Thesis proposal

Lynette Joosten

February 22, 2017

Abstract

In this document I describe my thesis proposal for the master thesis project of the master Information Studies: track Data Science. My thesis will be written during an internship at CBS, the Central Bureau of Statistics.

1 Personal details

My email `mailto:lynette.joosten@student.uva.nl`

My supervisors email `mailto:e.kanoulas@uva.nl`

The wiki on my github account `https://github.com/LynetteJoosten/Thesis/wiki`

2 Research question

At CBS there is a big dataset containing the scanner data of supermarkets. Scanner data of supermarkets are large files containing lots of data regarding the sales and turnover of various products. As such, these files provide valuable information on local economic phenomena. By comparing data from different shops, I will be studying potential new indicators indicative of changes in economy or related to producer and/or consumer confidence.

3 Related Literature

4 Methodology

4.1 Resources

This research will be focused on a data set provided by the Albert Heijn supermarket chain. Weekly the CBS gets a new data set from Albert Heijn consisting of all products, their price and the amount of products sold. This data set consists normally of .. rows and .. columns. It is already processed, so I will not have to do any preprocessing or cleaning of the data set myself. Python and R will be used to test the different approaches mentioned in the section about the research question. Scala or PySpark will be used if the data needs to be processed on a cluster.

4.2 Methods

It is not yet clear which specific methods will be used, since figuring out a new method is basically the goal of this thesis. At the moment, random sampling is considered as one method, but the other methods are not yet clear.

4.3 Evaluation

The results will be evaluated in two ways:

- Can the new consumer price index show the same differences over time as the old consumer price index?
- Is the absolute number of the new consumer price index the same or relatively close to the old consumer price index?

5 Risk assessment

6 Project plan

Week	Deliverables
1	..
2	..
3	..
4	..
5	..
6	Mid-term results
7	..
8	..
9	..
10	..
11	Hand in thesis
12	Defense