MINIPEDIA PROJECT

1 Introduction

Since several years, Wikipedia became one of the most important source of knowledge in the world. Thanks to its easy and free access and its completeness, it quickly became an unavoidable source of knowledge for any person in search of knowledge. However, Wikipedia remains inaccessible in zones without a good internet access. Because of its huge size, it's very hard to store all Wikipedia's containing in a USB key.

In this perspective, we will try to construct a smart and small network from a given node. We will see two methods for the generation. The first one based on the similitude between a node and all its hyperlinks with relation to the number of views on the page. The second one based only on the content of the pages' summaries.

2 Graph generation

In this part, we test two differents algorithms to generate our graphs. Both use the same starting node (Global Warming) but spread in different ways.

2.1 Network generation

2.1.1 M.S.E algorithm

In this algorithm, we generate the graph by comparing the evolution of views during a certain period (between the 01/10/2018 and the 30/10/2018) between a page and all of its hyperlinks. The comparison is made using MSE. We decided to implement this algorithm after having observed that the evolution of views between closed pages seems much correlated (Figure 1). After that comparison we keep only the hyperlinks having an mse higher than a certain threshold. To populate our graph we start with the node n0, find the correlated (as explained above) pages, add links from the parent to these pages and iterate until the number of nodes equals 500.

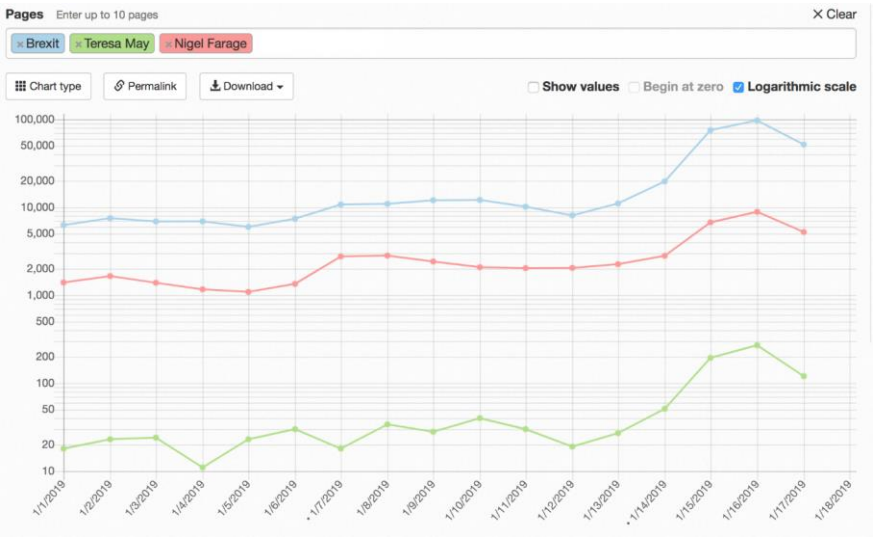

Figure 1 - Views evolution for three very closed pages

2.1.2 Header algorithm

This algorithm consists by keeping only the hyperlinks in the summary on a page. The motivation was thinking about the huge quantity of hyperlinks per page and how could we reduce them easily while following the habit of users. We start with the node n0, find all the hyperlinks of the summary, add links from the parent to these pages and iterate until the number of nodes equals 4'000.

Figure 2 - Graph of our MSE network

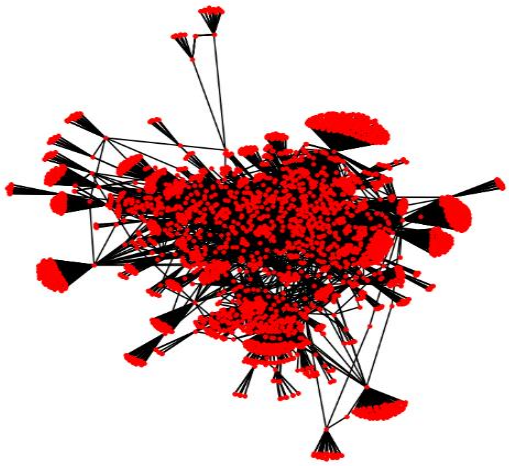

Figure 1 - Graph of our header network

2.2 Graph analysis

The main information of our networks are given in the following table. The degree distributions are also given. As we can see in the Figure 2 and 3, we can observe our networks are scale free networks.

