Project plan: Searching the space of word temporal profiles on Le Temps Newspaper

Due on Tuesday, March 10, 2015

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

Contents

Abstract

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 purpose of this project is to investigate how we could invert the problem and automatically explore the space of temporal curves in search for words. For instance, we would be interested in asking the system to retrieve all the curves “similar” to a given one. This implies defining a way of describing temporal profiles and classifying them according to a predefined distance.

Le Temps Newspaper Corpus 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.

1 About the project

1.1 Project name

"Searching the space of word temporal profiles on Le Temps Newspaper Corpus", a project proposed by the Digital Humanities laboratory.

1.2 Team members

1. Sidney Bovet (leader)
2. Valentin Rutz (leader)
3. Zhivka Gucevska
4. Mathieu Monney
5. Florian Junker
6. John Gaspoz
7. Joanna Salathé
8. Ana Manasovska
9. Fabien Jolidon

2 Goals

1. Given a word, give the words that have the similar temporal profile/curve
2. Potential add-on: given a curve, give list of words having the same temporal profile
3. Potential add-on: intersect with other database (countries, artists ...) in order to define ontologies (clustering of words)
4. Potential add-on: define general theme from words with similar curve
5. Potential add-on: merge all the metrics

3 Methodology

3.1 Task 1: Creating the 1-grams (2 weeks)

Once we get the data:

1. Parse the XML file, extract the words to obtain raw text
2. Then, we can just do a word count, and store the output as CSV
3. Compute the word temporal profiles

3.2 Task 2: Clustering of the word temporal profiles (5 weeks)

To achieve this, we'll try some of the following techniques:

1. Fourier transform
2. Machine Learning/Artificial Intelligence
3. Time series
4. Smooth the curves (at least for visualization, since it can be tricky for periodic events)
5. Ways to compare curves
   (a) exactly the same curve
   (b) same pattern (not same year)/ different frequencies
   (c) same year difference amplitude
   (d) mean

3.3 Task 3: Create an interface to see the curves/output (2 weeks)

4 Risks for the success of the project

4.1 Risks

- Lack of mathematical background = don't think of a better solution compared to the "easy"/brute force ones
- Wrong intuition
- Wrong/"not good" result and don't notice it
- Difficulty of evaluating our overall result (can compare two curves visually but for a lot of them . . .)
- Getting stuck for lacking of ideas (don't think of other fitting curves solutions)

4.2 Avoided risks

- Data is provided by DHlab
- The data is not dirty
- Our goals are computable (in some ways if we take the good track)

5 Milestones and Deliverables

Duration of the project: 9 weeks (without Easter break)

1. now-10.03.2015 (Week 1)
   Grab information/research the web
2. 10.03-23.03.2015 (Week 2-4)
   Parsing and computing 1-gram.
3. 31.03-04.05.2015 (Week 4 to 10) (with sub-milestones)

- finding and testing ways to compare the curves
- performance testing
- add-ons

4. 04.05-12.05.2015 (Week 11) User interface (web) and writing the final report.
5. 13.05-19.05.2015 (Week 12) Preparing the presentation.

6 Work packages and assignation to team

6.1 First milestone

- *Parsing Data* (Zhivka, Flo) [parsing the data, then continue with research for the 2nd part]
- *MapReduce*: make the 1-Gram and put them in CSV: Word, #occ, #year (Fabien, John)
- *Begin research of 2nd part* (Mathieu [Spark testing, Code style] Ana [think of machine learning techniques we could use based on the PCML class she took], Valentin [TBA])
- *No research solutions*(Sidney [mean idea], Joanna [exactly same curve])

6.2 Second milestone

[TBA]

6.3 Third milestone

[TBA]