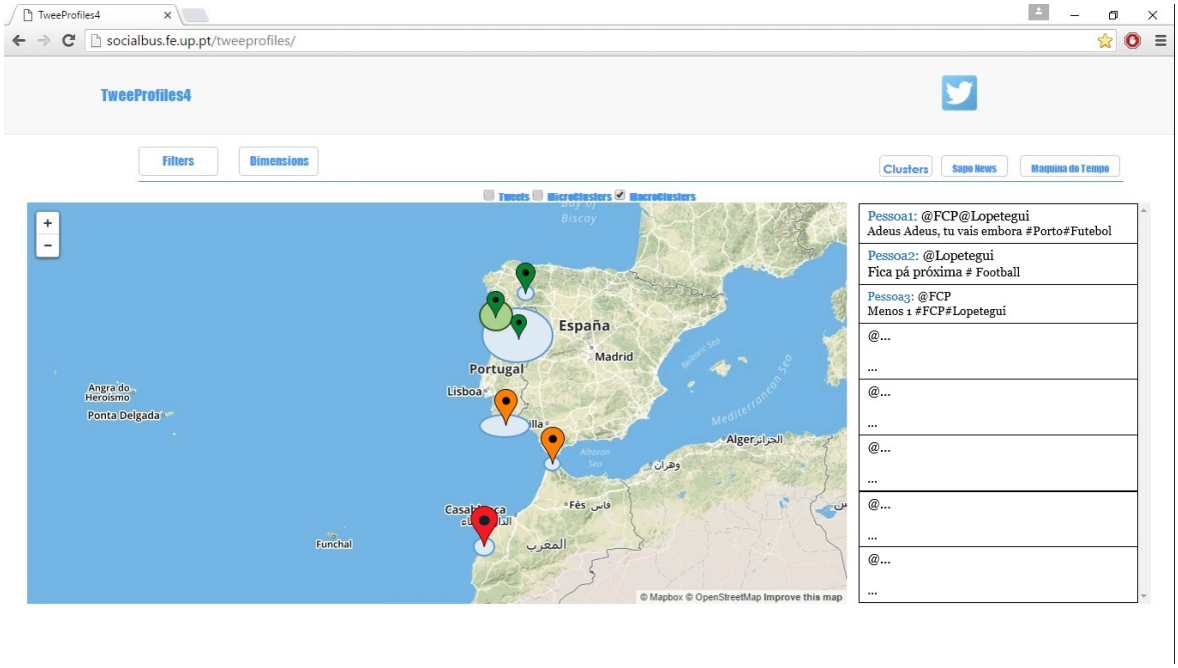
Caso 1: Exploração baseada em Clusters

Tab “Clusters” com a escolha de clusters

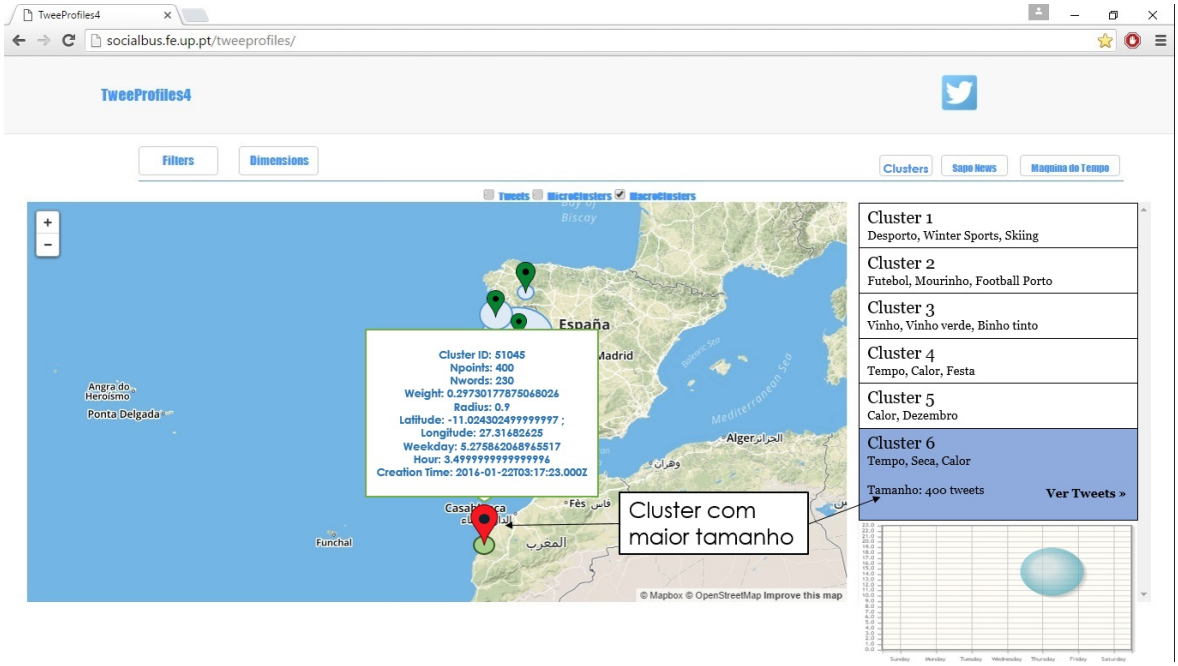
Cluster 2 escolhido



“Ver tweets” no submenu do Cluster3 escolhido



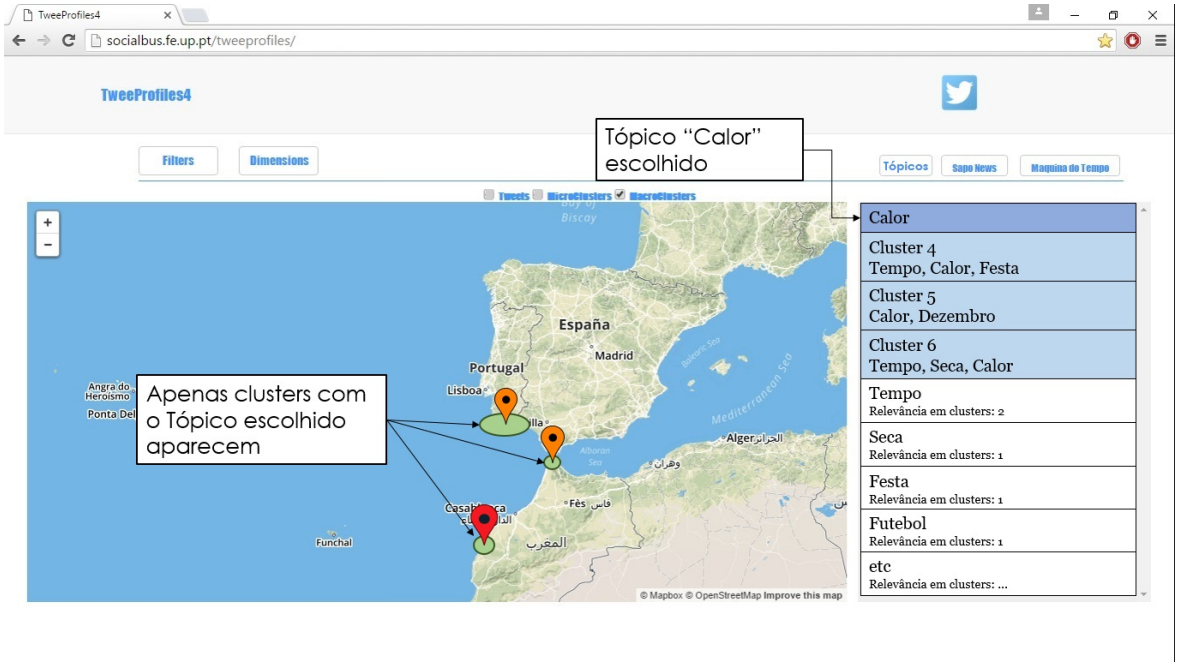
Cluster6 escolhido, visto ter maior tamanho



