Elaborato finale in **Intelligenza Artificiale**

**Incomplete knowledge and Reasoning**

Anno Accademico 2015/2016

Candidato:

**Antonio Pizzata**

matr. N46002005
Contents

1 Introduction ................................. 1
  1.1 Economic Analysis ......................... 1
  1.2 Sentiment Analysis ......................... 2
  1.3 Political Analysis ......................... 2
  1.4 Critics .................................. 2

2 State of Art ................................. 3
  2.1 Text Analysis Tools ....................... 3
    2.1.1 TaLTaC ............................ 3
    2.1.2 GATE ................................ 4
  2.2 Text Mining ............................... 4
  2.3 Social Network Analysis .................. 6

3 Methodology ................................. 7
  3.1 Overview .................................. 7
  3.2 Text Pretreatment ......................... 7
    3.2.1 Normalization ......................... 7
    3.2.2 Correcting Spelling Errors ............ 7
  3.3 Lexical Analysis ......................... 8
  3.4 Textual Analysis ......................... 8

4 Experimental Campaign .................... 9
  4.1 Description ............................. 9
  4.2 TaLTaC 2 Analysis ....................... 9
  4.3 GATE Analysis ........................... 11

5 Experimental Results ...................... 13
  5.1 TaLTaC 2 results ....................... 13
  5.2 Gate results ............................ 16

6 Conclusions ................................ 18

Bibliography ................................ 19
1 Introduction

Social networks analysis is an emerging trend among scholars and researchers in the last years. In literature there are various instruments and project to achieve different kind of analysis, yet in our work we focused on the analysis of the content of the text obtained from the social media. Through our investigation it was possible extracting the hot topics for the different, in our case, reviews studied obtained from the social network Yelp.

But what are social networks? Social networks are web sites where a registered user can do many things: share something that he finds interesting, chat with friends, play games etc. The first everyone recall is of course Facebook, born in 2004 form Mark Zuckerberg[WTFEa], but it is not the only one, indeed others important sites are Twitter, Instagram, Reddit, LinkedIn and more to come. But why do social networks are so widespread? Surely their easy access is a key feature, but the main reason is because people want to reach friends at any time in any places to share ideas, news anytime, anywhere. The use of these web sites changed also our culture and our vocabulary: we are now used to understand sentences like “tweet me” or acronyms like “lol” and “tgif”. In other words Social Networks are predominant in our everyday life.

1.1 Economic Analysis

For their spread many companies are interested in social networks data mining. The data that can be obtained from Facebook, Twitter or any other are very important to organize an advertising campaign, consider that in the last years born the “social media marketing”[WTFEc] a branch of marketing totally focused on social networks. In this way the companies can reach a bigger number of stakeholders and create a more modern concept of advertising: quicker, immediate and sometimes matched with catchy hashtags. Through this new process companies can even track and analyze online conversations about their brands, if their advertising campaign are good enough and they can even support costumers in order to lessen the negative effects that online complains can provide to their brand.[WTFEc]
1.2 Sentiment Analysis

Important information that analyst are interested in are how people feel about an event occurred in their lives or around them. It has been proved[KWM11] that the use of emoticon or hashtags can identify the sentiments the person feels and the use of these forms of communications are the easiest way to share them on the net. This led researchers and scholars to extend the study of sentiment analysis[WTFE1b] to the digital world, mining information through forums, blogs, social media to understand how the people react to and event (i.e. a new law, a terrorist attack). Though nowadays is hard achieving many information from even a small line of text, the studies carry on to create an automated system able to mine emotions from shared texts.