	Header algorithm	MSE algorithm
Average degree	4.29	2.34
Average clustering coefficient	0.18	0.0
Diameter	6	11
Node's number	3'987	554
Edge's number	8'570	648

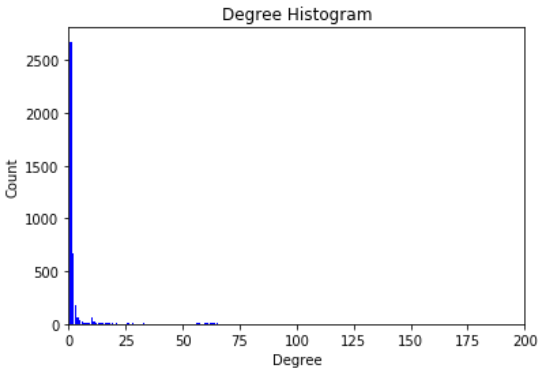

Figure 2 - Degree histogram of the header algorithm

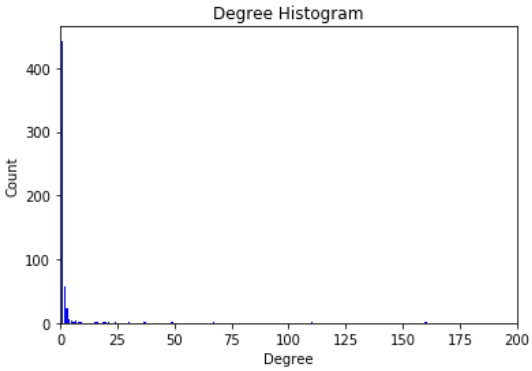

Figure 5 - Degree histogram of the mse algorithm

3 Signal

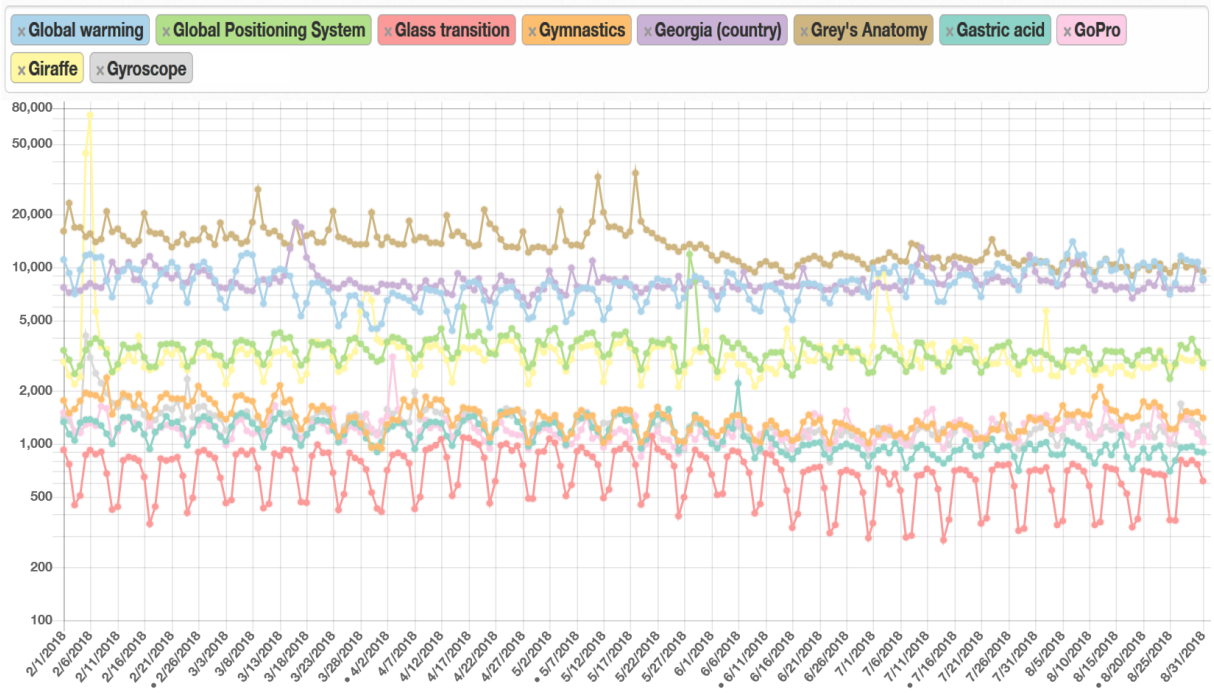

Figure 6 - Views evolution for random articles

As we can see on the image above (Figure 5), if no special events happen, there is a background page view constant signal value that changes from page to page. As we want to see variation we will subtract it from our node signals. We can also see that there is a 7-day periodic fluctuation in every page signal, we think it is an artefact because of its precise periodicity and the fact that it is synchronous between pages. Indeed it could be that views are not recorded for some portion of the day, when there is a system backup for example.

NB: We see a huge spike for "giraffe", at the beginning of February. We searched thoroughly the web but alas we couldn't find the reason behind it.

