POLITÉCNICA

UNIVERSIDAD
POLITÉCNICA
DE MADRID

“The Sanitary Evolution of London”

Natural Language Processing

Master of Science in Health & Medical Data Analytics (EIT Health)

Carlos Alcoba Núñez

Intelligent System

2021-2022

Introduction

One of the branches of the Artificial Intelligence in these days is the use of techniques to perform Natural Language Processing (NLP). The aim of these works is to use the advantages that computer technologies offer to address problems or situations where the data is presented in text formats. These techniques will allow to extract the relevant information from these documents and perform different task on it. Some of the task that can be done are classification of the text (such as differentiated between good and bad reviews for restaurants), assess the difference between two texts (such as understand the difference between two texts of different periods about the same topic), perform keyness analysis in a text (such as identifying the key concepts behind a speech) or create automatedly different texts that will adapt to different situations (such as an artificial intelligent chatbot that is able to react in real time to the situation).

Problem to solve

In this case the task will consist of performing an analysis on a book titled "The Sanitary Evolution of London" from Henry Lorenzo Jephson. This book provides a description of the London sanitary evolution between 1855 and 1906. Where each chapter is devoted to a different period of 1855-1906 years in succession. The aim of the work is to apply different NLP techniques to extract the information related to the health situation from each period and make comparison among them.

This book can be found online in: https://www.gutenberg.org/cache/epub/47308/pg47308.txt.

Methodology

As a general guide this task has been structured in the following way. After the identification of the text of interested it will be downloaded. A first pre-processing will be applied to the text to get it ready for the analysis. Secondly, the different analysis techniques will be applied to obtain the results seek. Finally, the results obtained will be discussed and future lines will be presented.

Loading and pre-processing of the text

The following process will be done with R programming languages and different libraries.

The first step will be loading and read the book from the internet repository. Once the complete text is loaded, the lines where the book starts, and ends will be identified, and the rest will be discarded. Next it will be checked that all characters belong to UTF8, in case not the necessary actions will be carried out to ensure it. Finally, all lines will be collapse into a single line.

Secondly the text will be divided in the different chapters. For this each the complete line from the end of the first part will be divided in the different chapters. After this process is done the different characters that are unnecessary will be removed (such as the blank line character, or the quotes character). Finally, the last part of the chapter will be removed as in this book it corresponds to Footnotes. This will result in a list of objects each one containing a chapter.

Analysis of the text

Sentences Size

The first analysis to be performed will be the size of the sentences from each chapter. For this a histogram from each of the chapter as well as a general one will be computed.

Chapter Distances

The second analysis will be to compute the distance between the chapters. This will be displayed by means of a dendrogram.

Most Frequent Words

This analysis will consist of identifying the most used word. Again, it will be done for the complete text and for each of the chapters separately. In this case stop words will be removed.

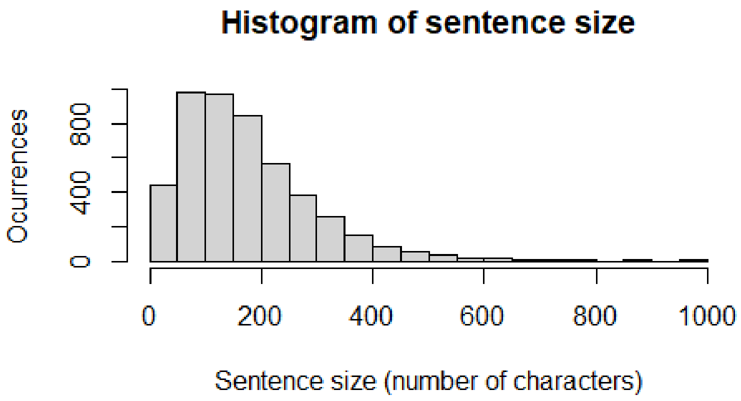
Keyword Analysis

The last part of the analysis will be to compute the Keyness from each chapter relative to another one. This will be done to try to identify the differences between each pair of chapters. This analysis will be done for each pairs of chapters.

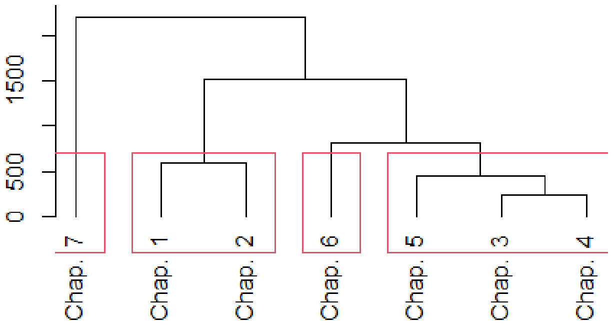
Analysis of the results

Sentence Size

There are no clear differences between the different histogram for each of the chapter and the global histogram. Some of the chapters have more prevalence of sentences around 200 characters while others are around 150. This difference could imply that some of the chapters are more descriptive as they use longer sentences while the others don't. Also, all of the chapter have much longer sentences that go up to 800 character, but in all cases they are residual. This could imply that these sentences have a higher description content that need more words to be clear.



