DE SIMONI Pietro
SESSI Michela

Stockholm University

Report of Laboratory Exercise 1 :

Preprocessing and Utilizing Web Data

Group 15

DE SIMONI Pietro
SESSI Michela

Stockholm University

Introduction

Now, in the 21st century, data are used as engine for empowering business in many ways. More common way to analyze data is to build a dataset, or looking for an existing one. By the way, creating a dataset (and sometimes also consulting one) can be very expensive. Nevertheless, Internet provides a big amount of information that are not structured and it needs methods and tools to manage this data.

The aim of this Laboratory Exercise is to practice and train with the pre-processing, and building a word space model.
A clarification needs to be done: this is just an exercise and the performance of the model are smaller than expected. The training of this models is usually done on hundreds of millions of words. Instead, in this case the size of the data is obviously reduced.

Content & Discussion

The lab exercise is a step by step process, aimed to highlight the usefulness of each process. In fact sometimes we were asked to skip some crucial steps in order to notice that something more needs to be done.

The first things that we discover is the difference between what we are used to see as a web site and the downloaded data: the text is represented in an HTML format. It means that is not so immediate to catch the part of the document containing information. In order to manage the real text, we need to apply some transformations highlighting useful data. The usage of Perl's text-extractor seems to remove a relevant part of the text too. We shouldn't allow our data cleaning process to let us lose relevant information.

For analyzing text a first approach is to get a list of the most frequent words in the document, so that we know which are the main topics. By the way, the result is a list of the main stopwords (very common but not characterizing words, such as articles, prepositions and conjunctions), with just some common words related to the topic of our group of documents (for example language, speech, information). Lemmatizing words before this passage is a good method to get a better result: for example "language" and "languages" are now considered as the same lemma.

We train and test a model with our text in order to obtain the most important words and, for each, the ten closest matches. As we noticed before, if we don't pre-process the text in advance, the result is affected by a list of stopwords. In fact sometimes stopwords happen to be close to some other words, just due their high frequency.
We train and test the model again, but this time we remove the stopwords. We could notice the difference showing the result for the word *"matrix"*: in the first model we obtained relationships with the words - *call, by, to, have, give, if, that, thus, system, function.* This list is full of non relevant words. In the other hand, the list with the new model removes the unrelevant word, showing new interesting relationships as the ones with the word *"invertible"*, *"determinant" or "approximation".* The same reasoning could be done to the other words in the list of the results.

Conclusion

Throughout the exercise we learnt how web crawlers can understand what a web page is about. This is a good way for search engine to index pages and to give better results to the users.
We have used a little collection of documents. To get a better list of closest matches we should analyzed a large number of texts. Also the stopword list was created by hand, but there are reliable lists of stopwords available that could lead us to better results.
This is a useful example to learn in a simplified way techniques and understand the powerful and the difference from using pre-process methods or not.