1.3 Political Analysis

Various studies proved that voters can influence each other on social networks. Users of these platforms can easily avoid searching independently information about who to be voted thanks to the spread of ideas and thoughts through social medias.[BdA16] Moreover the support that candidates got during campaigns on online social networks is an important indicator of the candidate success, so nowadays we can assert that having a good online campaign helps in the final results.[WG08]

1.4 Critics

Many have criticized social networks, because these platform can, for example, lead to disparity: not every person can access to them, and so this can induce in many form of differences in the culture or habits of who has them and who does not. As already said many people get information from social medias, but not always those are true. The presence of fake news websites is very popular nowadays, due to the easy way to earn money thanks to the visit they get. That brings us to say that the possibility of earning money in any way on the net induces many to commercialize anything with harsh advertising, spam campaign and clickbaits that are not frowned upon by the consumers. But the main complain that is moved to the social networks infrastructures is the lack of privacy, indeed many data can be collected from user profiles by third party applications or the network itself. Every time somebody publishes, shares or comments something this information is no longer private and it can be used for market surveys or any applications. This problem affects in particular teenagers, who are much more likely to share their personal information and lots of their private lives online. [Soc]
2 State of Art

2.1 Text Analysis Tools

Our experimentation has been realized thanks to the use of the software: TaLTaC² and Gate. The data set analyzed comes from the social network Yelp, founded in 2004, which publishes crowd-sourced reviews about local businesses.[WTFEd] and it is made up of 50 tuples structured this way:

- anonymized user name
- anonymized reviewed place
- date of the review
- review

In order to utilize TaLTaC² functionalities the data are divided in fragments, one for each tuple, and in sections, one for each elements listed before.

2.1.1 TaLTaC²

TaLTaC² (Trattamento automatico Lessicale e Testuale per l’analisi del Contenuto di un Corpus, Lexical and Textual automatic processing for analyzing the Content of a Corpus)[Mor] is a software able to perform on documents and data written in natural language operations like: Text Analysis, Text Mining and Corpus Analysis. It has been developed in Italy from the conjunction of the University of Salerno and University “La Sapienza” of Rome. The first task to do in our environment is creating a “Work Session”, which is the file that is going to contain all our information, then we build the “Corpus”, our main object, which will be analyzed with various instruments and operations. The Corpus is then divided in two parts: fragment and section. The first is identified by four asterisks (****) a name and some variables, if needed, identified themselves with an asterisk, a name and a value. The latter is defined with four plus signs (++++) and a name. Every Corpus can contain more fragments and sections. One of the main functions of TaLTaC² is the extraction of significant information from the Corpus (Text Mining), such task uses endogenous and exogenous resources: the former is composed of the number of fragments and the categorical variables which can be associated to the text in order to identify fraction of the corpus logically related. Thanks to this kind of resource TaLTaC²
is able to perform a “Specificity Analysis”. The exogenous resources are lists which contain the frequency of a term or lexical unit, thanks to these lists the software is able to identify peculiar language of the text.

2.1.2 GATE

GATE (General Architecture for Text Engineering) [pt] is a open source free software which excels at Text Analysis. The first version was written in the mid-1990s and it has reached currently version 8. The main resource of this program is ANNIE (A Nearly New Information Extraction System), which provides a lot of information extraction techniques such as the English Tokeniser which splits the text into annotations of type Token or the POS Tagger that assigns to every token an annotation that describes it characteristic (i.e NNP for Proper Noun in singular form) and more. The Gate software is a family made up of:

- an IDE, Gate Developer
- a web app, Gate Teamware
- a framework, Gate Embedded

Due to Operative System compatibility on our machine we are going to use version 6.1.

2.2 Text Mining

Effective and efficient access to domain relevant information requires the ability to automatic process and organize the information especially if these are contained in huge repositories of data[BDS14, BDS11]. The most used approaches in Big Data processing are based on the graph algorithms, parallel and distributed architecture. Some Big Data infrastructures deal with Apache Hadoop [Apa] software for data-intensive distributed applications, based in the MapReduce programming model and a Distributed File System (Hadoop). MapReduce job splits the input data set into independent subsets that are dealt with map tasks in parallel. This step of mapping is then followed by a step of reducing tasks. These reduce tasks use the output of the maps to compute the result of the job. Some open source tools for Big Graph mining are proposed, as Pegasus, a big graph mining system built on top of MapReduce.

