Official work with private tools

Rehan Mulakhel {rehan.mulakhel}

Brice Repond {brice.repond}

Kenneth Nguyen {kenneth.nguyen}{@epfl.ch}

Abstract

Identifying topics from a relatively large amount of textual messages could not be done efficiently a few decades ago. In this paper, we will compare results obtained from a naive algorithm based on key words system coupled with external knowledge against traditional machine learning to better understand the sensitive and sometimes confidential information contained in Hillary Clinton's emails.

1 Introduction

The Hillary Clinton email controversy arising from the use of private email server for official communications during her tenure as Secretary of State hit the headline of the media in 2015. However the content of the emails was not covered properly and reading relevant information related to it remains a difficult task. We will use the content of the emails and the network of the connection to draw conclusions.

2 Datasets description

Kaggle provides the same data in two different formats: four csv files VS one sqlite file. We only consider the former because the operations we will perform require the built in functions of pandas. The raw data with dimensions have the following shapes:

- `emails.csv` (7945, 22)
- `persons.csv` (513, 2)
- `aliases.csv` (850, 3)
- `email_receiver.csv` (9306, 3)

Theses tables form an entity relationship diagram. One would expect the emails to be linked somehow to the persons and the persons to be linked to the aliases. However, the raw data is based on the fields of the emails rather than the structure of a well designed database. Indeed, emails fields use (textual!) aliases to join the tables. Therefore, it is not possible to get the value of the persons directly from the emails.

The reasons for this inconvenient choice is likely due to the fact that the data were generated before the creation of the structure storing it. The one who gathered the data stopped the process at an early stage, leaving us to fend ourselves. Actually, the few extracted values such as the content are not useful because of the poor quality.

3 Data exploration

In general, it is not possible to organize the work in a cascade fashion when working with potentially dirty data. Going back and forth cannot be avoided and is required to understand the data. Understanding the data is necessary to improve the cleaning actions or to recover from missing or unusable values. There are easy mistakes or bad design choice we can observe by simply over reading columns. There are hidden mistakes which cannot be spot without carefully reading and *expecting* some values. We describe the recovering process of the missing and/or dirty fields in the notebook.

3.1 Emails

The subject of an email cannot always be deduced from its content because all messages have an implicit context. The same sentence does not necessarily have the same meaning all the time. It depends on too many parameters like the person who wrote it and the moment when it was written and the social conventions at the time when it was written...

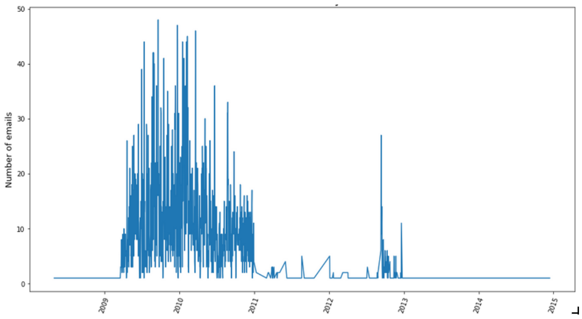


Figure 1: Time distribution of all the emails

3.1.1 Distribution over time

The Figure 1 shows three periods we call 'contexts' we identify them as:

- `ctx0` from 2009 to 2011
- `ctx1` from 2011 to 2012 June
- `ctx2` from 2012 June to 2013

We should keep in mind that some emails may have not been or could have not been disclosed.

It is hard to give an interpretation without introducing external knowledge. We can map the second periods to the Libya intervention in 2011 and the third one to the murder of US ambassador in Libya.

3.1.2 Sender and Receiver

The number of persons in the network is roughly 300 after recovering dirty values. The frequency of the messages sent or received changes in time. The topics and words changes too as we will see in section 4. However, we see in Figure 2 some names appear in the three contexts. Actually, these nodes are mostly political advisers of Hillary Clinton and they are the strongest links in the network.

3.1.3 Content

A typical email is composed of a sender, a receiver, a time stamp and a content. Some emails are made of many replies. Sometimes the replies come from the same person. We decided to keep those emails as single emails. This will not hurt the results too much since the thread of conversations should be related to the same topics in most cases.

During the pre-process, we remove the unnecessary lines and common words which appear in all emails. We keep only sentences having at least three words and join them with the char '|'.

During the post-process, we remove the frequent words by setting the threshold of the tf-idf at 3.4. We got this value by manually testing and observing the output of the words which would be removed.