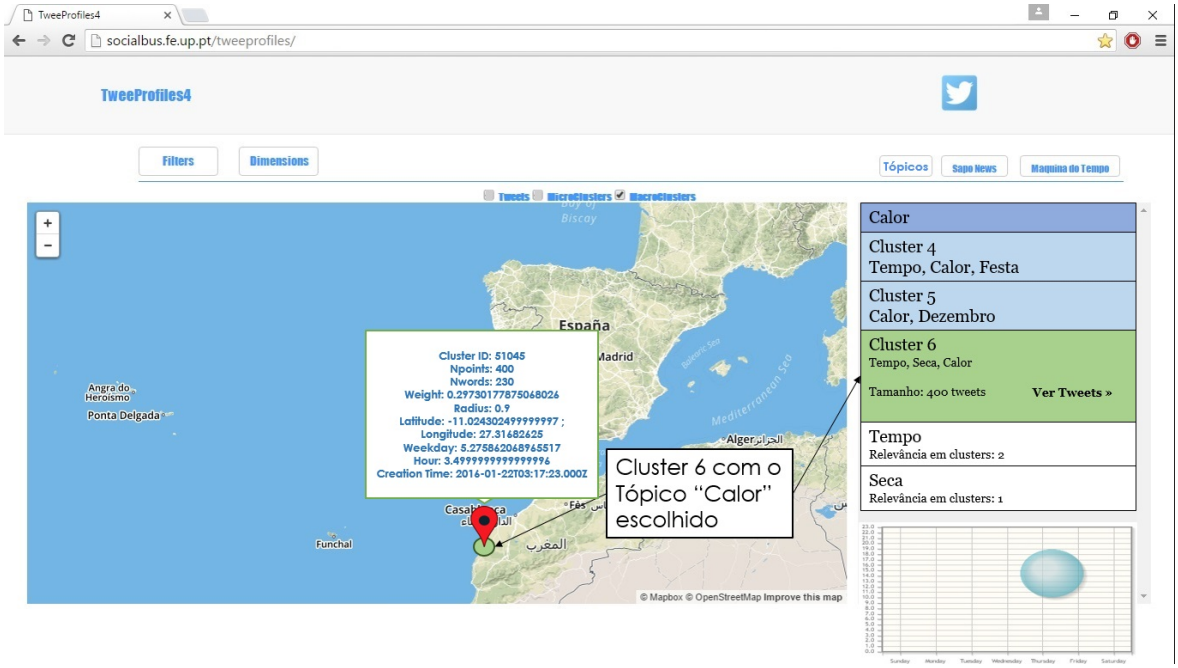
Caso 2: Exploração baseada em Tópicos

Tab “Topics” com a escolha de tópicos

Label “Calor” escolhida

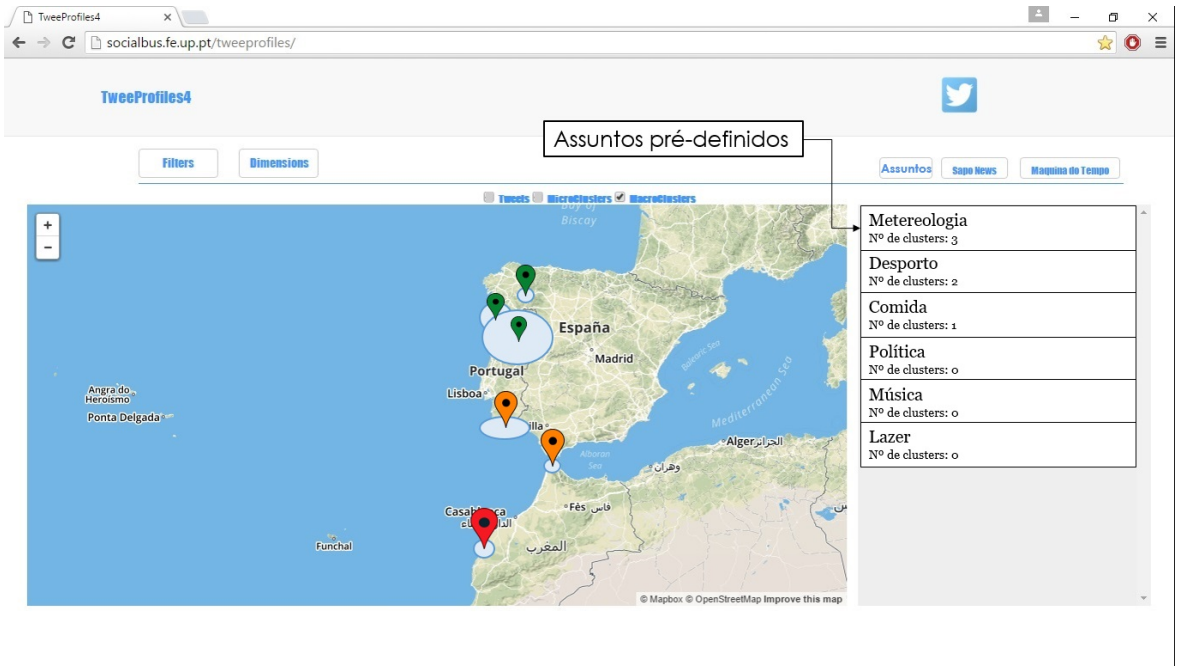


Cluster 6 escolhido

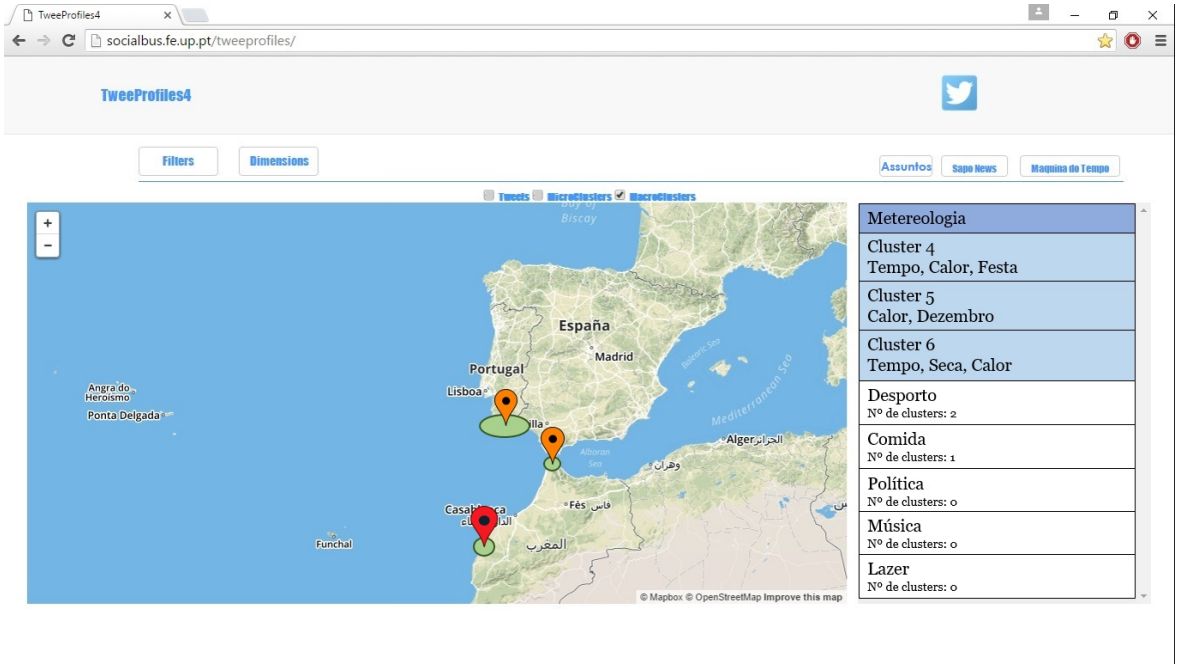


Caso 3: Exploração baseada em Assuntos

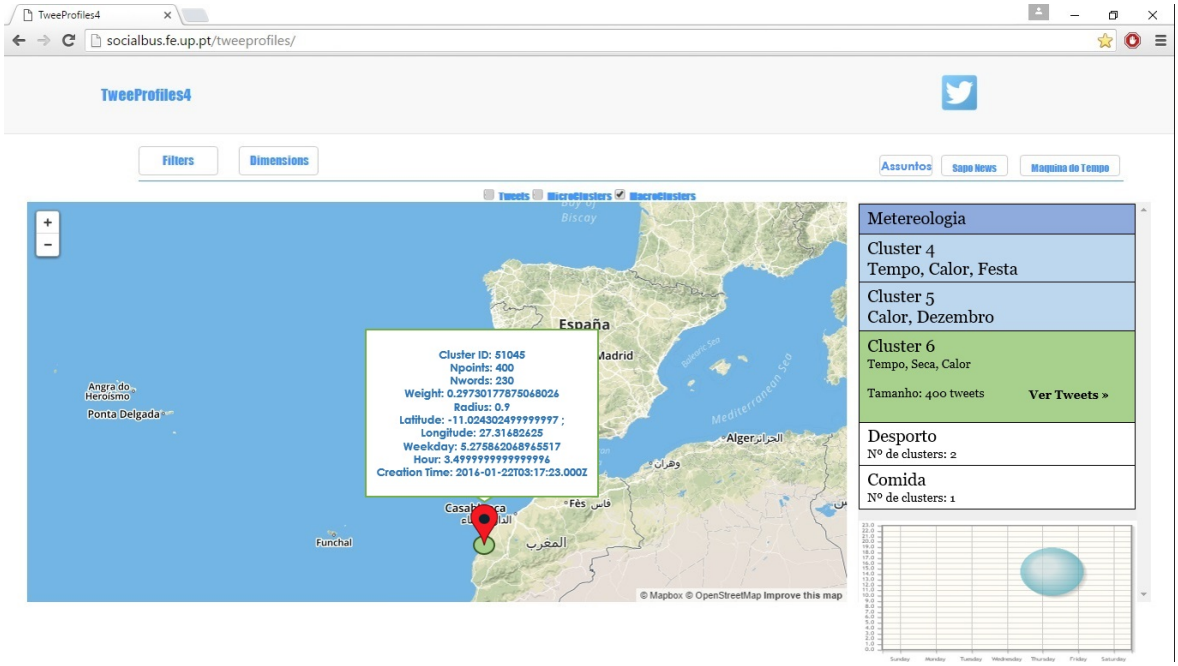
Tab "Assuntos" com a escolha de assuntos pré-definidos



Metereologia escolhida



Cluster 6 escolhido



10. **1- Qual o cargo que considera mais beneficiar desta ferramenta?**
Marcar apenas uma oval.