It allows to find patterns and anomalies in massive real-world graphs. Another Big Data Mining initiative is Apache a scalable machine learning and data mining open source software based mainly on Hadoop and a collection of hardware, software and design patterns for managing very fast large-scale data at very low cost and using BIDMat an interactive matrix library that integrates CPU and GPU acceleration.
For what concerns the text analysis, Morphosemantic approaches similar to the one proposed here have been already proposed for many languages and applied to the medical domain. Works that deserve to be mentioned are Pratt on the identification and on the transformation of terminal morphemes in the English medical dictionary; Wolff on the classification of the medical lexicon based on formative elements of Latin and Greek origin; Pacak et al. on the diseases words ending in -itis; Norton e Pacak on the surgical operation words ending in -ectomy or -stomy; Dujols et al. on the suffix -osis. Between the nineties and the 2000, many studies have been published on the automatic population of thesauri, we recollect among others Lovis et al., that derives the meaning of the words from the morphemes that compose them; Lovis et al. that identifies ICD codes in diagnoses written in different languages; Hahn et al. that segments the subwords in order to recognize and extract medical documents; and Grabar e Zweigenbaum that uses machine learning methods on the morphological data of the thesaurus SNOMED (French, Russian, English). Several works focused the problem to the definition and the implementation of a comprehensive architecture for information structuring, while the work is dedicated to resolve the issue of ensuring semantic interoperability of different entities by mapping the content of different corpora on a set of shared concepts.

For what concerns the decision support system in literature they are usually categorized in two typologies, Knowledge-based and non-Knowledge-based [Coi15], [MMG14]. The firsts are accurately described in [Mil94]. AAPHelp, created in 1972, was an early attempt to implement automated reasoning under uncertainty. Other systems are Asbru, EON and PRODIGY [VCP+95]; PROforma, SAGE [TCG+07]; and the Clinical Reminder System [KCF12]. The last one is based on the [SRG+96] Evidence-based medicine and provides evidence-based clinical guidelines. A more detailed and systematic overview on many other CDSS is described in [GAM+05]. Recently, many studies focused on medical information extraction from structured or unstructured texts. Fette [FEW+12] presents a IE systems that integrates medical unstructured information into a clinical data warehouse to transform into a structured format the information inserted by physicians in a clinical information system. Rink [RHR11] proposes a method for the automatic extraction of medical concepts and relations from electronic medical reports. Medical concepts are extracted with supervised machine learning algorithms. Several knowledge sources are used for feature extraction: a semantic role labeler, a POS tagger, a phrase chunk parser, WordNet, Wikipedia and the General Inquirer lexicon. Doan [DBK+10] introduced an automated system to extract medications and related information from discharge summaries. The researchers developed an integrated system adapting some existing NLP tools. In order to efficiently process huge amount of data, several approach regarding hardware implementation of data processing tools are developed. In particular, in [BMV15] and [BDPG+15] authors proposed an hardware implementation of a Decision Tree based multi-classification system traffic analyzer. The system is able to deal with a huge amount of data and tight constraints, such as power consumption and hardware resources. In [BDBMV15] a traffic analysis hardware
accelerator, based on the Decision Tree model, is presented through an infrastructure which collects data from mobile devices and provide them update versions of the analyzer by exploit new traffic information. Moreover, in the field of data protection, in [CBM14] authors proposed a secure infrastructure to protect intellectual property installed on the FPGA by means of partial dynamic configuration.[Ama16]

2.3 Social Network Analysis

The predominant approach to analyze social network is the graph theory, even though it is largely debated. This theory derives from the studies of Euler and provides us a way for studying Networks of any kind. In social networks the single user or groups of users are represented as a point and their relation are represented as lines. The data obtained from these graphs are then recorded in matrix form, in this way we can study directly the data without drawing the graph, that helps a lot when we are facing large social network data sets. To the lines connecting points in the diagram we can assign a direction in order to determine which point influence the other and to that influence we can also assign a value to represent the strength of that relation.[Sco11]
3 Methodology

3.1 Overview

In order to achieve a fine analysis of our data we applied different types of operations.

