TAL - Politic review

Ismail Erradi, Björn Goriatcheff, Enzo Hamelin

06 May 2017

1 Introduction

In this project, analyzing language syntax and semantic is used in order to review, rate and classify ideas extracted from the program of every presidential candidates. On 5 different topics, we analyzed what was the opinion of each candidate and how those ideas were correlate.

2 Goals

- Extracting/parsing opinion of candidates on 5 different subjects.
- Weighting them using appropriate metric.
- Emphasizing the emergence of a group.
- Implementing an ask-answer system.

3 Functions

3.1 Data preparation

Dictionary The listing of every candidates have been implemented in a dictionary to create aliases. The listing of subjects have also been implemented in a dictionary containing the synonyms and other words associated to the subject. Both dictionaries are included in the file "settings.py"

Programs Every presidential program have been previously downloaded and imported into the folder "projects".

3.2 Parsing

In order to parse the programs, we used textract library to parse pdf encoded files out to text. Additionally, this text was tokenized using nltk library.

3.3 Weighting

Weighting correctly generated tokens was the most difficult part of our project. In order to do so, we semantically analyze a sentence of tokens. If this sentence contained a keyword of our dictionary then the proposition was weighted according to the grammar (affirmative sentence or negative sentence). Every similar proposition is weighted the same way and counted as a positive or negative proposition.

3.4 Classify

To classify the position of a candidate on a precise subject we count how many times this subject was recall in the program and how was the recall (positive or negative) compared to the original opinion. A simple calculus is executed to know if we can trust the opinion using the following formula:

$$trust = \frac{positive}{positive + negative + neutral}$$

4 Results

5 To Achieve