- ◯ Editor
- ◯ Colunista
- ◯ Pauteiro (chefe de reportagem)
- ◯ Outra: .....

11. **2- Qual o caso que melhor refletiria a exploração que o cargo faz/pretende fazer usando redes sociais ?** *

Marcar apenas uma oval.

- Caso 1: Exploração baseada em Clusters
- Caso 2: Exploração baseada em Tópicos
- Caso 3: Exploração baseada em Temas

12. **3- Quão adequado crê serem os casos apresentados numa exploração de redes sociais?** *

Marcar apenas uma oval por linha.

	não adequado	pouco adequado	adequado	muito adequado
Caso 1: Exploração baseada em Clusters				
Caso 2: Exploração baseada em Tópicos				
Caso 3: Exploração baseada em Temas				

13. **4- Atendendo às diferentes formas de exploração apresentadas, a que cargo atribuira cada caso?**

Marcar apenas uma oval por linha.

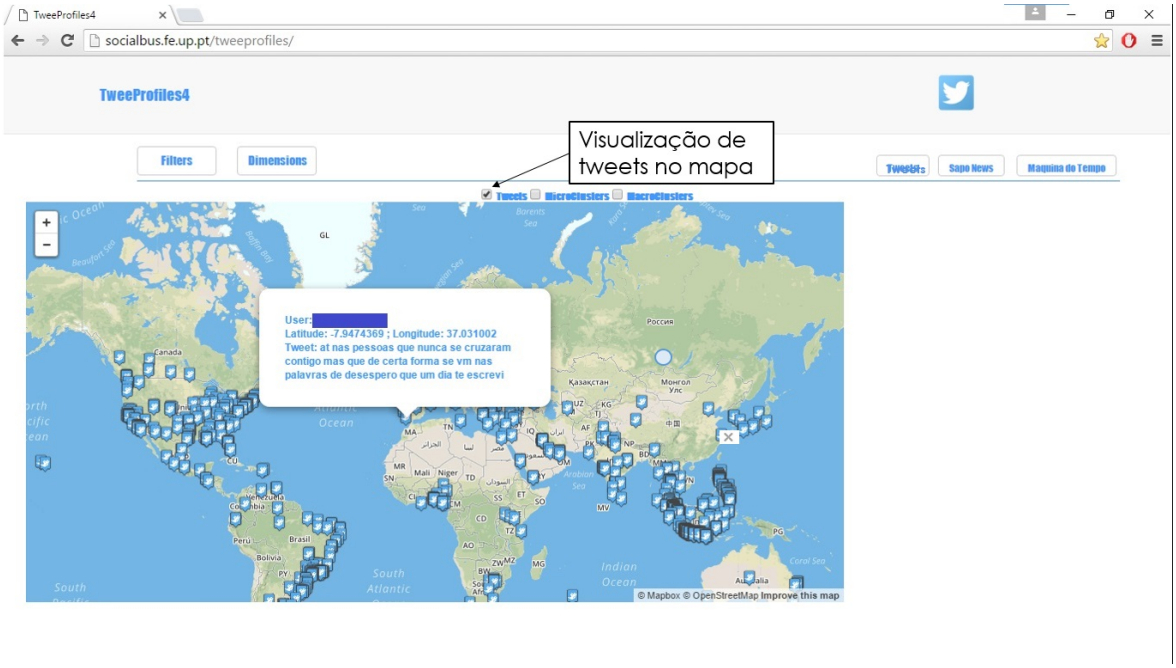
	Editor	Colunista	Pauteiro (chefe de reportagem)	Outro
Caso 1: Exploração baseada em Clusters				
Caso 2: Exploração baseada em Tópicos				
Caso 3: Exploração baseada em Temas				

Casos de uso

De seguida são apresentadas algumas funcionalidades da ferramenta, implementadas e por implementar.

Tweets no mapa

Escolha de visualização de tweets individuais

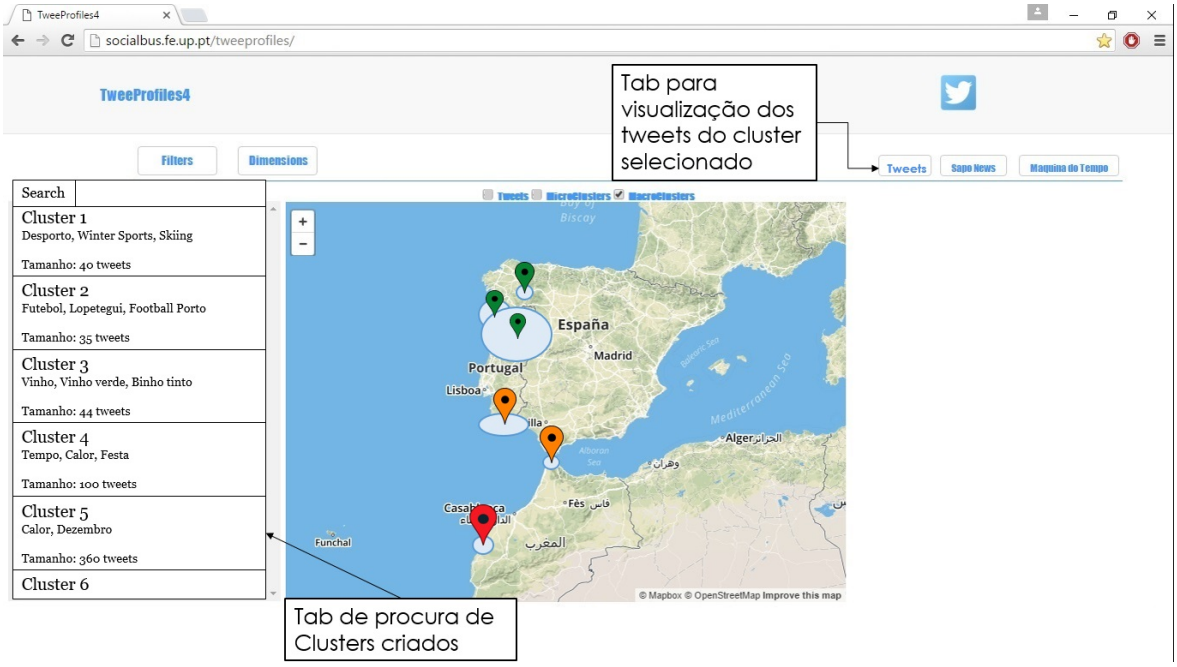


14. **Gostaria de manter a funcionalidade de poder ver tweets individuais no mapa?**
Marcar apenas uma oval.