3.2 Text Pretreatment

This is the first process to be applied on text, in order to obtain a clear analysis in the successive processes. It is made up of:

3.2.1 Normalization

This phase allows to remove any data duplication and it normalizes the writings of names, acronyms and other entities. In order to achieve these goals we have to execute various tasks:

- Change apostrophes into stresses (for the words that is needed), in order to determinate the right word.
- Label words/sequence of words so that they can assume the right meaning and are not mistook with others expressions (i.e. a name can be mistake for a noun, Rose or rose).
- Change of capital letters into lowercase for the words that are not labeled, if one is labeled we need to analyze what the label says: if it is a name (of person, of a city or a general proper name) then it will keep the capital letter, in other case it will not.

3.2.2 Correcting Spelling Errors

This phase consists of comparing a misspelled word with the system dictionary in order to correct and analyze it in the right way.
3.3 Lexical Analysis

Lexical analysis analyzes the segments of the Corpus (a segment is a sequence of graphic forms separated by a strong divider), but how to determine a segment? Well it can be easily be defined by choosing a set of dividers (i.e. punctuation such as “.”,”;”) and then separating the sentences between these elements. Once we obtained our segments we can estimate various analysis parameter such as the IS index: this measures the level of absorption of the segment regard the single elements which it is made of. Other important operations in this part of the analysis are the Tagging, which links to every word a description of the grammatical or semantic characteristics, and the Lexation which identifies the sequence of words defined during the pretreatment as one unique entity. Last we define the Corpus’ key words by studying the repetition rates , the ones that have a noticeable standard deviation (considering only the integers) can be assumed being meaningful.

3.4 Textual Analysis

The first step in this analysis is the study of the Concordances in which we can examine the contest where every word or segment we choose is. Then we calculate the TFIDF rate which sorts the research’s results according to the frequency and distribution of the search keyword in the documents provided. The TFIDF is equal to

\[ tf \times \log\left(\frac{N}{n}\right) \]

where \( tf \) is number of occurrences of an element, and the remaining part is the logarithm of the ratio between the number of documents building the Corpus \( (N) \) and the number of documents which present that element \( (n) \). Another import part of Textual Analysis is the co-occurrences identification, where with co-occurrences we identify those couple of near elements that repeat in the text. This identification is useful to define the primary concepts contained in the Corpus.
4 Experimental Campaign

4.1 Description

Our experimentation aims to analyze online social networks’ datasets in order to derive as much information as possible.

4.2 TaLTaC$^2$ Analysis

Prior starting the analysis we pass through the Text Pretreatment, and so after the parsing of the Corpus we normalize (Figure 4.1) it and compute the sub-occurrences.

The next phase is the Textual Analysis, here we start with the identification of the segments (Figure 4.2), these are saved in two files located here: C:\Documents and Settings\PROGETTO\Desktop\testo\progetto, the former is “Lista dei Segmenti (con indice IS)” which contains the segments with their relative number of occurrences, number of elements forming the segment and the IS index. The latter, named: “Lista dei segmenti Significativi” is a list of the significant segments. Then we analyze the specificity of our Corpus together with the computation of the TFIDF index, these data are saved in the same location under the name of: “Vocabolario” and “Lessico”. The last operation of our study is the Textual Analysis with the concordances and co-occurrences computation (Figure 4.3), the latter can be found in the file “Cooccorrenze e collocazioni significative”.