3.1 Signal scraping

In this part, we will try to do some signal processing on our network. Do to this, we scrape information about number of views on a daily basis from the 28-10-2018 to the 24-11-2018 using a Python script. The resulting signal is launch to our network, and then, analyse. We also wrote a Python script to obtain the number of modification per day, given a page, but we didn't analyse this signal.

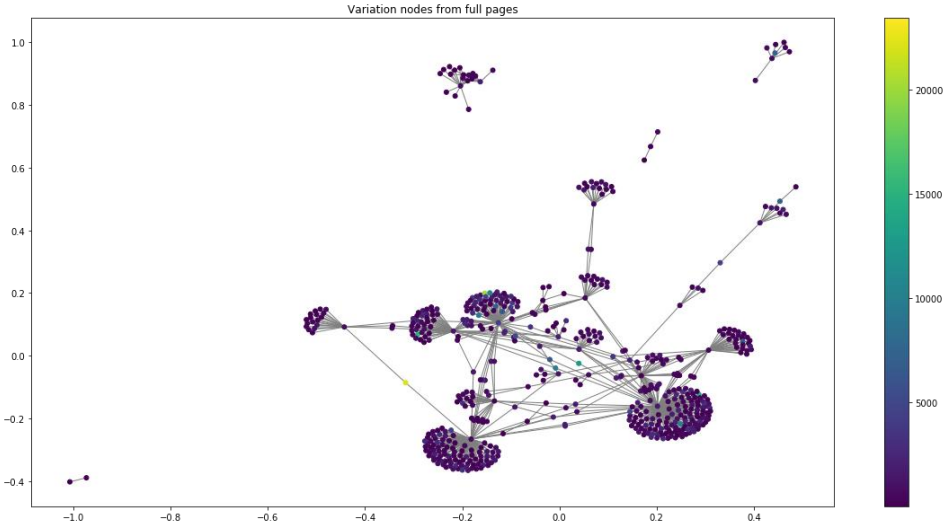

Figure 7 - Views per page of our header network (date: 30-10-2018)

3.2 Heat Kernel Comparisons on our graphs

What we see is that the graph constructed by only using the links contained in the summary of an article is much more connected, whereas the graph constructed by filtering out nodes has a lot of bottlenecks, and we actually had to search a lot to find a node (a node about "Rain shadow"), on which the heat kernel could diffuse far away.

This big difference issues from how the graphs were built in the first place and it is an expected result. The graph that didn't filter any edges is very similar to the one used in the dataset for the milestones, and both are extremely connected.