- Sim
- Não

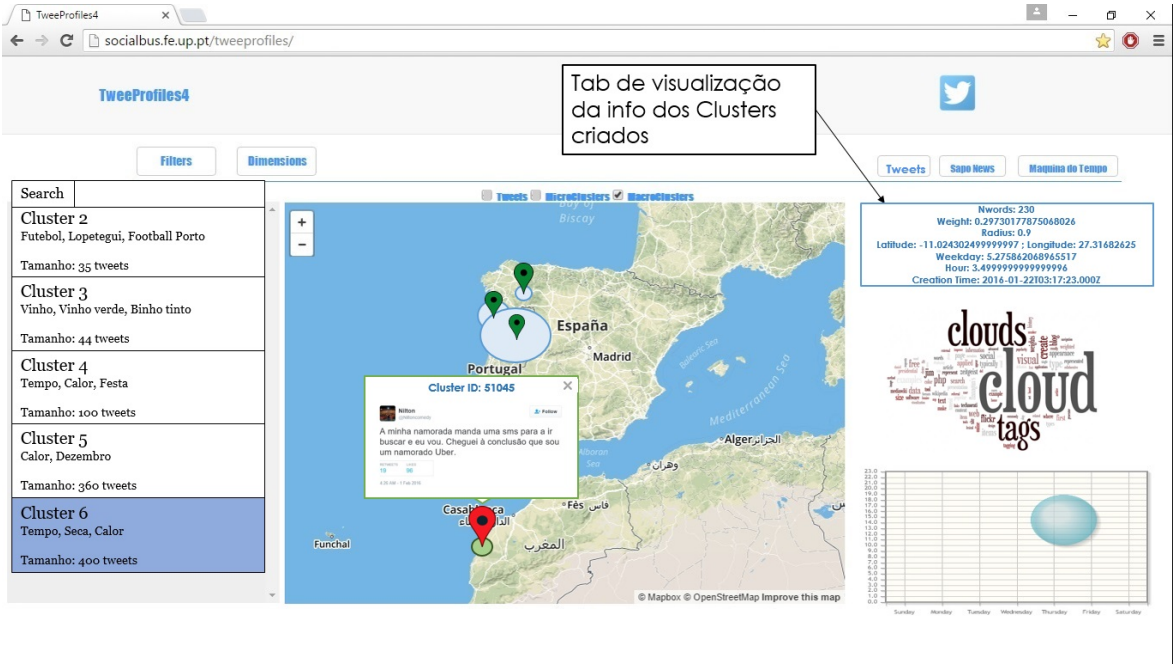
Diferente escolha de clusters e visualização

Escolha de Clusters na esquerda



Visualização dos dados na direita, tweets visíveis na info-

window e na tab Tweets

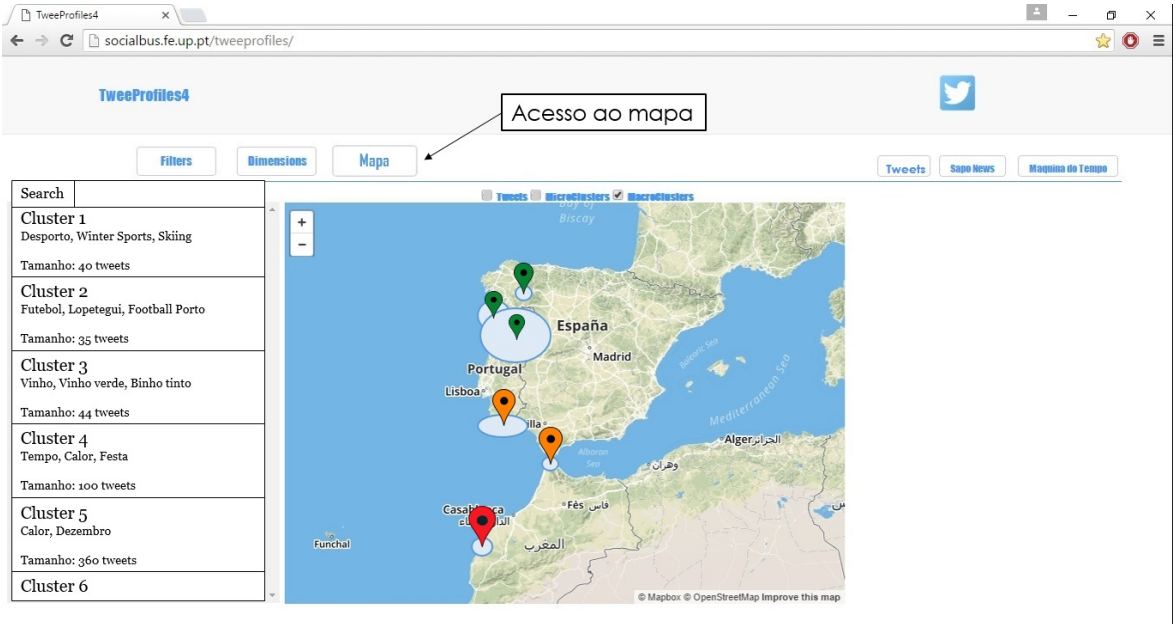


15. **Acha que a implementação desta funcionalidade poderá melhorar a utilização?**
Marcar apenas uma oval.

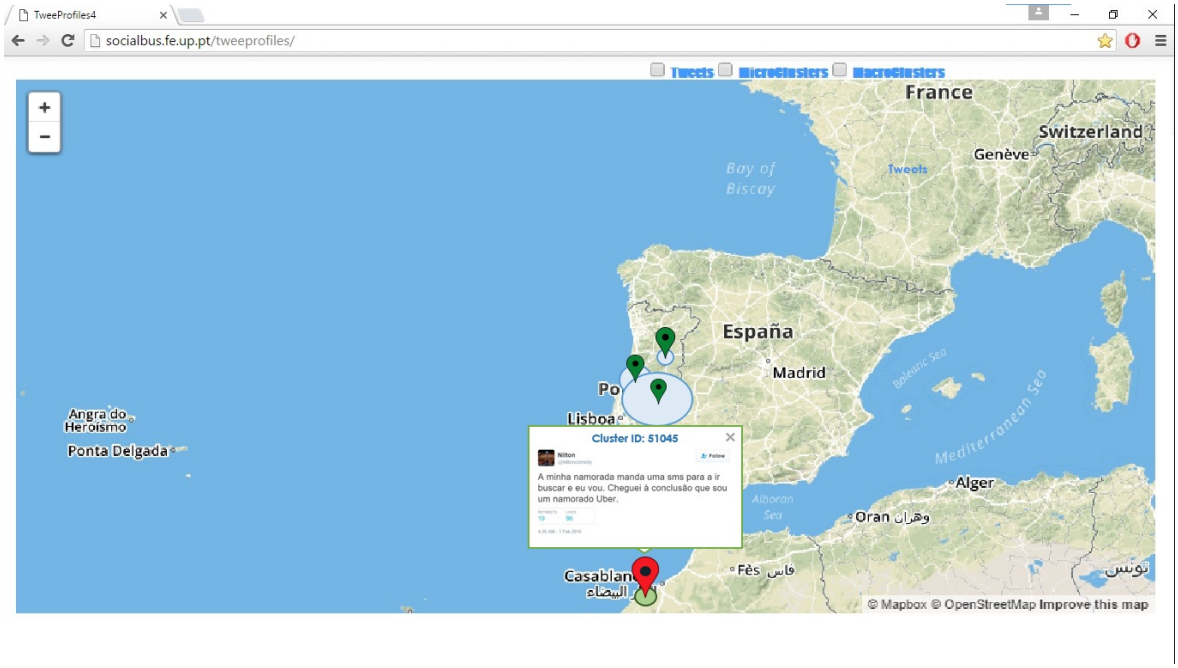
- Sim
- Não

Foco no mapa

Novo botão para acesso à visualização apenas mapa



Foco no mapa, para uma exploração rápida



16. **Acha que a implementação desta funcionalidade poderá melhorar a utilização?**
Marcar apenas uma oval.

- ◯ Sim
- ◯ Não

Opinião

17. **Gostaria de ter alguma funcionalidade não presente nas ilustrações?**

.....

18. **Alteraria algum aspeto da interface?**
cor, forma como o mapa mostra informação, facilidade de aceder a Tweets, etc

.....

19. **Observações e Sugestões**

.....

.....

.....

.....

.....

Contacto

Esta secção tem como propósito estabelecer um canal de troca de informação. As suas perguntas são opcionais, contudo, a sua resposta possibilitaria um futuro impacto na ferramenta. Ao contrário da informação anteriormente introduzida, as informações de contacto não serão disponibilizadas no documento final da dissertação.

20. **Qual é o seu nome?**

21. **Qual é o seu contacto de email?**
    poderá introduzir o seu email pessoal ou institucional

Appendix D

First JPN Inquiry Analysis

Estudo do inquérito conduzido no dia 24 Fev 2016 para avaliação da ferramenta TweeProfiles para apoio à atividade jornalística

Bruno Miguel Alves Vieira

ee07154@fe.up.pt

29/04/2016

Índice

1. Introdução

O inquérito realizado enquadra-se no projeto de dissertação intitulado "Journalism 3.0: Multidimensional Cluster Visualization and Labelling on Twitter Data for Data Journalism", no âmbito do Mestrado Integrado em Engenharia Electrotécnica e de Computadores, da Faculdade de Engenharia da Universidade do Porto com a colaboração do JPN - JornalismoPortoNet.

O projeto pretende ponderar a exploração realizada no TweeProfiles, uma ferramenta de recolha e visualização de tweets, com o intuito de a adequar ao processo jornalístico acarretado por profissionais. O projeto será desenvolvido por Bruno Vieira e supervisionado pelos Engenheiros Carlos Soares e Jorge Teixeira.

