French Presidential Election Candidates Tweets

Emma Kerinec Nicolas Derumigny

ENS Lyon

11 May 2017

Introduction

- Dataset: more than 3‘000 tweets of the 11 French presidential election candidates.
- Language: Python 3
- Package: Sklearn, Numpy, Matplotlib

Working on words

Preprocessing

- Keep only relevant words $\rightarrow$ words whith more than 5 letters.
- Distinguish hashtags with words.
- Find common points between candidates without semantic analysis

- Hard to display on a graph $\rightarrow$ need a function *words* $\rightarrow \mathbb{R}$
- Harder to see correlations

Simple Data

Compute the most used words for a given candidate.

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: Most used words and hashtags for Phillipe Poutou

Distance between tweets

Distance of sets of words

Measure the proportion of words that are different between two set of words S_1 and S_2:

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

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

Distance between two candidates

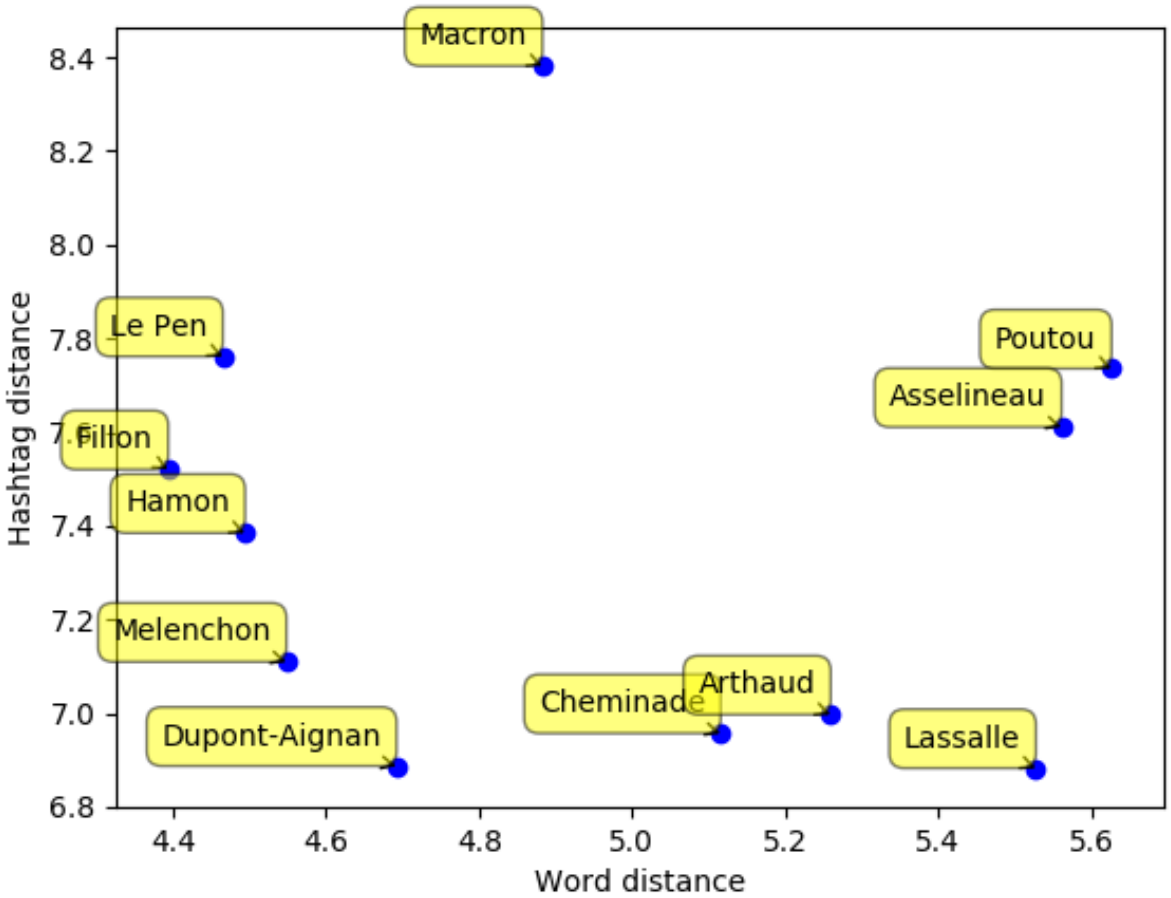

Figure: Sum of the distance to the other candidates, all words on x and hashtags on y

Kmeans and Hierarchical clustering

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

Figure: Word Similarities

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

Figure: Hashtag Similarities

Variation over time

Using Hierarchical clustering on hashtags.

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

Figure: Before the campaign

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

Figure: During the campaign

A priori algorithm

```
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
```

- Lots of auto-citations
- Very few real expressions for the main candidates, except "Front National"

Conclusion

Thank you for listening.

Does anybody have questions?