Chapter Distances

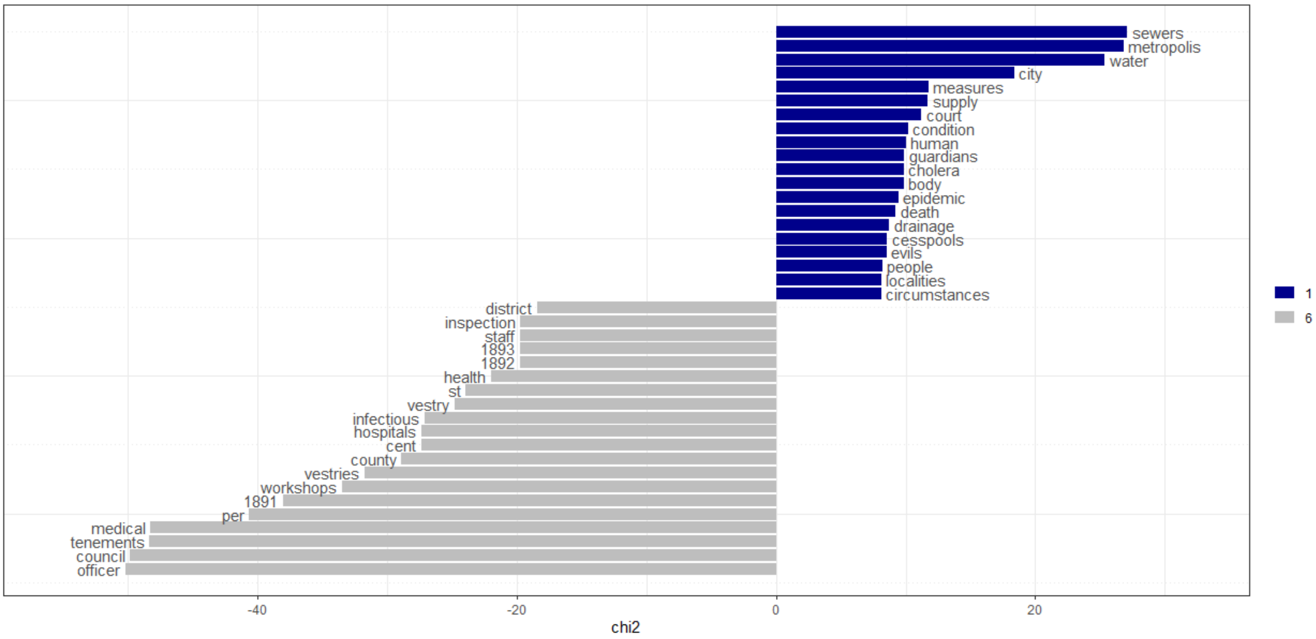


In this dendrogram it can be appreciated that the most similar chapters are the 3 and 4 followed by the 5. This could indicate that the situation during the period that this chapter refers to are very similar and little changes occurred. They are also similar to the chapter 6, but in this case, there is more distance between them, which could indicate that an event could have occurred that has motivated a change in the situation. Chapter 1 and chapter 2 are also similar, this is surprising as the first chapter is an introduction and does not talk about a specific period. Which could imply that the data referring to the first period, presented in Chapter 2, is not very specific motivating that a global view like the one that could be provided in Chapter 1 could be similar. Lastly the last Chapter is the one with the bigger distance between the others. There are two possibilities for this, the first is that at that time there are more data about the sanitary situation and more concrete data is provided. Or it could be that in this chapter a summary and general conclusion about the book is drawn.