The results of these operations show that the system without knowing anything of the data submitted can retrieve meaningful information, such as the TFIDF index that shows how much a word is important in the document, or the co-occurrences which show the main concepts of the text thanks to the couple of words that recur all the time. Unfortunately TaLTaC$^2$ can not perform all the analysis on our data set due to the fact that it is in English and so our work is not totally complete. Even though our results are few, they are very significative.
Figure 4.1: Text Normalization window

Figure 4.2: Segments individuation window
4.3 GATE Analysis

Our first operation with this software is the initialization of ANNIE with Defaults, once all the Processing Resources are loaded we run ANNIE and start in sequence (Figure 4.4):

- **Document Reset PR**
  The document reset resource enables the document to be reset to its original state, by removing all the annotation sets and their contents.

- **ANNIE English Tokeniser**
  The tokeniser splits the text into very simple tokens such as numbers, punctuation and words of different types.

- **ANNIE Gazetteer**
  The role of the gazetteer is to identify entity names in the text based on lists.

- **ANNIE Sentence Splitter**
  The sentence splitter is a cascade of finite-state transducers which segments the text into sentences.

- **ANNIE POS Tagger**
  The tagger produces a part-of-speech tag as an annotation on each word or symbol.

- **ANNIE NE Transducer**
• ANNIE OrthoMatcher

The Orthomatcher module adds identity relations between named entities found by the semantic tagger, in order to perform coreference.

Figure 4.4: ANNIE Pipeline

Thanks to these operations we can overcome TaLTaC2 limits and run a Grammatical Tag on our data set. It must be said that our original data have been modified, indeed we are going to analyze only the review section due to the fact that in GATE the other information are meaningless.
5 Experimental Results

From our initial data set, made up of more than 5000 tuples, we extracted, as said, our 50 samples from which we obtained these results:

5.1 TaLTaC\textsuperscript{2} results

Thanks to TaLTaC\textsuperscript{2} functions we can say that our peculiar lexicon, shown in Figure 5.1, is composed by words with highest occurrences and TFIDF like: “games”, “beer”, “good”, “great”, “food”, “place” so it is reasonable to assume that the reviews analyzed talk about a place to eat food, drink beer or play some games and that the main audience thinks that it is a good place.

Other meaningful data are the co-occurrences which are reported in Figure 5.2, these elements strengthen our thoughts about the place reviewed with expressions as: “the bar”, “a good”, “a place”, “beer selection”, “the games”. Last we took a meaningful word, “food”, and studied its concordances through all the Corpus (Figure 5.3), in order to understand the context of the lemma.

