ALMA MATER STUDIORUM – UNIVERSITÀ DI BOLOGNA

DEPARTMENT OF STATISTICAL SCIENCES “PAOLO FORTUNATI”
Bachelor in Statistical Sciences - Curriculum Stats&Maths

APPLICATION OF THE POIROT METHODOLOGY TO RESTAURANTS’ REVIEWS.

Thesis in
Data Mining, Text Mining And Big Data Analytics

Supervisor
Prof. Claudio Sartori

Co-supervisor
Prof. Gianluca Moro

Presented by
Matteo Rossi Reich

Second Graduation Session
Academic Year 2020 – 2021

KEYWORDS

NLU
LSA
pLSA
LDA
POIROT

So long, and thanks for all the fish.

Introduction

Natural Language Processing is the domain of science concerned with analyzing and processing textual information, as the name suggest this kind of data is not codified and ready to use, but indeed natural. This poses various challenges, words can have different meanings based on the context in which they are used or on the user and frequently to understand text is necessary to have some knowledge of what lies out of it. One particular application of NLP is to find explanations for a phenomena based on related text, the POIROT method was developed by professor Gianluca Moro for this exact purpose. This thesis presents the methodology and to shows one of its possible applications, given a dataset of restaurants review made available by Yelp, the aim is to find which are the reasons behind a bad review. It was chosen to focus on extremely negative reviews because they are the ones that even in a small amount can influence a restaurant making it fall off the charts, for this reason it is my belief that there is a lot to be learned from them.

Index

List of Figures

Chapter 1

The Language Models

This chapter explains three different language models that can be used for the POIROT methodology.

1.1 Latent Semantic Analysis

LSA is a technique which allows to extract meaning of words from a corpus of text and to produce measures of similarity on these words. It is based on a specific concept of meaning for which the meaning of a word comes from the context in which it is used and hence its representation in the space will depend on the words around which it did or did not appear [1].

> *The complete meaning of a word is always contextual, and no study of meaning apart from context can be taken seriously.*[2]

Latent Semantic Analysis is a statistical method which creates a vector based representation of text in a vector space in which latent semantic relationships is captured, similar words should end up in near neighbors of the latent space, and hence the similarity between a query and a documents can be computed also when the two don't share terms. This method leverages on the corpus-based representation of terms given by the term-document matrix, the term weighting applied to it and on singular value decomposition which allows to compress all this information in a smaller space [3]

1.1.1 Singular Value Decomposition

SVD is the linear algebra technique behind the "latent" part of LSA, indeed it tries to make latent relationships between the terms evident by re-orienting the vector space and ranking its dimensions [4]. This latter property makes it also possible to reduce dimensionality still being sure that such representation is the best for that dimension. SVD decomposes a matrix C into the product of three matrices as:

$$C = U\Sigma V^T \tag{1.1}$$

where the rows of U are the term vectors in the latent space and the columns of V are documents vectors while the diagonal matrix Σ holds the ranked eigenvalues. An important part of this procedure is to chose the right dimensionality k to which to shrink the space, this has to be done empirically. To choose the right number of dimensions one has to pick a knee point, as described in [5] this point will coincide with one of the minima given by the curvature function as shown in the picture below. SVD ensures that the newfound matrix C_k is determined with the optimal decomposition in the the Frobenius norm.

1.2 Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis is a statistical model developed to deal with dyadic data, data whose observation are made in *dyads(x,y)*[6], like the co-occurrences of the term-document matrix. To overcome the problem of sparseness in the term-document matrix pLSA uses a finite mixture model which assumes the existence of a latent variable. Given a corpus of text $D = \{d_1, ..., d_n\}$ and a vocabulary $W = \{w_1, ..., w_m\}$ it associates with each observation an unobserved variable $Z = \{z_1, ..., z_k\}$ and defines a joint probability model over $D \times W$. The mixture can be expressed in two forms, in this thesis the symmetric one is going to be used:

$$P(d, w) = \sum_{z \in Z} P(z)P(d|z)P(w|z) \tag{1.2}$$

To maximize the predictive power of the model an Expectation Maximization(EM) algorithm is used, this method comprises of two steps:

- an Expectation step which estimates posterior probabilities for the latent variables are computed
- a Maximization step where parameters are updated

1.2.1 Advantages

By comparing pLSA to LSA by representing 1.2 in matrix notation as $P = U\Sigma V$, given $U = (P(d_i|z_k))_{ik}$, $V = (P(w_j|z_k))_{jk}$ and $\Sigma = diag(P(z_k))_k$, it is possible to note the biggest difference. While LSA uses the Frobenius norm as objective function to optimize the decomposition of C, pLSA maximizes a likelihood function of multinomial sampling, this has the advantage of giving to P a clear probability distribution and it is possible to try and give a meaning to directions in this latent space. Moreover the dimensionality choice in a pLSA model is done making use of the statistical theory for model selection. A further advantage is the better performance in the context of polysemous words[7] which is one of the main drawbacks of LSA.

1.3 Latent Dirichlet Allocation

Latent Dirichlet Allocation is a further generative statistical model which generalizes pLSA and tries to solve its shortcomings. pLSA provides no probabilistic model for the documents, documents are represented by their mixing proportions for the topics which are strictly related to the training set [8]. This leads to a problem of linear growth of the size of the model, creating over-fitting problems, and no direct way to assign probability to a document outside the training set. To overcome said issues LDA uses a k-parameters hidden random variable as the topic mixture weights.

Chapter 2

The POIROT methodology

The aim of this chapter is to explain the POIROT methodology for interactive phenomena explanation as described by Frisoni and Moro in [9], to later use it.

2.1 Preprocessing

As in every work making use of data there is the need to preprocess it in order to make it usable, text data is particularly problematic since grammatical errors are the norm and linguistics variations make the same word change in meaning and spelling. Hence this first section will cover the techniques applied to make further analysis possible.

As already pointed out getting rid of grammatical errors is crucial since it will make correlations more evident and reduce dimensionality, moreover stopwords, extra withe spaces and punctuation can be removed. Lemmatization does standardize all forms of a word into one single lemma making correlation between similar words stronger. Furthermore tokenization is central to the construction of the vocabulary, for the purpose of this method a unigram approach will be used.

A discretionary step is the use of a Named Entity Recognition (NER) to classify terms into categories and to ease the feature selection.

The POIROT methodology works on labeled text, hence one last step is to classify the documents if no class is readily available. This can be done for instance by means of an opinion mining task, or by classifying the documents according to and underlying category.

2.2 Modeling

2.2.1 Term-Document Matrix

The different choices of language models have already been discussed in Chapter 1, however to use a language model on a corpus of text is necessary to generate a term-document matrix, such matrix is sparse and has as columns the documents and as rows the terms, each cell will then represent how many times that term appeared in that document. On this matrix will then be applied a transformation which makes each term more relevant the more it appears in the document but at the same time less relevant if said term has a overall high frequency in the corpus. This term-weighting transformation known as TF*IDF allows to determine the importance of each term in a document, given its frequency in the document (TF) and its relative collective frequency (IDF). Following suggestion given in [9] in this thesis will be used the following equation:

$$w_{t,d} = log(1 + tf_{t,d}) \times (1 - shannonEntropy(tdm)) \tag{2.1}$$

2.2.2 Feature Selection

In order to remove irrelevant terms it is possible to perform feature selection, this is done by keeping only the terms with a frequency above a certain threshold. This has to be done with caution in order not to weaken latent correlation relations and possibly a different threshold has to be used for entities since they might be relevant but underrepresented in the document.

2.2.3 Graphical Representation

Once a language model is applied the term-document matrix is mapped into a latent vector space whose first few dimensions capture most of the variability. For this reason a two dimensional representation of the data will still be representative of their distribution in the latent space. However, one might still want to use t-SNE [10], a technique used to represent high-dimensional data in a two dimensional space, this is however optional as it is possible to see from the rapidly decreasing values of the eigenvalues in Σ_k.