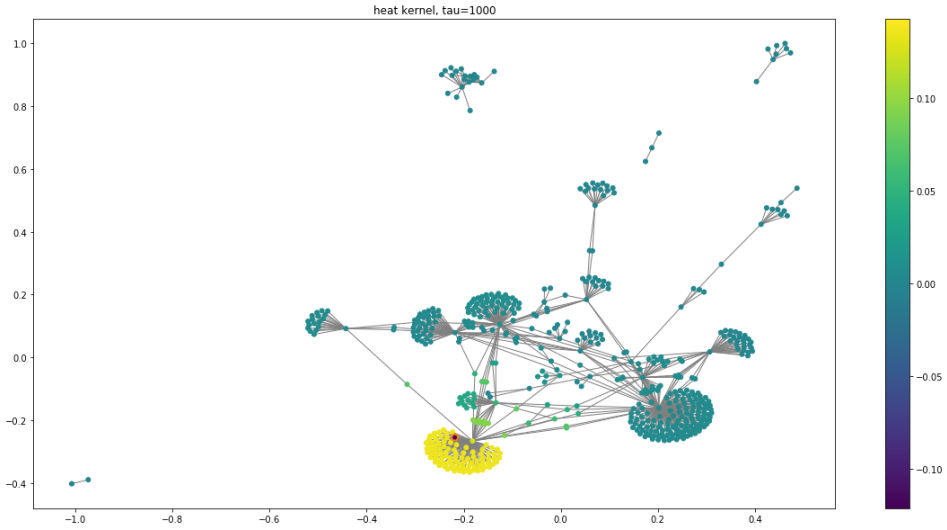

Figure 8 - Views per page filtered

3.3 Further Improvements on GSP

To go even further we can either implement a moving average on the individual signal using classical signal processing, or either implement a low pass filter on the graph signals, to filter out page view signal oscillations. However it is not necessary since the oscillations are very small and when a page become

trending due to an event in real life, the spike in view counts is easily detectable by subtracting the background average page view count.

4 Eigen map

With the Eigen maps, we verify that the different clusters have a link with the starting theme, and if it would be possible to delete a few nodes. The MSE network doesn't seem to have relevant result. That's why we only present the headers 'one.

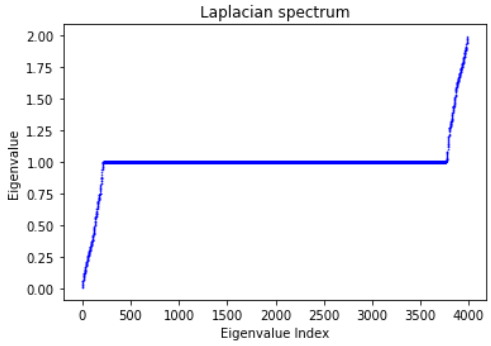

Figure 9 - Laplacian spectrum

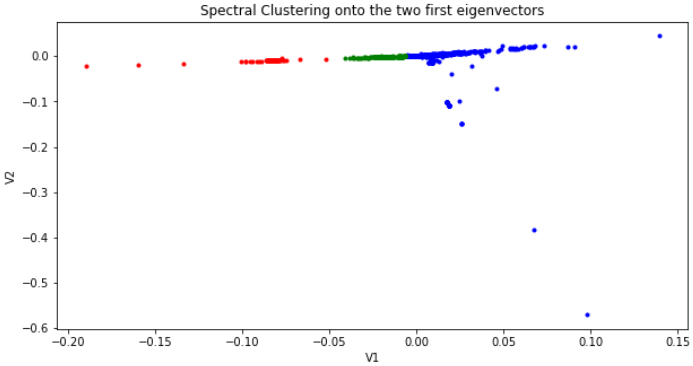

Figure 10 - Clustering

We try to increase the number of clusters, and one result is that one cluster was always present, on the theme of statistics. Maybe we found a very dense component of Wikipedia…

5 Conclusion

To conclude, the two main things that transpired throughout this study are:

- Creating a graph with the summary of the pages is not a good representation of closeness of separate themes in different articles. That said, with this technique we can see (through clustering) that the majority of the pages stay in the same theme.
- Looking at the created signal of the views per page, we can see that each page has a more or less constant 'background' number of views. Looking at the difference between the views and the 'background' we can see the influence of daily life events on the views of a page. From that we can see how people propagate in the network, and how they are influenced. Knowing how people 'propagate' could also be an interesting way to create a network, leading to Minipedias, a meaningful way to group information offline.