<table>
<thead>
<tr>
<th>Vocabularies</th>
<th>Occurrence Total</th>
<th>Longhezze</th>
<th>TFIDF</th>
<th>Occurrence Total</th>
<th>Longhezze</th>
<th>TFIDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>was</td>
<td>83.03</td>
<td>3.77217</td>
<td></td>
<td>there</td>
<td>38.04</td>
<td>1.83534</td>
</tr>
<tr>
<td>is</td>
<td>62.01</td>
<td>2.73499</td>
<td></td>
<td>has</td>
<td>21.04</td>
<td>1.432042</td>
</tr>
<tr>
<td>for</td>
<td>71.00</td>
<td>2.61615</td>
<td></td>
<td>these</td>
<td>19.04</td>
<td>1.76896</td>
</tr>
<tr>
<td>games</td>
<td>34.05</td>
<td>2.46866</td>
<td></td>
<td>there</td>
<td>24.03</td>
<td>1.79625</td>
</tr>
<tr>
<td>as</td>
<td>37.02</td>
<td>2.36757</td>
<td></td>
<td>get</td>
<td>20.03</td>
<td>1.70192</td>
</tr>
<tr>
<td>beer</td>
<td>24.04</td>
<td>2.09485</td>
<td></td>
<td>great</td>
<td>20.05</td>
<td>1.72396</td>
</tr>
<tr>
<td>are</td>
<td>34.03</td>
<td>2.10508</td>
<td></td>
<td>as</td>
<td>17.02</td>
<td>1.77996</td>
</tr>
<tr>
<td>to</td>
<td>34.02</td>
<td>2.07666</td>
<td></td>
<td>with</td>
<td>37.04</td>
<td>1.85699</td>
</tr>
<tr>
<td>in</td>
<td>37.01</td>
<td>2.06400</td>
<td></td>
<td>had</td>
<td>27.03</td>
<td>1.75749</td>
</tr>
<tr>
<td>good</td>
<td>37.04</td>
<td>2.06401</td>
<td></td>
<td>it</td>
<td>17.05</td>
<td>1.79809</td>
</tr>
<tr>
<td>selection</td>
<td>25.08</td>
<td>2.00825</td>
<td></td>
<td>place</td>
<td>47.05</td>
<td>1.75902</td>
</tr>
<tr>
<td>you</td>
<td>45.02</td>
<td>2.04466</td>
<td></td>
<td>our</td>
<td>27.03</td>
<td>1.72304</td>
</tr>
<tr>
<td>were</td>
<td>37.01</td>
<td>1.98888</td>
<td></td>
<td>but</td>
<td>131.01</td>
<td>1.75042</td>
</tr>
<tr>
<td>there</td>
<td>29.05</td>
<td>1.97596</td>
<td></td>
<td>run</td>
<td>13.03</td>
<td>1.70798</td>
</tr>
<tr>
<td>that</td>
<td>57.04</td>
<td>1.85741</td>
<td></td>
<td>of</td>
<td>18.04</td>
<td>1.65514</td>
</tr>
<tr>
<td>Great</td>
<td>7.05</td>
<td>1.02191</td>
<td></td>
<td>was</td>
<td>52.03</td>
<td>1.63081</td>
</tr>
<tr>
<td>in</td>
<td>56.02</td>
<td>1.91502</td>
<td></td>
<td>from</td>
<td>24.04</td>
<td>1.65308</td>
</tr>
<tr>
<td>on</td>
<td>47.02</td>
<td>1.90979</td>
<td></td>
<td>against</td>
<td>9.05</td>
<td>1.64295</td>
</tr>
<tr>
<td>food</td>
<td>34.04</td>
<td>1.87461</td>
<td></td>
<td>too</td>
<td>18.02</td>
<td>1.62190</td>
</tr>
<tr>
<td>of</td>
<td>25.01</td>
<td>1.66714</td>
<td></td>
<td>is</td>
<td>20.01</td>
<td>1.59425</td>
</tr>
<tr>
<td>the</td>
<td>91.02</td>
<td>1.88000</td>
<td></td>
<td>have</td>
<td>46.04</td>
<td>1.58942</td>
</tr>
<tr>
<td>LexCorolbashiL12070a1b0c1d0e1f0g0h0i0j0k0l0m0n0o0p0q0r0s0t0u0v0w0x0y0z0</td>
<td>26.22</td>
<td>1.08947</td>
<td></td>
<td>but</td>
<td>41.03</td>
<td>1.57269</td>
</tr>
<tr>
<td>all</td>
<td>18.00</td>
<td>1.06001</td>
<td></td>
<td>is</td>
<td>11.03</td>
<td>1.55776</td>
</tr>
<tr>
<td>and</td>
<td>19.00</td>
<td>1.07490</td>
<td></td>
<td>is</td>
<td>11.03</td>
<td>1.58117</td>
</tr>
</tbody>
</table>