2.2.4 Identification

By means of the graphical representation it is possible to proceed in the last step leading to phenomena explanation: the identification of the areas in which the analysis has to be conducted. Such portions of the space are characterized by a higher concentration of documents of the class of interest. This is done visually by assigning to each class a different color, so that clusters of documents will be easier to spot.

2.3 Phenomena Explenation

A key feature of the LM seen in Chapter 1 is that they allow the transposition of new vectors into the latent semantic space, by leveraging this it is possible to iteratively build up a query containing all the terms which stand out in the area previously identified, such query can then be interpreted to explain the phenomenon of interest. The first term is selected among those central to the identified area and with the highest norm, it will be the starting point of the process. A query q can be considered as a document in C and hence, after having undergone preprocessing, can be transformed as a row of the matrix V_k as: $q_k = q^t U_k \Sigma_k^T$. Once q lies in the latent space it is possible to check whether there is a correlation between it and the class c by means of a chi-squared test used in combination with R-precision. In order to pick the next term to be added is necessary to define a similarity measure. This is done by calculating the cosine of their scalar products $C_k C^T = U_k \Sigma_k^2 U_k^t = (U_k \Sigma_k)(U_k \Sigma_k)^t$ or $V_k V^T = U_k \Sigma_k^2 V_k^t = (V_k \Sigma_k)(V_k \Sigma_k)^t$. It is worth noticing that while for LSA the similarity is expressed in the range [-1,1], for pLSA and LDA is in the range [0,1] allowing a probabilistic interpretation. This way we can choose the next term as the one with high similarity and norm, at this point the process just needs to be iterated till the next term will show no significant correlation with the class c. The final product is going to be a vector of terms all highly correlated with the class of interest, by interpreting it is possible to try and find an explanation for the phenomenon.

2.4 Evaluation

The most straightforward way to evaluate whether the resulting explanation makes sense is to compare it with preexisting reports and expert opinions. In [9] a more formal way based on gold standards is described, this requires experts creating a set containing terms with positive and negative correlations on the topic of interest. Such correlation are going to be evaluated in the same way as the correlations between queries and classes were, in this way a confusion matrix can be created. This method allows the evaluation of the framework in which the explanation is constructed and to indirectly asses the correctness of the explanation.

Chapter 3

Case Study and Experiments

3.0.1 Dataset

Yelp is an American company with headquarters in San Francisco, it runs the website Yelp.com where millions of reviews for the business are hosted. In the past years it held the Yelp Dataset Challenge for which it released a dataset containing a subset of its reviews available for academic purposes, in this thesis a sample of that dataset containing restaurants' reviews and their stars rating is going to be used. It was chosen because it has a lot textual data in combination with a label given by the rating, moreover there is a phenomenon to be understood, the motivations behind the satisfaction or dissatisfaction of a customer, this reasons make it a perfect candidate for the use of the POIROT methodology. The code was run on a Google Colab notebook with 12GB of RAM, this imposes a stringent constrain on the size of the dataset which can be used without crashing the session, hence a much smaller sample is created. The files of interest are "business.json" containing information about the business, most important it allows to select only restaurants due to its "categories" data and "review.json" containing the actual reviews and the stars.

3.0.2 Data Loading

To read .json files it necessary to make use of the "jsonlite" package, while in order to select only reviews of restaurants I resort to an SQL approach using the package "sqldf". The fact is that the categories are specified as tags like the following:

```
## American;Nightlife;Bars;Sandwiches;Burgers;Restaurants
## Italian;Restaurants
## Vegetarian, Vegan, Restaurants, Indian
## Hair Stylists;Hair Salons;Men's Hair Salons;
## Dentists;General Dentistry;Health&Medical;Oral Surgeons
```

