1 Introduction

Dataset: more than 3‘000 tweets of the 11 French presidential election candidates. We have worked in Python and used package: Sklearn, Numpy, Matplotlib.

2 Preprocessing

We have choose to keep only relevant words, a lot of very used words do not bring any information (e.g. "le"). We have also distinguish hashtags and other words as they do not have the same meaning in a tweet. Because semantic of words is really context dependant and difficult to process we have choose to not consider it.

Word	Frequency
contre	1.25%
gouvernement	0.49%
travail	0.45%
paris	0.40%
droite	0.38%
solidarité	0.38%

Hashtag	Frequency
#npa	11.03%
#loitravail	2.98%
#grèce	1.73%
#migrants	1.66%
#poutou2017	1.45%
#hollande	1.18%

Figure 1: Most used words and hashtags for Phillipe Poutou

2.1 Simple Data

Our firts idea has been to compute frequencies of each word and each hashtag for each candidate. For a lot of candidate the few most frequent words confirm the view that we have about their opinions. We can see it with Philippe Poutou (cf images **??**).

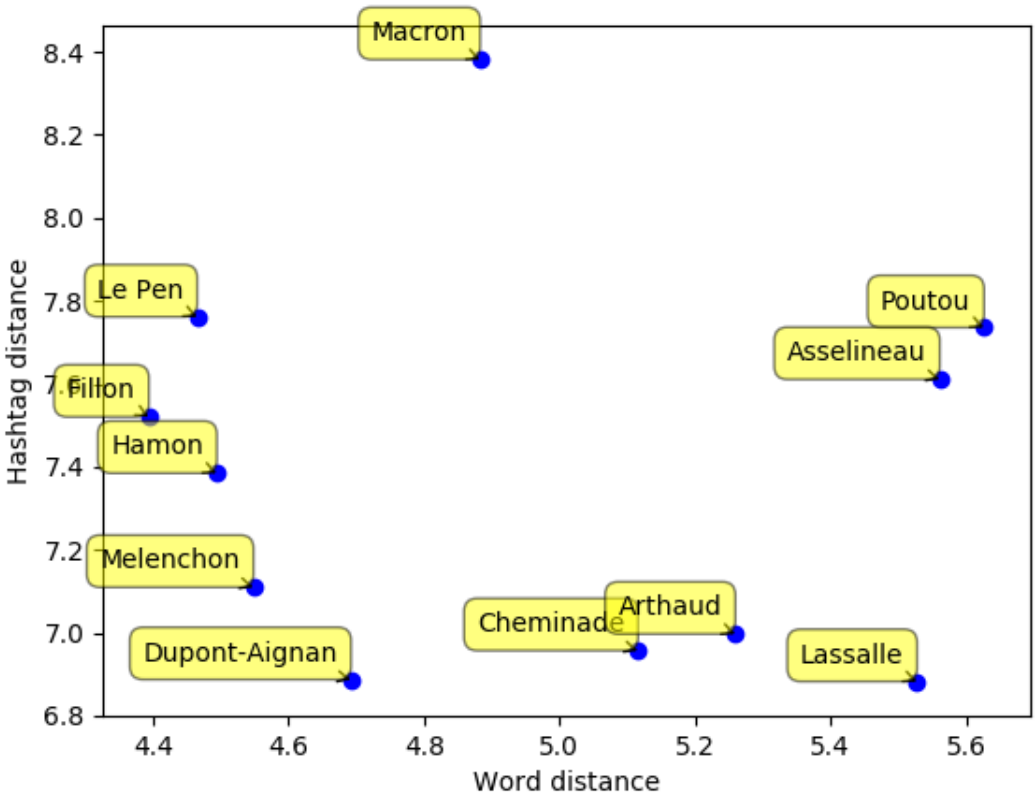

Figure 2: Sum of the distance to the other candidates, all words on x and hashtags on y

3 Distance

Definition 3.1. Distance bewteen two sets of words
Measure the proportion of words that are different between two set of words S_1 and S_2:

$$d(S_1, S_2) = \frac{1}{2} \cdot \Big(\sum_{\substack{w \in S_1 \\ w \notin S_2}} f(w) \quad + \quad \sum_{\substack{w \in S_2 \\ w \notin S_1}} f(w) \Big)$$

where $f(x)$ is the frequency of apparition of the word x.

In that way we can compute the distance in words and the distance in hastags between two candidates.

It is interresting to see that Emmanuel Macron is the most different in hashtags (cf images **??**).

Poutou	Melenchon
Cheminade	Fillon
Arthaud	Hamon
Lassalle	Le Pen
Asselineau	Macron
	Dupont-Aignan

Figure 3: Words Similarities

Marcon	Melenchon
	Fillon
	Hamon
	Le Pen
	Dupont-Aignan
	Poutou
	Cheminade
	Arthaud
	Lassalle
	Asselineau

Figure 4: Hashtags Similarities

4 Data Mining

4.1 Kmeans and Hierarchical clustering

We have implemented k-means and Hierarchical clustering. For two clusters weobtain the same result with both methods.

We can find that for the words there is a repartition minor/leading candidates and that for the hashtags Macron is alone (cf images **??** and **??**).

Melenchon	Fillon
Poutou	Le Pen
Cheminade	Macron
Hamon	Asselineau
Arthaud	Dupont-Aignan
Lassalle	

Figure 5: Before the campaign Hierarchical clustering on hashtags

Melenchon	Poutou
Fillon	Cheminade
Hamon	Arthaud
Le Pen	Lassalle
Macron	Asselineau
	Dupont-Aignan

Figure 6: During the campaign Hierarchical clustering on hashtags

4.2 Variation over time

We have observed this partition for different periods, since our data cover a few years.

We can see that before the campaign the clustering if rigth/left candidates but during the campaign it is major/minor candidates (cf images **??** and **??**).

```
Rule:  ('#fillon',) → ('français', 'dupontaignan') , 0.149
Rule:  ('#fillon', 'dlf_officiel') → ('#macron',) , 0.222
Rule:  ('#fillon', 'dlf_officiel') → ('dupontaignan',) , 0.988
Rule:  ('#le79inter',) → ('dupontaignan',) , 1.000
Rule:  ('#legrandjury',) → ('dupontaignan',) , 1.000
...
Rule:  ('judiciaire',) → ('casier',) , 0.937
Rule:  ('judiciaire',) → ('vierge',) , 0.875
Rule:  ('judiciaire',) → ('vierge', 'casier') , 0.875
Rule:  ('élection',) → ('dupontaignan', 'dlf_officiel') , 0.556
```

Figure 7: Use of the a priori algorithm

4.3 A priori algorithm

We have done the a priori algorithm.

We have observed that there is a huge number of auto-citations and very few real expressions for the main candidates, except "Front National"(cf images **??**).

5 Observations

The drowback of not using semantic is that there is no link between words with the same meaning, and even between singular and plural words.