O objetivo do inquérito foi aferir a sensibilidade das pessoas para ferramentas de apoio ao jornalismo, com foco na investigação dos dados disponíveis e partilhados nas redes sociais, em particular do Twitter. O conhecimento adquirido será usado para moldar a ferramenta, sendo disponibilizado e discutido com os colaboradores do JPN.

O inquérito ocorreu dia 24 de Fevereiro com início às 14:30 com uma duração de aproximadamente 30 minutos no qual participaram 19 estagiários do JPN, a maioria dos quais são estudantes de jornalismo. A apresentação da ferramenta e posteriores esclarecimentos foram realizados por Carlos Soares (orientador do projeto) e Bruno Vieira (estudante), tendo a sessão sido supervisionada pela Filipa Silva (editora JPN).

A sessão foi aprovada pela Isabel Reis (diretora JPN), Filipa Silva (editora JPN) e Sérgio Nunes (coordenador técnico JPN), aos quais agradecemos esta oportunidade, bem como aos estudantes que participaram.

2. Análise dos dados do inquérito

Os dados, obtidos pelo google forms, apresentam 19 entradas a 26 variáveis. Atendendo às variáveis, o primeiro campo corresponde ao timestamp da entrega, adicionado pelo google forms, sendo os restantes referentes às perguntas feitas. Quanto às entradas, foi notado que 3 delas tinham todos os campos vazios, concluindo que, atendendo à designação de campos obrigatórios no inquérito, estas foram geradas por erros de submissão, possivelmente devido a erro técnico ou humano na submissão das respostas.

De seguida é apresentada a análise do número de respostas de cada jornalista (Figura 1) e de cada pergunta Figura 2.

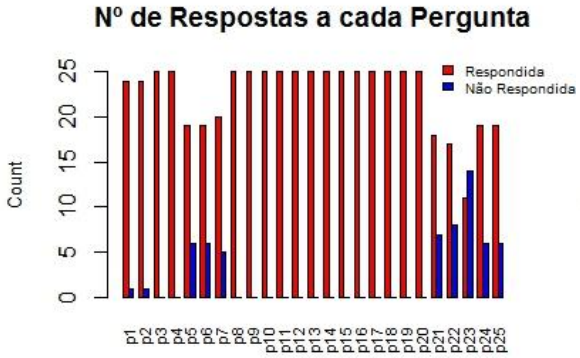


Figura 1. Número de Respostas de cada Jornalista

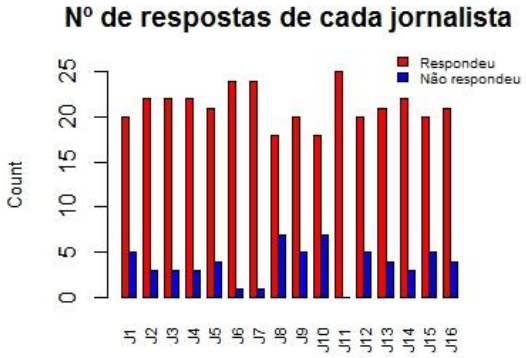


Figura 2. Número de Respostas a cada Pergunta

Analisando as figuras anteriores, os jornalistas aceitaram responder a um número significativo de perguntas, tendo em conta que só 3 das 25 perguntas eram obrigatórias. Sobre as perguntas com menos afluência, na Figura 2 observam-se tendências localizadas em p5/p6/p7 e p21 e posterior. Analisando a primeira tendência, p5 inquiria sobre a experiência profissional prévia (Sim ou Não),

sendo que a resposta "Não" bloquearia as subsequentes p6/p7. Incidindo nas perguntas p21 e posteriores, estas incidem na secção de perguntas abertas sobre a opinião e na secção de pedido de contacto, p24/p25.

Em suma, os dados mostram que o inquérito foi adequado para promover o envolvimento dos participantes.

3. Caracterização dos participantes

Os inquiridos têm idades compreendidas no intervalo dos 20 aos 32 anos, com maior concentração nos 20 anos (Figura 3), com habilitações académicas divididas praticamente em partes iguais entre licenciados e outro, ou seja, a acabar a licenciatura (Figura 4).

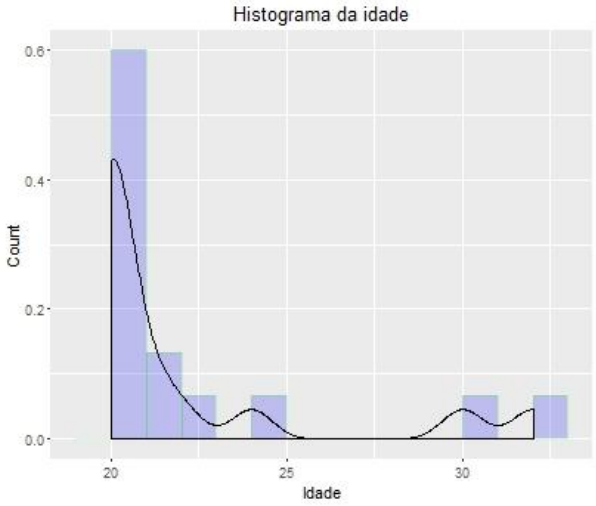


Figura 3. Idade de cada Jornalista

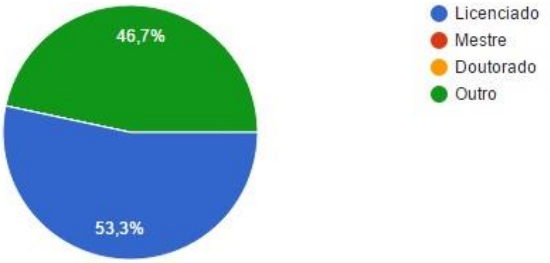


Figura 4. Habilitações académicas de cada Jornalista

A profissão foi indicada como resposta aberta e, como tal, foi preciso tratar as respostas. O resultado pode ser visto na Figura 5, mostrando uma distribuição semelhante entre profissões.

As respostas referentes à experiência numa área relacionada com jornalismo/comunicação social denotam que mais de metade já foi confrontado com os seus desafios (Figura 6).

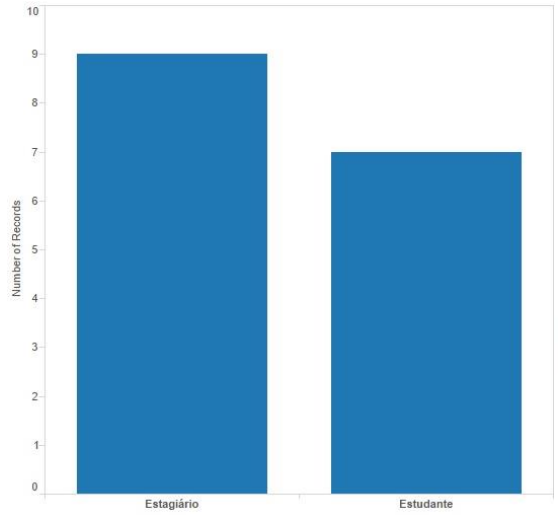


Figura 5. Profissão atual de cada Jornalista

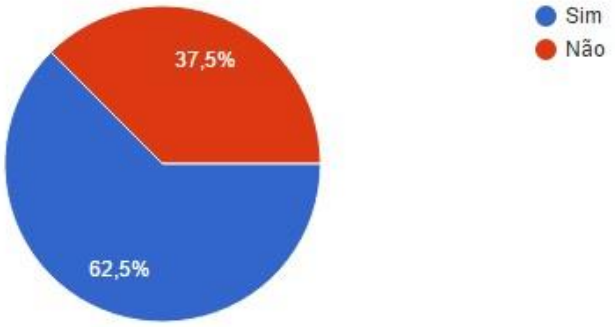


Figura 6. Trabalhou numa área relacionada com o jornalismo/comunicação social

A uma resposta positiva à pergunta anterior, foi pedido esclarecimento quanto ao tipo de trabalho e por quanto tempo o realizou. A função predominante foi a de Jornalista, pouco mais de metade, sendo as restantes de Editor e Redator. Os anos de serviço indicados podem ser vistos na Figura 7 com predominância em 0.1 anos, ou seja, semanas a um mês, o que coincide com a sua estadia no JPN.