This means that in order to effectively subset restaurants' reviews "LIKE" is very handy. This is also the moment when the dataset gets sampled and split into smaller random chunks, for this project a sample of 250,000 reviews has been used. It is worth noting that while reading the .json file a custom handler function was used in order to extract only the relevant information.

```
reviews <- NULL
stream_in(
  file("/content/drive/MyDrive/TESI/yelp_academic_dataset_
      review.json"),
  pagesize=100000,
  handler=function(x) {
    reviews <<- rbind(dr, x[,c("stars","text", "business_id")])
  }
)

business <- NULL
stream_in(
  file("/content/drive/MyDrive/TESI/yelp_academic_dataset_
      business.json"),
  pagesize=100000,
  handler=function(x) {
    business <<- rbind(dri, x[,c("categories" ,"business_id")])
  }
)

final <- merge(reviews, business, by = "business_id")


n_spl <- 10
cat("Make", n_spl, "splits\n")
splits <- sample(1:n_spl, nrow(final), replace=TRUE)

for(i in 1:n_spl) {
  cat("Saving split", i, "splits\n")
  write.csv(final[splits==i,],
      file=gzfile(paste0("bus_rev.",i,".csv.gz")), row.names = FALSE)
}
```

```
bus <- read.csv("/content/drive/MyDrive/TESI/bus_rev.1.csv.gz")
Restaurants_only <- sqldf("select * from bus where categories LIKE
   '%restaurant%'")
Restaurants_only1 <- Restaurants_only[, 2:3]
write.csv(Restaurants_only, "restsurants_only_1.csv.gz", row.names =
   FALSE)
```

3.0.3 Preprocessing

This section makes use o the package "tm" for the creation of a corpus and to carry out preprocessing operations such as stemming, removing white spaces, punctuation and stop-words, as well as the creation of a term-document matrix and performing feature selection. Here "txt" is a character vector containing the text of the reviews. Performing tf–idf weighting is a bottleneck for the entire process, if the sample is too big this operation will consume all the available memory in a few seconds and crash the session.

```
char_vec <- txt
corpus <- VCorpus(VectorSource(char_vec), list(language = "eng"))
corpus <- tm_map(corpus, content_transformer(tolower))
corpus <- tm_map(corpus, removePunctuation)
corpus <- tm_map(corpus, removeWords, stopwords("english"))
corpus <- tm_map(corpus, stemDocument)
tdm <- TermDocumentMatrix(corpus)
tdm <- removeSparseTerms( tdm, 0.99 )
words <- rownames(tdm)
tdm <- as.matrix(tdm)
tdmle <- lw_logtf(tdm) * ( 1-entropy(tdm) )
```

3.0.4 Phenomena Explanation and Code

After having applied the LSA model to the weighted tdm it is possible plot documents and terms into the latent space, doing this allows for a visual identification of the first term. In this case as described in [9] it is recommended to print them on the second and third dimension since the data will be less concentrated an easier to tell apart. As it is possible to see from 3.1 the term "speak" is a good choice since it appears to have high norm, because it is far away from the origin, and at the center of a cluster of one star reviews.