Figure 5.1: Peculiar lexicon by occurrences
Figure 5.2: Co-occurrences report
<table>
<thead>
<tr>
<th>ID Fr.</th>
<th>Ingrido metric</th>
<th>Forna gratica</th>
<th>Intropo desire</th>
</tr>
</thead>
<tbody>
<tr>
<td>fragm...</td>
<td>...and authentic. If you're looking for good fish...</td>
<td>food</td>
<td>and a cold pint, you can't go wrong at the Pour House</td>
</tr>
<tr>
<td>fragm...</td>
<td>...worth seeking out. They have some of the best fish...</td>
<td>food</td>
<td>I've had in Pittsburgh. The colcannon is awesome and...</td>
</tr>
<tr>
<td>fragm...</td>
<td>...out of this world. If you're not looking for fish...</td>
<td>food</td>
<td>...then try the grilled cheese and make sure you ask...</td>
</tr>
<tr>
<td>fragm...</td>
<td>...there on a Saturday night with a mind to try the fish...</td>
<td>food</td>
<td>...Apparently, we were out of luck. I've always thought...</td>
</tr>
<tr>
<td>fragm...</td>
<td>...neces are of restaurant success is to actually stock...</td>
<td>food</td>
<td>...for people to eat. He told us before we ordered that...</td>
</tr>
<tr>
<td>fragm...</td>
<td>...were out... At that point, realizing that the only...</td>
<td>food</td>
<td>...to be lodged in the place was what was cranked on the...</td>
</tr>
<tr>
<td>fragm...</td>
<td>...head.982961becWxv9393p4j 2011-08-20 Best fish...</td>
<td>food</td>
<td>...in the Bush. Great bar food too. The service is...</td>
</tr>
<tr>
<td>fragm...</td>
<td>...2014-08-23 Best Irish food in the Bush. Great bar...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...Excellent wings and sandwiches, generally good...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...1-pCu.O065h5H676H6w6H2p... PUBLIC_27.jpg 2011-09-15...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...wonderful character and ambiance of this place, but the...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...2 people walking all the tables in both rooms. Our...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...were pretty good, also. If you are looking for good...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...beer selection, but it won't be for a while. The...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...op35we4AM9pQp.2011-05-15...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...Miss, I've only gone to Duke's one time, but the...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...very attentive and it didn't take long to get our...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...place to catch a game, drink some beers and get great...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...The biggest downfall of Duke's, though, is the...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...favorite, I wish I could purchase it. I am not a big bar...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...back tonight after a hoot and nothing out on getting...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...and required at least a second side of dressing. The...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...beer but I think where they really shine is in the...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...this is definitely one of your best choices. The...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...a full bar and some great craft beer options. The...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...selection of appetizers that go beyond the standard...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...LPUB1n6G681nJa580qy80l40 2011-06-10 Love duke...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...wouldn't plan to any work gatherings. Their...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...how to serve and the place is actually stock...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...suggest it to everyone who likes good beer and good...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...I 2008-01-27 . DBS is nice once in a while but the...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...I know how much you are spending. Menus for drinks and...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...But most people don't. Come to Dave &amp; Blues is too...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...you to buy a glass to pick up your Bill. Next, the...</td>
<td>food</td>
<td>food</td>
</tr>
<tr>
<td>fragm...</td>
<td>...cheap but make for great dinner days. As for the...</td>
<td>food</td>
<td>food</td>
</tr>
</tbody>
</table>

**Figure 5.3:** Concordances of the word “food”
5.2 Gate results

Once all the Annie computations have run our Corpus should look like Figure 5.4 where we can see that thanks to the Tokeniser we can distinguish spaces from words, and thanks to the Gazetteer and POS Tagger every word in our dataset has a description, as shown in Figure 5.5.

![ANNIE Tokeniser result](image-url)

**Figure 5.4:** ANNIE Tokeniser result
5.2 Gate results

Figure 5.5: Description of the word provided by ANNIE Gazetteer and POS Tagger

At the end of our analysis with this software we can assert the grammatical features of our data and we can give even more meaning to the analysis described in section 5.1.
Social networks analysis is an emerging trend among scholars and researchers in the last years. In literature there are various instruments and project to achieve different kind of analysis, yet in our work we focused on the analysis of the content of the text obtained from the social media. Through our investigation it was possible extracting the hot topics for the different, in our case, reviews studied.

We are very satisfied of our work and we can easily assert that this study can be the starting point of further analysis on the domain of cybersecurity, through the detection of text containing dangerous messages, viral market advertising, thanks to an analysis of the feedback from the users or costumers, or information crawling.
Bibliography


Chapter 6 Bibliography


Bibliography