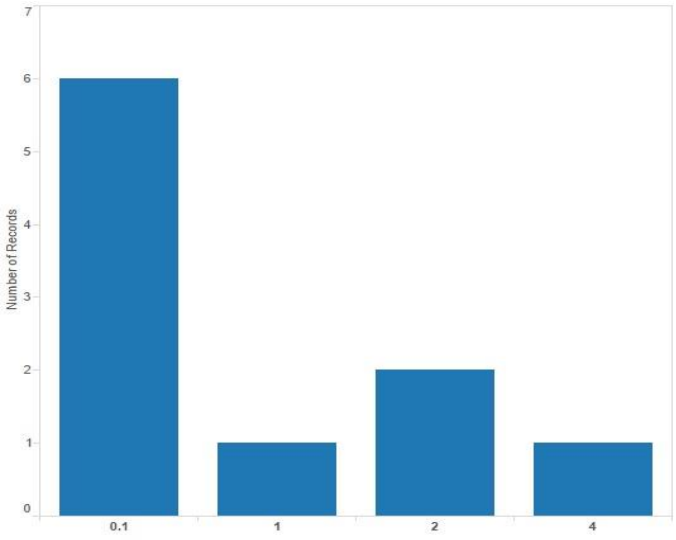


Figura 7. Anos de trabalho na área

A pergunta referente ao uso de redes sociais num contexto profissional revelou um forte apreço pelo Facebook, ferramenta usada por todos os que responderam. Outras ferramentas a notar foram

o Linkedin e Twitter, tendo o Twitter ocupado o 3º lugar. A menção do Linkedin é especialmente interessante, denotando a procura de informação para além das redes sociais generalistas. A sua distribuição pode ser vista na Figura 8.

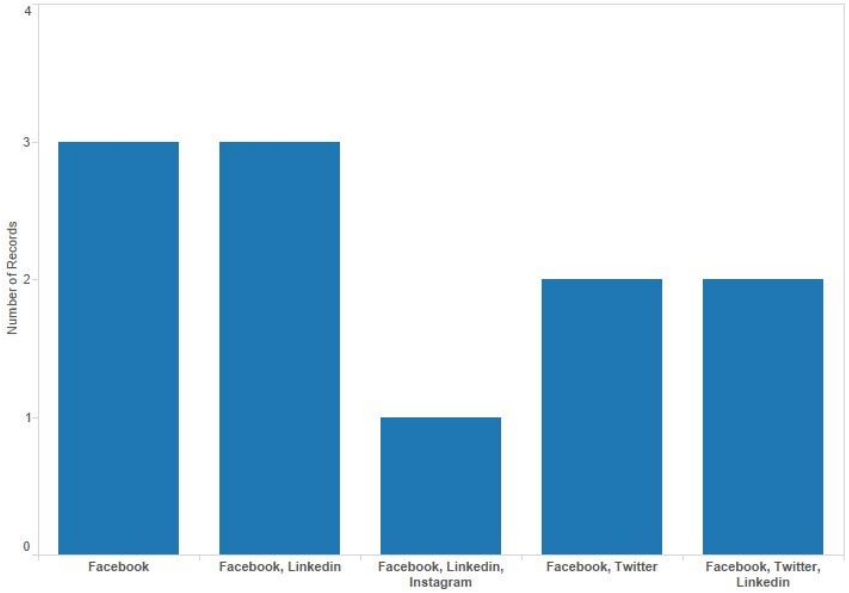


Figura 8. Respostas sobre o uso de ferramentas de exploração das redes sociais no âmbito profissional

As perguntas referentes ao tipo de informação que seria privilegiada na investigação de redes sociais denotam um grande interesse em eventos (Figura 9). Esta informação engloba localização, tópicos, menções (Tweets que mencionam o evento) e ordem cronológica (sequência de eventos). Outras informações de relevo são pessoas e especialmente Organizações.

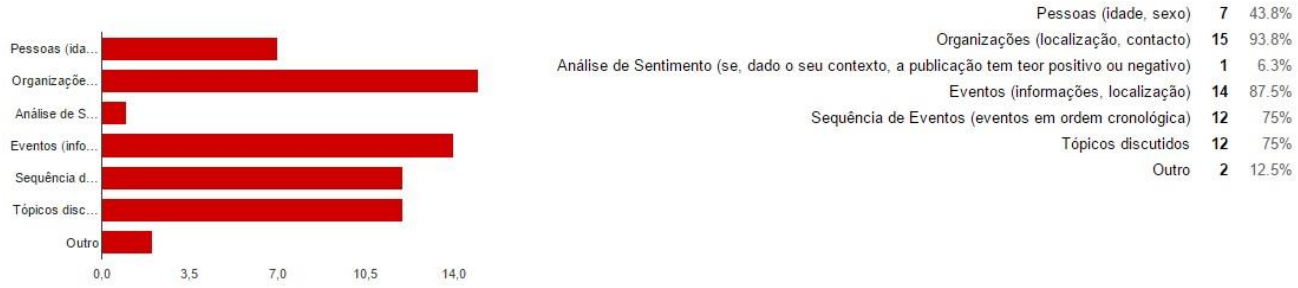


Figura 9. Interesse em informação na ferramenta

4. Escolhas referentes aos Casos de uso

Após inteirar os inquiridos com os casos de uso, mostrando diferentes modos de exploração, a sua adequação a cargos existentes em Jornalismo foi questionada (consultar Anexo 1 para a interpretação dos cargos). A atribuição da conduta que esta ferramenta proporcionaria divide-se essencialmente entre o Editor e o Pauteiro.

Todos os casos apresentados foram considerados maioritariamente adequados e muito adequados à tarefa de exploração de redes sociais.

Quanto ao caso de uso mais apropriado à perceção de trabalho dos inquiridos, a exploração com foco em Clusters e Temas apresenta maior afluência. A Figura 10 mostra a distribuição obtida.

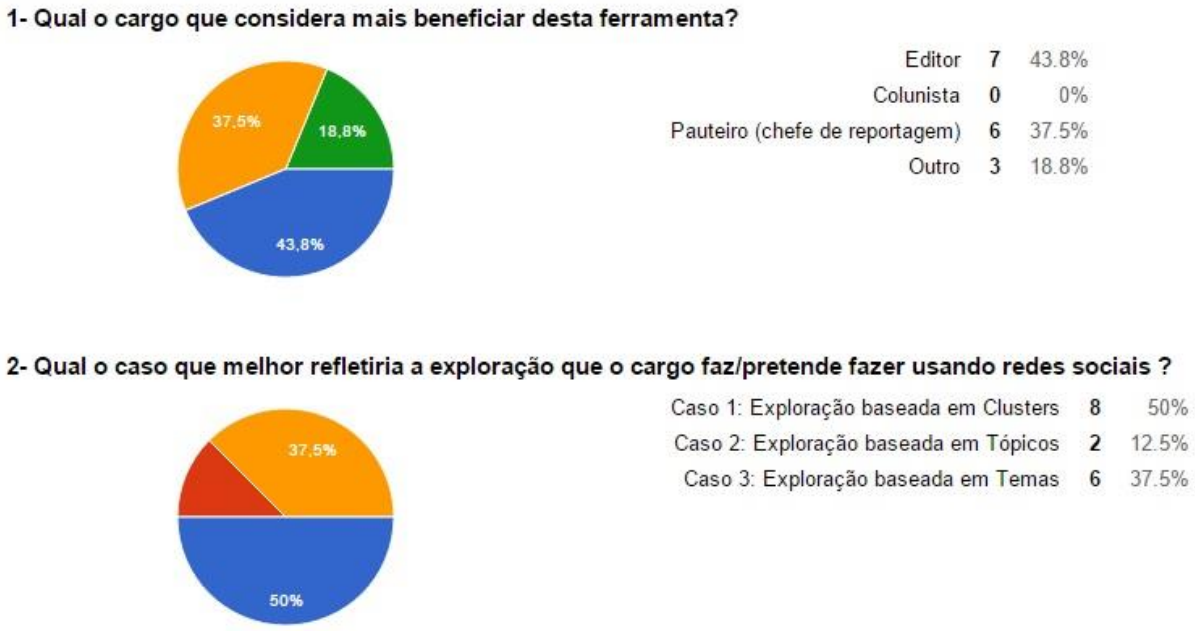


Figura 10. Cargo atribuído e Caso de uso preferido

Tendo em conta que os inquiridos têm níveis variados de experiência, é importante analisar as respostas tendo essa experiência em conta. A escolha do caso de uso por experiência pode ser visto na Figura 11.