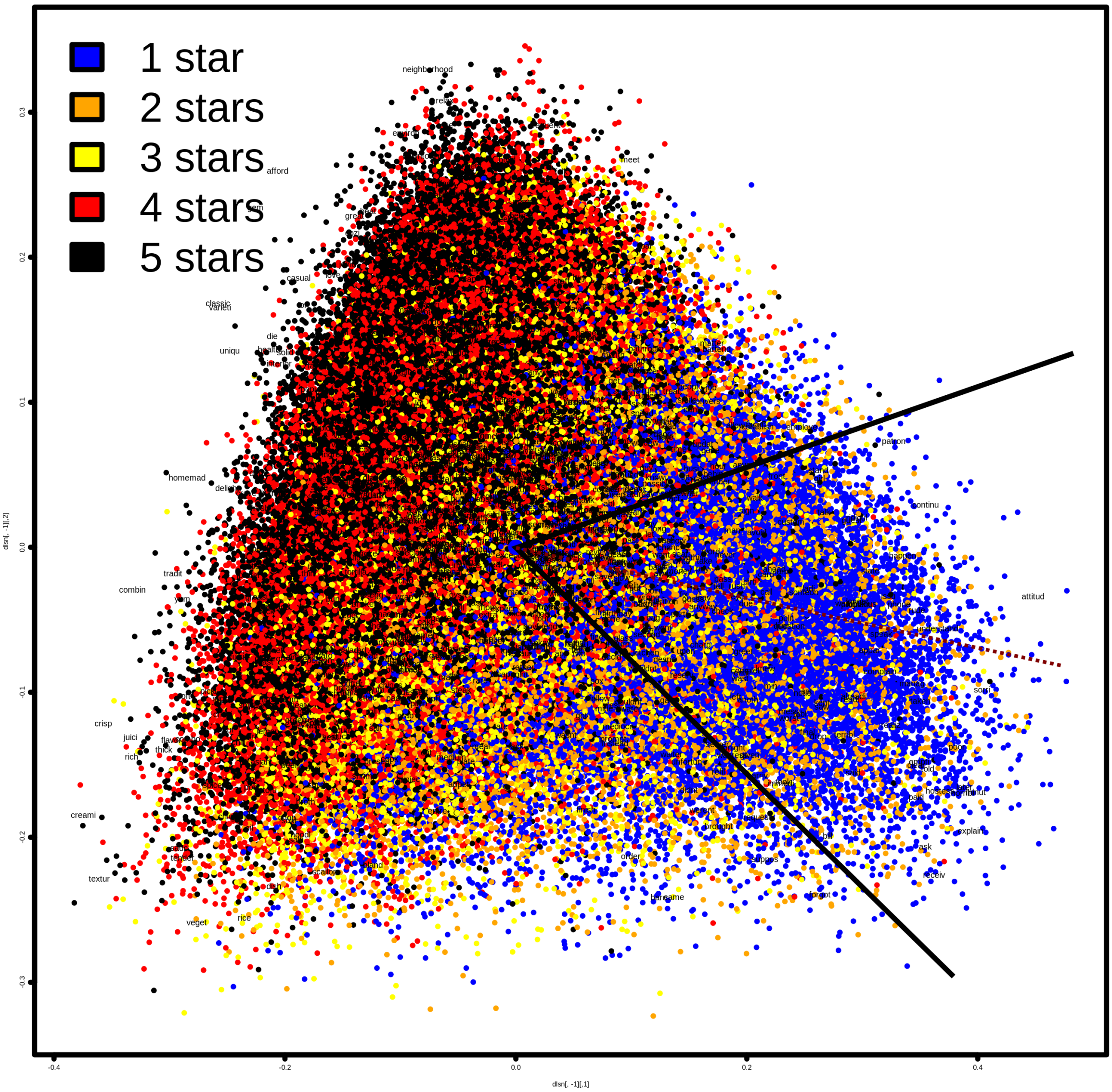

Figura 3.1: Graphical representation of terms and documents in the second and third dimension.

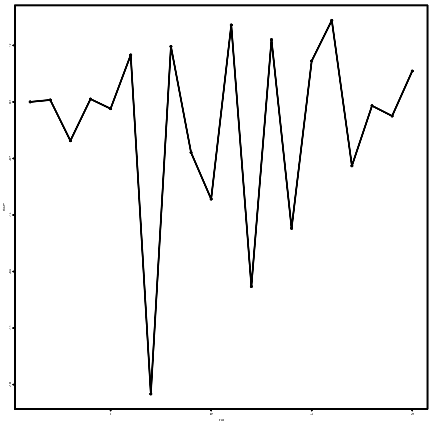

Figura 3.2: Graphical representation the knee points used for dimensionality reduction

Before proceeding further it is important to make the right dimensionality choice, therefore six dimensions will be picked accordingly to 3.2 Now that the first term is chosen it has to be transformed into a query in the latent space and then a chi-square test has to be applied to it to see whether it has a correlation with the one star reviews.

```
q1 <- makequery("speak", tdm, lsar)
q1x <- table(1:nrow(dls) %in% top(cosines(dls[,1:6], q1$ls[1:6]),
   28321), stars == 1 )
chisq.test( q1x, correct=FALSE )
```

