Data Representation,Reduction and Analysis

Juan Jose Soriano Escobar

Redona Brahimetaj

Master in Applied Computer Science and Engineering
Vrije Universiteit Brussels
Brussels Belgium
January 15, 2017

List of Figures

A Introduction

B Cleaning of Tweets

One of the most crucial parts to start with and that has a big influence in the result of our project, is data processing step. A csv file containing 2000 tweets was provided.It was required to preprocess the data to make them good enough for the 'learning' step.

B.1 Pre-Processing

We cleaned the data by applying several filters to them. We would like to mention that for this part, we have adapted the code that we already did before on Distributed Computing and Storage Architecture project. We removed the stop-words like determiners, the coordinating conjunctions and prepositions. Another preprocessing step that we did was trying to keep only the root of the words by applying stemming. In this way the clustering would be correctly implemented since it would be easier to cluster and find similarities beetween words that are the same.

Codescreenshot needs to be put..

C Noise Removal

D Clustering

E Visualization

F Conclusion