Event Stream Detection

BigData2015 - CS422
Digital Humanities Laboratory

Laurent ANADON
Antoine BASTIEN
Antoine BODIN
Matias CERCHIERINI
Nina DESNICA
Louis FAUCON
Damien HILLOULIN
Christian MOUCHET
Sami PERRIN

Professor: Christoph KOCH
Teaching Assistant Immanuel TRUMMER
DHLab Supervisor Yannick ROCHAT

Spring 2015

Contents

1 Introduction

Along with the 2015 Big Data course, we are leading a project adressing topic detection in news streams for the DHLab[1] as part of a course project.

We aim at detecting articles that talk about the same topic over a set of issues contiguous in time, and across two newspapers over 200 years:

- *Journal de Genève* (JDG) from 1826 to 1998,
- *Gazette de Lausanne* (GDL, under different names) from 1798 to 1998.

To do this, we are looking into clustering, hierarchical clustering and correlations detection techniques. One of the main challenges here is the huge amount of data: we are considering articles over a huge time span, which is why we need the algorithms we implement to be scalable.

In order to achieve this goal, we focus upon previous studies such as [1] and try to use this in the context of the big data and the requirements of Spark. First of all, we parse articles and store them inside convenient distributed data structures through Spark RDDs, and secondly we extract releveant themes among articles over a well choosen time period with parallelized machine learning algorithms. Once this is done, we find correlations between these themes to build the evolutionary graph. Eventually, we analyze the life cycles of themes and measure their strength over time.

[1] Digital Humanities Laboratory

References

[1] Qiaozhu Mei and ChengXiang Zhai. Discovering evolutionary theme patterns from text: An exploration of temporal text mining. In *Proc. of KDD'05*, pages 198–207, 2005.