4 Topics detection

The size of the dataset is too small to get valid results using common machine learning tools. We have $|\{emails\}| << |\{words\}|$. On the other hand, there are potentially a huge number of topics appearing in the few emails we have. The conditions are met for an overfit. We will use two different techniques to discover information: one basic and naive way based on key words, and one more complex based on logistic regression, hoping to reach the same conclusions.

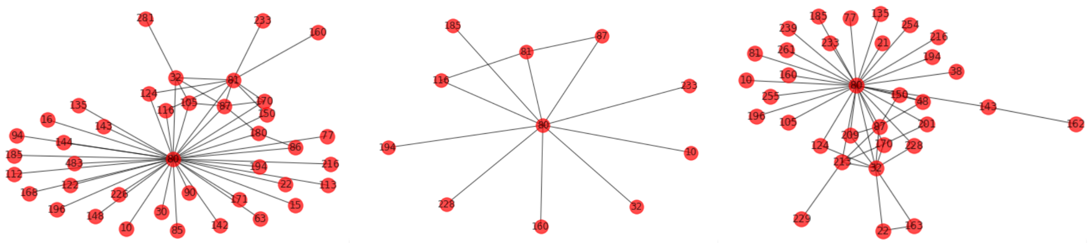
4.1 Key words

Enriching data is a common task in data science. It could be useful to map an email to one or many places on the Earth. We consider the following: North America, Latino world, Europe, Africa, Middle East, Central Asia, Far East, Russia (including Ukraine). Each of these columns is a boolean and one email can be situated in more than a single place. One naive system can determine the position based on some key words appearing in the content of the emails. The results of this solution need to be taken with cautious because it contains a lot of false positive when frequent words are used as key words, and a lot of false negative when important words are missing. For our case, we can expect 'Obama' to appear frequently even when the content is not related to the United States.

4.1.1 Emails distribution per region

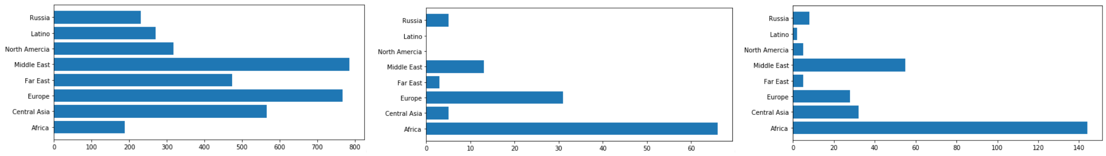
If we consider all the dataset and not the feature 'North America', emails are dominated by the Middle East and Europe, followed by Central Asia, Far East then Africa. Latino world and Russia do not seem to be concern to much by the emails compared to the others. We will neglect that and consider only the distribution within what we call `ctx` as shown in Figure 3.

If we take the plot in Figure 3a, then it not surprising to observe the same distribution as the one we have for the whole dataset. It is because the majority of emails were sent during this moment. What is surprising is the share Europe takes.



(a) ctx0 with $w(e) \geq 8$ (b) ctx1 with $w(e) \geq 0$ (c) ctx0 with $w(e) \geq 0$

Figure 2: The evolution of Hillary's network from 2009 to 2013.



(a) Distribution in ctx0 (b) Distribution in ctx1 (c) Distribution in ctx2

Figure 3: The evolution of emails distribution from 2009 to 2013 based on key words.

For the ctx1 in Figure 3b, the place dominating is the Africa, followed by Europe and Middle East.

For the last context in Figure 3c, we see that Africa is again dominating but Middle East beats Europe.

The plots based on key words lead us to the hypothesis we have drawn based on the distribution of the emails date. Indeed, Europe looses width from ctx1 to ctx2 against Africa. It seems safe to consider NATO operation in Libya as the main topic in the middle.

4.1.2 Word cloud and N-grams

We create clouds to observe the dominant terms in each context in Figure 4. Unfortunately, it does not help much for the ctx0. However, it confirms our hypothesis: ctx1 is related to Libya and 'quaddafi' while ctx2 is related Libya and 'embassi' (because of the stemming process).

We can do better using the n-grams. We decided to choose arbitrary $n = 2$. This time we apply the algorithm on the regions rather than the contexts.

For Africa, we get: 'terrorist attack', 'muslim brotherhood', 'al quaeda', 'chris steven' (US ambassador), 'civil war'.

Africa Human rights, Libya, policy transition, revolution of 2011.

Central Asia Human rights, Al Qaida, Ben Laden. Pakistant, Saudia Arabia.