Most Frequent Word

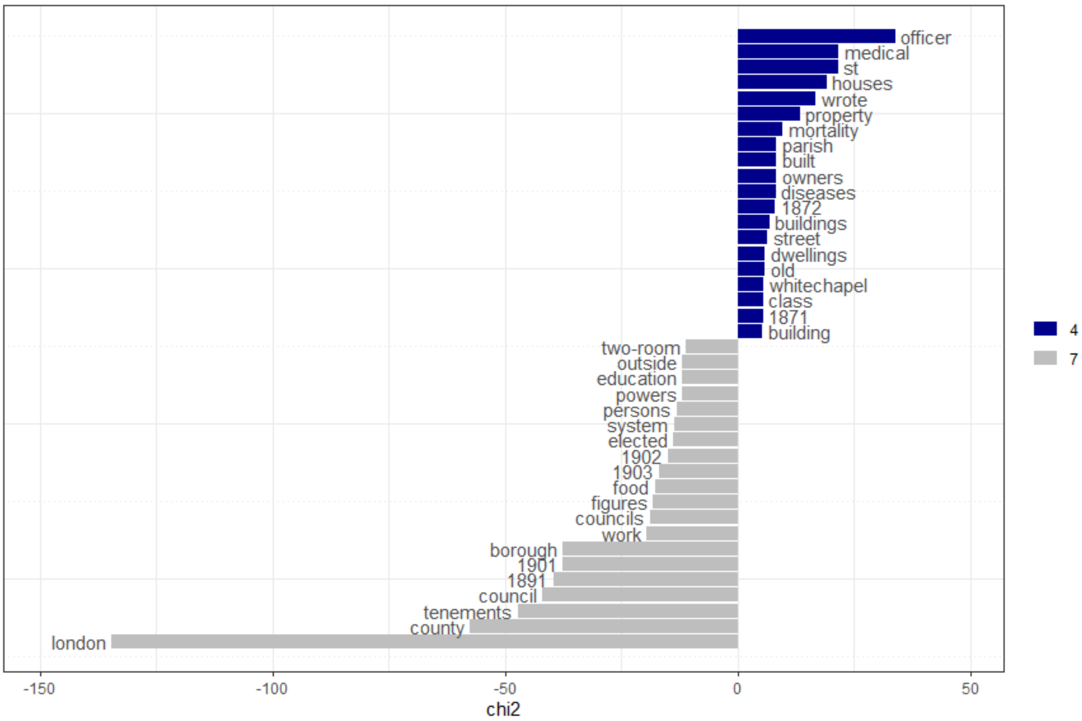
In a global view it can be found that the top 5 word used in this book are: health, London, houses, sanitary and medical. This is expected as it is a book about the London sanitary evolution. Looking at the top 5 for each of the chapter, it can be appreciated a similar distribution but with some changes. In chapter one appears people, which could mean that this chapter is more focused on how the sanitary situation affect the people. In the second chapter appears officer, and that could indicate that officers play and important role in this period. For chapter 3, 4 and 6there are no changes respect to the global. In the Chapter 5 and 7 appears board, which could be understood as a govern that has made some decision to cause an effect in the sanitary situation, but in the case of the Chapter 5 the rest of the global words are also present. This information can be compared with the dendrogram, and it can be found that chapters 3, 4 5 and 6 are the more similar which agree with their most frequent words being the more similar while 7 is the less similar and it has the bigger differences

Keyness Analysis



Two different Keyness analysis have been selected to be included in this report. The first one represetn the difference between the 1 and 6 Chapter. It can be appreciated that the keyword related to the first chapter are more general and more related to health field, like cholera, epidemic or measures. While in the case of the 6 chapter the keywords are more related to buildings and localities, like dsitric, hospitals or tenements. This could mean that in the first chapter there is a more gloval view focusing on what diseases has affected London and how it has spread and have been fought (sewers, measures, supply…). While in the 6 chapter the focus could have been more on the implications of the sanitary conditions that affected this period and the period itself (inspection, each year, medical, tenements or council). It could be that this chapter also mention different measures taken by the council or other government levels.

In the second case the comparsion has been made between chapter 4 and 7. In this case it can be appreciated that in the chapter 4 the focus was set to in building and spread of the diseases (medical, houses, mortality, property, street..) while in the chapter 7 the focus was more in different personalities (persons, selected, figures, councils…). An explanaition for this different could be that in the period related to the chapter 4 there where less measures taken in order to stop the advance of a epidemic, as for that time it is possible that there where no eniught information about it and they where just describing the situation. This would lead to period, presented in chapter 7, where they had the information and the focus was set on how they could mitigate the effects, and this is why institution like the coundil or selected person has more importance as they where the ones in charge of the measures.



Conclusion

The analysis of this text has allowed to reveal how the data is presented and which period where more similar. Although the initial objective that was to identify the prevalent symptoms and disease in each period, has not been completed, the work has allowed a better compression about the period that the book talk about. With this analysis it has been possible to identify the structure of the book. Where the first chapter is more of an introduction and the following chapter will present how was the situation for each period. Once there is enough data, different measures had been taking to address it. And a final chapter that has the biggest difference where the keyword analysis shows that it is possible that it focusses more on the measures and how took those measures in order to improve the sanitary condition of London.

Future work

The first objective proposed for this work has not been achieve, but a future line for achieving it could be the use of a specific medical corpus. Comparing this corpus with the corpus of each chapter could show which where the symptoms and disease more prevalent and their evolution. Also, with this information it can be stated if the measures taken where effective to eradicate or decrease each of the diseases or symptoms.

Code availability and more graphical material

A repository with this project can be found at: https://github.com/calcoba/NLPLondonSanitary. This repository also holds more graphics not selected for the report.