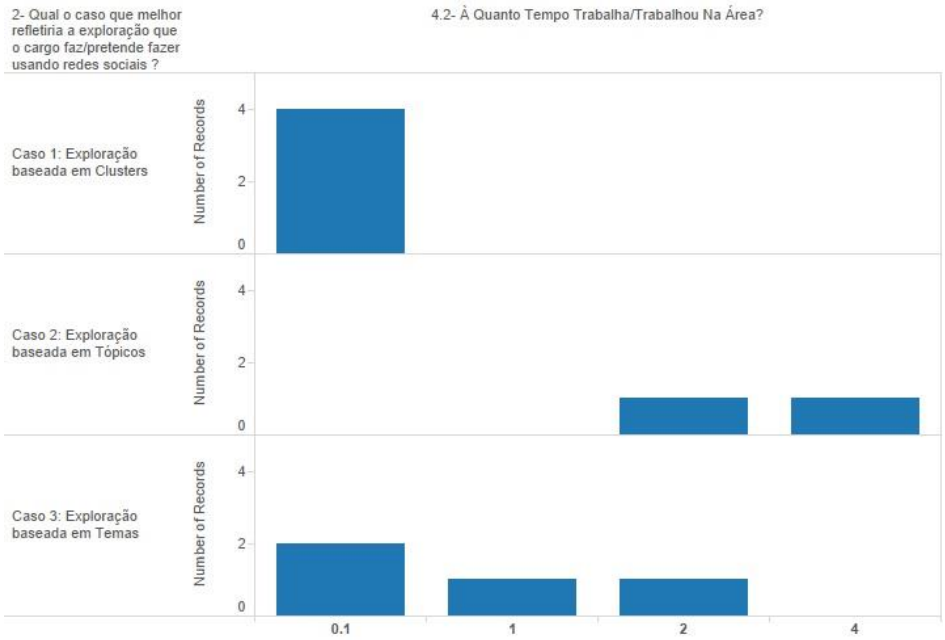


Figura 11. Caso de uso VS Anos de trabalho

Na Figura 11 é notado que o interesse pela exploração de clusters é proveniente de pessoas com pouco experiência profissional. O apreço pelos Casos 2 e 3 é maior consoante os anos de experiência, possivelmente por estarem mais de acordo com a sua atividade corrente. Esta análise não pretende afirmar que a opinião dos inquiridos irá mudar com os anos de trabalho, apenas que, de momento os casos 2 e 3 poderão estar mais de acordo com o dia-a-dia de um trabalhador que usa redes sociais.

As respostas referentes à associação de um cargo a cada caso de uso denotam que, a exploração baseada em Clusters e em Assuntos é maioritariamente atribuída ao Editor, enquanto a exploração baseada em Tópicos é maioritariamente atribuída ao Pauteiro. Tendo em conta as tarefas de cada cargo, os Casos 1 e 3 seriam usados por alguém que decide o que é ou não notícia e o Caso 2 seria usado por alguém na procura de candidatos a notícia. A Figura 12 apresenta os resultados obtidos.

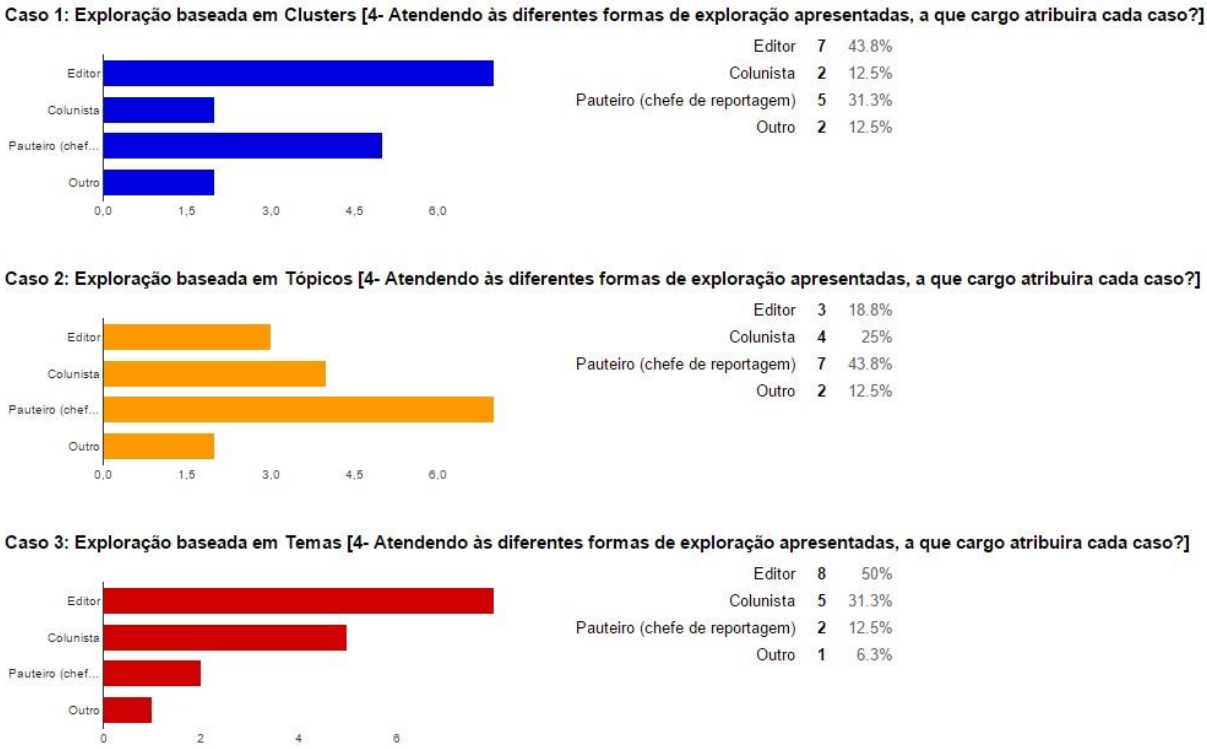


Figura 12. Atribuição de cargos a cada caso de uso

A informação obtida de atribuição maioritária das funções ao Editor vai de acordo com a hipótese referida na primeira reunião no JPN.

As três funcionalidades apresentadas, nomeadamente ver Tweets no mapa, divisão de tarefas na visualização e foco no mapa, foram apreciadas, quase de forma unânime, como de interesse para a ferramenta.

5. Opiniões dos inquiridos

Nesta secção foi pedido a opinião, em resposta livre, referente a três aspetos: funcionalidades não presentes, aspeto da interface e sugestões.

- Funcionalidades

- Gráfico com percentagem de utilizadores a aceder em tempo real aos clusters
- Mais estatística, quem está a twittar determinado assunto por género, idade
- Se for possível analisar também os links partilhados, seria interessante agrupar todos os tweets que partilhassem o mesmo link - quer fosse longo, quer fosse short url.
- Hiperligação para o Twitter dos autores dos tweets apresentados.

As primeiras duas funcionalidades mostram um interesse por mais informação que, até agora, era usada pelo processo de clustering, tendo então sido considerada de pouco interesse. A terceira indica, mais uma vez, que os pressupostos do desenvolvimentonão eram adequados, dado que os URLs são eliminados antes de se analisar o conteúdo dos tweets com o algoritmo de clustering. A quarta funcionalidade está de momento implementada, mas não apresentada.

- Aspeto da interface

- Tornava o template mais apelativo com cores mais relacionadas com o Twitter
- forma como o mapa mostra a informação
- Aspeto mais apelativo
- Diminuia os contrastes entre linhas e investia na interface Visualização dos dados na direita, tweets visíveis na info-window e na tab Tweets

Os aspetos da interface notados estão de acordo com o trabalho que um designer teria, devido a noções de cor e contraste. Quanto à forma como o mapa mostra informação, o comentário não tem informação suficiente para podermos refletir sobre o processo de desenvolvimento da ferramenta.

- Observações e Sugestões

- Excelente ideia! O único desafio será ter a explicação prévia: talvez com recurso a video tutorial ou infográfico, para melhor explicar o conceito e funcionalidade, a uma classe de trabalhadores que por definição tem muito pouco tempo disponível.
- A ferramenta não será útil se não tiver possibilidade de tradução dos tweets e dos hashtags. Os temas de interesse internacional não são escritos apenas em inglês. Clusters, temas ou tópicos não serão identificáveis se estiverem em línguas diferentes.

6. Alterações a considerar no processo de desenvolvimento da ferramenta

Com a informação obtida, as alterações no processo de desenvolvimento incidem na prioridade de implementação dos casos de uso, nos dados apresentados e no aspeto geral da ferramenta.

Quanto aos casos de uso, o Caso 1 – Exploração baseada em clusters terá prioridade acima dos restantes.

Os dados apresentados, referentes a cada cluster, terão de ser reconsiderados de modo a proporcionar a informação extra requisitada pelos inquiridos, deixando de ter o foco operacional.

O aspeto da ferramenta deverá ser alterado para uma melhor associação ao Twitter, forma como os tweets são apresentados.

7. Conclusões sobre o inquérito

O inquérito procurou ponderar a exploração realizada no TweeProfiles, uma ferramenta de recolha e visualização de tweets, com o intuito de a adequar ao processo jornalístico.

O número de respostas obtidas pelos inquiridos foi considerada aceitável à exploração, validando o método e contexto no qual o inquérito decorreu.