Europe Climate change, European Union, UK, human rights.

Far East Petrol, climate change, human rights. Related to specific countries as North Korea, South Korea, Saudi Arabia.

Middle east Human rights, peace, missile defense. Related to Israel, Palestine.

North America Health care, human rights are some subjects.

Latino Contains only around 250 emails about subjects completely different.

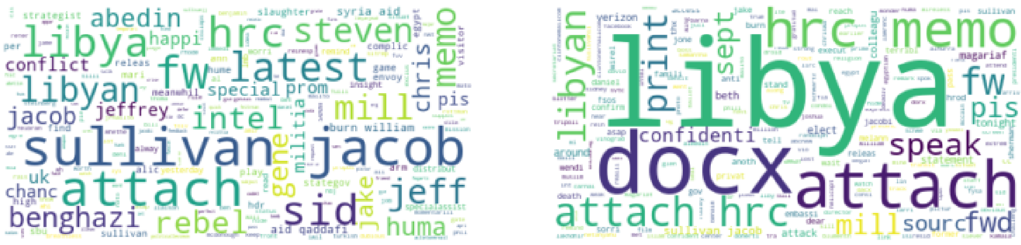
Russia Missile defense, human rights, North Korea, arm control are some interesting subjects.

One final word on the words... We have not seen anything related to Wikileaks yet. It is hard to think Julian Assange name or work were not discussed. Actually, Wikileaks appears in the word cloud per region for Europe, Russia, Middle East, Central Asia and Far East.

4.2 Machine learning

Document clustering and topics extraction can be done using unsupervised machine learning algorithm. We visualize the clusters in Figure 5 where we group *emails*. The ctx0 is not computed because it contains to many elements.

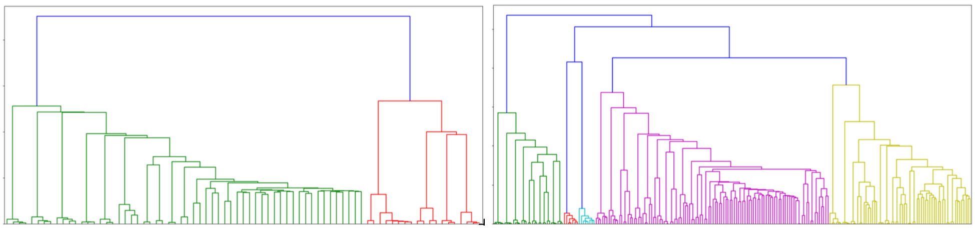
We know how to extract topics from different books. We will simply adapt our case to make



(a) Word cloud for `ctx1`

(b) Word cloud for `ctx2`

Figure 4: Words dominating in each context.



(a) Emails clusters in `ctx1`

(b) Emails clusters in `ctx2`

Figure 5: Dendograms grouped documents by cosine distance.

it work. We create our 'books' by concatenating *sentences* within the same context.

The inputs of our linear regression are not emails but sentences. The target value will be the three contexts we observe in Figure 1. The goal is to learn the topics covered in each context.

The regularization parameter is taken at the moment when we see the (cross validation) score decreasing.

The result we get are the following:

Context 0 Afghanistan, China, Iraq

Context 1 Libya (Qaddafi, NATO, fight, rebel), Egypt (Morsi)

Context 2 US Ambassador killed in Benghazi, Morocco, Tunisia, Islam and Muslim (probably because of the movie Innocence of Muslims), Romney

These topics are convincing.

The Figure 5a shows two classes. One of them is likely to be related to North Africa which includes Libya (NATO, fly zone, ...). We can suppose the other class is related to Mohamed Morsi. We can also suppose Egypt and Libya are in the same cluster and the other one contains topics unrelated to this part of the world and unrelated to each other too.

The Figure 5b shows at least two topics dominating. We guess there is one cluster containing emails related to the murder of the US ambassador. Another one probably contains events related directly or indirectly to the movie 'Innocence of Muslims'.

5 Conclusion

The naive algorithm based on key words works surprisingly well in this specific case. But it requires a lot of external knowledge and is prone to miss interesting information. On the other hand, the unsupervised algorithm is easier to implement and produces results which are more easy to interpret. In both cases, external knowledge helps to understand the context.

It is feasible to split the emails into private and professional category by identifying the personal contact in the network. Running the unsupervised algorithms could display embarrassing information. In the end, it is a bad idea to mix private and professional tools which can lead to dangerous leaks. This may also reduce the number of burn out and improve the well-being of people. But this is another subject.