Pearson's Chi-squared test
data: q1x, X-squared = 60127, df = 1, p-value < 2.2e-16

Such a small p-value suggest the presence of a correlation between the term "speak" and one-star reviews. It is now possible to visualize which documents fall near the query.

```
## Literally the worst customer service. The manager is beyond rude.
## Food was good. Customer Service is terrible. Cashier was very rude.
## I had a younger girl taking my order. She was somewhat rude and racists!
## My mind was blown with the rudeness this resturaunt to a customer.
## Horrible restaurant , called the place and the guy was rude on the phone
## I spent more time in the restroom due to food poisoning than I did eating.
## I would give them 0 stars if i could. the customer service is terrible.
## This girl Lexie answered tonight and was completely rude.
## Poor attitude of the waitress towards the customers.
```

```
head(sort( cosines( tksrs[,1:6], q1$dksrs[1:6] ) , decreasing =
    TRUE), 20)
sort(tnorms4[cosines(tksrs[,1:6], q1$dksrs[1:6])> 0.89], decreasing =
    TRUE)
```

The documents seem to be indeed all related to people speaking which is a good thing, using the code in 3.0.4 it is possible to identify the next word with high norm and similarity. By reiterating this process six words able to explain the phenomena have been found, which are "speak custom rude phone call worst".
The old saying "customer is allways right", seems to still hold a lot of wisdom today, indeed unsatisfied customers do write bad reviews complaining about how they received a rude treatment when speaking to the customer service, especially when trying calling by phone. It is interesting to see how most of the reviews that where related to the queries didn't even mention how the food was, they just focused on how the personal interaction with the staff went.

Conclusion

This thesis shows an application of the POIROT methodology, which allows to analyze textual content and extract information about possible causes for a phenomena of interest. Making use of it was possible to analyze a dataset of restaurants' reviews and to propose a possible explanation behind the negative ones, it seems that one star reviews are written by clients unhappy with the customer service, both in person and on the phone. This result should however be taken with some reservation since it wasn't possible to conduct an evaluation based on expert opinion or on gold standards. Being able to extract this kind of information from text with a relatively simple technique can prove very useful as shown from this application.

Bibliography

[1] Thomas K. Landauer and Susan T. Dumais. A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240, 1997.

[2] J. R. Firth. The technique of semantics. *Transactions of the Philological Society*, 34(1):36–73, 1935.

[3] Peter M. Wiemer-Hastings. How latent is latent semantic analysis? In Thomas Dean, editor, *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, IJCAI 99, Stockholm, Sweden, July 31 - August 6, 1999. 2 Volumes, 1450 pages*, pages 932–941. Morgan Kaufmann, 1999.

[4] T.K. Landauer, P.W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse processes*, 25:259–284, 1998.

[5] Giacomo Frisoni., Gianluca Moro., and Antonella Carbonaro. Learning interpretable and statistically significant knowledge from unlabeled corpora of social text messages: A novel methodology of descriptive text mining. In *Proceedings of the 9th International Conference on Data Science, Technology and Applications - DATA,*, pages 121–132. INSTICC, SciTePress, 2020.

[6] Thomas Hofmann and Jan Puzicha. Unsupervised Learning from Dyadic Data. Technical Report TR-98-042, International Computer Science Insitute, Berkeley, CA, 1998.

[7] Thomas Hofmann. Probabilistic latent semantic analysis. In Kathryn B. Laskey and Henri Prade, editors, *UAI '99: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden, July 30 - August 1, 1999*, pages 289–296. Morgan Kaufmann, 1999.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[9] Giacomo Frisoni and Gianluca Moro. Phenomena explanation from text: Unsupervised learning of interpretable and statistically significant knowledge. In Slimane Hammoudi, Christoph Quix, and Jorge Bernardino, editors, *Data Management Technologies and Applications - 9th International Conference, DATA 2020, Virtual Event, July 7-9, 2020, Revised Selected Papers*, volume 1446 of *Communications in Computer and Information Science*, pages 293–318. Springer, 2020.

[10] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.