Os participantes são estudantes e estagiários, em igual número, dos 20 aos 32 anos cuja experiência profissional incide maioritariamente no intervalo de algumas semanas a 1 mês. A sua experiência no uso de redes sociais num âmbito profissional incide, fortemente, no Facebook com especial menção do Linkedin. A informação privilegiada numa ferramenta de investigação de redes sociais consiste na ligação a pessoas, eventos ou organizações. Após inspeção dos dados, foi notado o interesse em informação adicional como a natureza do vínculo (colaborador permanente, colaborador a tempo parcial, estagiário), o tipo de meio de comunicação social (local/nacional), mídia (tv, imprensa escrita em papel, imprensa escrita online) e o nome da instituição.

O Caso de uso considerado como mais adequeado foi o Caso 1 – Exploração baseada em clusters, com a atribuição da função ao cargo de Editor. As funcionalidades apresentadas, com o intuito de melhorar a experiência de utilização, foram consideradas favoráveis.

As opiniões denotam interesse em informação adicional que tem sido ou considerado como auxiliar de operações ou descartada pelo não uso das funcionalidades, como URLs. A necessidade de apresentar o fluxo de operações da ferramenta foi notada tanto no local como nas opiniões fornecidas.

Anexo 1 – Interpretação dos cargos ligados ao Jornalismo

- O Editor tem cargo mais elevado no jornalismo, com poder de decisão sobre o que é notícia.

- O Pauteiro tem a tarefa de orientar os repórteres na apuração de informação e de os ajudar na prospeção de temas com potencial para virar notícia.

References

[1] Tiago Cunha, Carlos Soares, and Eduarda Mendes Rodrigues. TweeProfiles: Detection of Spatio-temporal Patterns on Twitter. Master's thesis, 2013.

[2] Luís Pereira and Carlos Soares. TweeProfiles4: a weighted multidimensional stream clustering algorithm. Master's thesis, 2015.

[3] A. O. Larsson and H. Moe. Studying political microblogging: Twitter users in the 2010 Swedish election campaign. *New Media & Society*, 14(5):729–747, 2012. doi:10.1177/1461444811422894.

[4] Charity Pradiptarini. Social Media Marketing : Measuring Its Effectiveness and Identifying the Target Market. *Journal of Undergraduate Research*, pages 1–11, 2011.

[5] Twitter. About twitter, 2015. URL: https://about.twitter.com/company.

[6] Kevin Makice. *Twitter API: Up and Running Learn How to Build Applications with the Twitter API*. O'Reilly Media, Inc., 1st edition, 2009.

[7] Matko Bošnjak, Eduardo Oliveira, Jose Martins, Eduarda Mendes Rodrigues, and Luis Sarmento. TwitterEcho - A Distributed Focused Crawler to Support Open Research with Twitter Data. *Proceedings of the WWW 2012, the 21st International Conference Companion on World Wide Web*, pages 1233–1239, 2012. doi:10.1145/2187980.2188266.

[8] REACTION. Socialbus documentation, 2013. URL: http://reaction.fe.up.pt/socialbus/index.html.

[9] André Maia, Carlos Soares, and Pedro Abreu. TweeProfiles3: visualização de padrões espacio-temporais no Twitter. Master's thesis, 2015.

[10] André Maia, Tiago Cunha, Carlos Soares, and Pedro Henriques Abreu. *TweeProfiles3: Visualization of Spatio-Temporal Patterns on Twitter*, pages 869–878. Springer International Publishing, Cham, 2016. URL: http://dx.doi.org/10.1007/978-3-319-31232-3_82, doi:10.1007/978-3-319-31232-3_82.

[11] Twitter. Twitter documentation, 2015. URL: https://dev.twitter.com/overview/documentation.

[12] João Pereira, Tiago Cunha, and Carlos Soares. TweeProfiles2 : real-time detection of spatio-temporal patterns in Twitter. Master's thesis, 2014.

[13] Ronen Feldman and James Sanger. *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, New York, NY, USA, 2006.

[14] Kazi Saidul Hasan and Vincent Ng. Automatic Keyphrase Extraction: A Survey of the State of the Art. *Association for Computational Linguistics Conference (ACL)*, pages 1262–1273, 2014. doi:10.3115/v1/P14-1119.

[15] Khaled M Hammouda, Diego N Matute, and Mohamed S Kamel. CorePhrase: Keyphrase extraction for document clustering. *Machine Learning and Data Mining in Pattern Recognition Proceedinds*, 3587:265–274, 2005. URL: http://www.springerlink.com/index/1a8adr1jmc756ajk.pdf.

[16] Su Nam Kim and Timothy Baldwin. Extracting keywords from multi-party live chats. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 199–208, 2012.

[17] Pucktada Treeratpituk, Pucktada Treeratpituk, Jamie Callan, and Jamie Callan. Automatically labeling hierarchical clusters. In *Proceedings of the 2006 international conference on Digital government research*, pages 167–176, 2006. URL: http://www.cs.cmu.edu/{~}callan/Papers/dgo06-puck.pdfhttp://portal.acm.org/citation.cfm?doid=1146598.1146650, doi:10.1145/1146598.1146650.

[18] Chengzhi Zhang, Huilin Wang, Yao Liu, Dan Wu, Yi Liao, and Bo Wang. Automatic Keyword Extraction from Documents Using Conditional Random Fields. *Journal of Computational Information Systems*, 4:1169–1180, 2008. URL: http://eprints.rclis.org/12305/1/Automatic{_}Keyword{_}Extraction{_}from{_}Documents{_}Using{_}Conditional{_}Rando pdf.

[19] Samhaa R El-beltagy and Ahmed Rafea. KP-Miner : Participation in SemEval-2. (July):190–193, 2010.

[20] Daniele Quercia, Harry Askham, and Jon Crowcroft. TweetLDA: Supervised Topic Classification and Link Prediction in Twitter. *Proceedings of the 3rd Annual ACM Web Science Conference on - WebSci '12*, pages 247–250, 2012. URL: http://dl.acm.org/citation.cfm?id=2380718.2380750, doi:10.1145/2380718.2380750.

[21] S Vijayarani, Ms J Ilamathi, and Ms Nithya. Preprocessing techniques for text mining-an overview. *vol*, 5:7–16.

[22] Teng-Sheng Moh and Surya Bhagvat. Clustering of technology tweets and the impact of stop words on clusters. In *Proceedings of the 50th Annual Southeast Regional Conference*, ACM-SE '12, pages 226–231, New York, NY, USA, 2012. ACM. URL: http://doi.acm.org/10.1145/2184512.2184566, doi:10.1145/2184512.2184566.

[23] Y. Matsuo and M. Ishizuka. Keyword Extraction From a Single Document Using Word Co-Occurrence Statistical Information. *International Journal on Artificial Intelligence Tools*, 13(01):157–169, 2004. doi:10.1142/S0218213004001466.

[24] Slobodan Beliga. Keyword extraction: a review of methods and approaches.

[25] Sungjick Lee and Han-Joon Kim. News Keyword Extraction for Topic Tracking. *2008 Fourth International Conference on Networked Computing and Advanced Information Management*, pages 554–559, 2008. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4624203, doi:10.1109/NCM.2008.199.

[26] Gonenc Ercan and Ilyas Cicekli. Using lexical chains for keyword extraction. *Information Processing and Management*, 43(6):1705–1714, 2007. doi:10.1016/j.ipm.2007.01.015.

[27] Kuo Zhang, Hui Xu, Jie Tang, and Juanzi Li. Keyword Extraction Using Support Vector Machine. *LNCS*, 4016:85–96, 2006.

[28] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022, 2003. arXiv:1111.6189v1, doi:10.1162/jmlr.2003.3.4-5.993.

[29] Wayne Xin Zhao, Jing Jiang, Jing He, Yang Song, Palakorn Achananuparp, Ee-Peng Lim, and Xiaoming Li. Topical keyphrase extraction from twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 379–388, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. URL: http://dl.acm.org/citation.cfm?id=2002472.2002521.

[30] A. Goldberg and Y. Zhou. *Algorithmic Aspects in Information and Management: 5th International Conference, AAIM 2009, San Francisco, CA, USA, June 15-17, 2009, Proceedings*. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2009. URL: https://books.google.pt/books?id=GRprCQAAQBAJ.

[31] Loulwah AlSumait, Daniel Barbará, James Gentle, and Carlotta Domeniconi. *Topic Significance Ranking of LDA Generative Models*, pages 67–82. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. URL: http://dx.doi.org/10.1007/978-3-642-04180-8_22, doi:10.1007/978-3-642-04180-8_22.

[32] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL: http://igraph.org.

[33] Bettina Grün and Kurt Hornik. topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, 40(13):1–30, 2011. doi:10.18637/jss.v040.i13.