Searching the space of word temporal profiles on Le Temps Newspaper

Due on Monday, May 18, 2015

Sidney Bovet Valentin Rutz Zhivka Gucevska Mathieu Monney
Florian Junker John Gaspoz Joanna Salathé Ana Manasovska
Fabien Jolidon

May 14, 2015

Contents

1 Motivation

Studies based on the visualization and analyses of temporal profile of word using a so-called “n-gram” approach have been popular in the last years. However, most of the studies so far discuss the case of remarkable words, most of the time found thanks to the researcher’s intuitions for finding “Interesting” curves using the n-gram viewer.

The motivation of this project is to invert the problem and to propose a simple tool for researchers to automatically explore the space of temporal curves in search for words. As researchers often analyse how words appear and relate to historical events, our goal was to provide researchers a way to extract words having the same temporal profile. These results would help the researchers to analyze the use of a given word and the events that they relate to.

In order to classify the temporal profiles various metric were implemented providing the user different ways to retrieve interesting information.

In section 2, we will explain how the data generation was made. From parsing the OCR’d articles to a cleaned and formatted data. In Section 3 we detail how the application is structured and its general pipeline. Section 4 describes in details the metrics implemented as well as their strengths and weaknesses. Section 5 and 6 describe how we proceeded to optimize the metric’s parameter and how the metrics perform against some test cases. Finally, section 7 provides hints about future works that could fit this project.

Our analysis is based on Le Temps Newspaper Corpus which is a database of 4 million articles covering a period of 200 years and composed of digitized facsimiles of "Le Journal de Genève" and "La Gazette de Lausanne" wherein each article has been OCR’d.

2 Data Generation

Data generation is a critical component as messy data will impact the accuracy of the metrics and the pertinence of its outputs. Hence, great care was taken in the data generation phase in order to make the data as useful as possible for the application. With the use of various statistics we were able to filter many garbage data.

2.1 Parsing

2.2 1gram Generation

To ease the development of the metrics we defined that for each word we would have a list of 159 Integer entries containing the number of occurrences for this word for each year in the range of our database (from year 1840 to 1998).

In order to compute this, two map reduce functions were created. The parsing phase had given a list of files, one file for a year, containing every word used in that year. The first mapreduce would process each of these files and output, in a file, each word associated with the year and occurrences for this year : [year_word occurrences].

The second mapreduce would take, as input, the output of the first one and generate the final format [word occurence_year1, occurence_year2, ...]. If not occurence of a word was found for a year a 0 value would be assigned to this specific year.

2.3 1gram Cleaning

3 Architecture

4 Metrics

4.0.1 Curve Comparison

4.0.2 Peak Detection

4.0.3 Dynamic Time Warping

5 Data Tuning

6 Experimental results

6.1 Metric Benchmark

7 Conclusion and Future work

As observed in the previous section, with the use of this application we were able to retrieve interesting information. For example, by testing the word "guerre" we detected various semantically similar words as output. And by looking at their temporal profiles we can analyse when these words were mostly used and observe that their pic of occurrences would match the historic event of the first and second war.

This a simple test-case but it shows that, giving a word, the user can retrieve useful information. Of course not all words would give such interesting results and, as we saw previously, not every metric performs well for any words.