Semester Project: Wikipedia Activity Analysis

Raphaël Madillo

January 2019

Contents

1 Introduction

1.1 Subject

Wikipedia is an open-source encyclopedia that includes more than 40 million articles written in 301 languages. Moreover it is one of the most popular website of the World Wide Web. And finally it provides an API to access different information of the website programmatically. So with that in mind, Wikipedia becomes an amazing tool for data scientist to freely analyze the behavior of millions of users surfing over the encyclopedia [1].

1.2 Project

This project has the goal to give a simple analysis tool for visualization over a Wikipedia sub-graph. This is done by reducing Wikipedia to a graph, nodes represent Wikipedia articles and edges represent hyperlinks from one page to the other (note that these edges are directed). Then the user can see the user activities across Wikipedia pages for a given month. It also provides the possibility for users to observe the difference of activity of Wikipedia sub-set across its different language (i.e en.wikipedia.org or fr.wikipedia.org). And finally the user will be able to modify the size of graph nodes in function of the variance and the mean.

1.3 Material

I had at my disposal for this project 3 different datasets:

- Wikispedia dataset:
  The Wikispedia dataset contains 4,604 Wikipedia articles with there plaintext associated and 119,882 hyperlinks between those. These articles have no particular subject in common, but their category is registered in the dataset files [2]. For the visualization I only took the science articles, it was 1,097 articles and 13,971 hyperlinks.
- Diseases dataset:
  The "diseases" dataset contains Wikipedia 1,385 articles and 20,460 hyperlinks between those. This dataset has been built by taking all articles that are at depth two starting from Wikipedia disease category page[3]. In this dataset we don't have the full category as with Wikispedia.
- XX Century dataset:
  The "XX century" dataset contains Wikipedia 823 articles and 11,289 hyperlinks between those. This dataset has been created from all articles that are at depth two from Wikipedia XX century category page[4]. This dataset doesn't have articles category neither.

I also had the activity of Wikipedia pages accessible from http GET request [5], and the different names for same Wikipedia article were also available through WikiMedia API [6].

2 Related Work

One related work for this project is the Wikipedia Graph Mining research paper and blog [7] from Volodimir Miz. His goal was also to try to analyze common behavior of Wikipedia users, how during different events the number of views of Wikipedia pages increased. For example just before the Superbowl, US football teams Wikipedia page views increase until the date of Superbowl in February.
The approach between both projects are a bit different. In their case they focus on one event and then observe increasing number of view for a cluster of page. In my project, we first observe some uncommon user activity and then try to relate this activity to a possible explanation. Moreover with my project we couldn't clearly see cluster dynamics as they did for US football teams.

3 Work

3.1 Tools

The algorithmic implementation has been realized in Python using different libraries:

- Panda to manipulate and clean the dataset.
- Networkx to compute the community and give them colors that will be used in visualization.
- Force Atlas 2 to compute the position of the nodes.
- Requests to make http GET request to obtain user activity and article names in different languages.

Then the visualization part has been made using Javascript and Sigma.js library.

3.2 Process

3.2.1 Algorithmic

The following description will describe the pipeline used to produce the visualization for disease and XX century. For Wikispedia the process is different as we have an additional category and the format is different.

Input:
- A csv file containing all Wikipedia articles identifiers, there name and further more information.
- A csv file containing the hyperlinks with its source and its target, where the sources and targets are part of the Wikipedia sub-set we want to analyze.

Pipeline:

1. We import both csv as Panda Dataframes.
2. We filter out Wikipedia articles that are not in of the hyperlink files to only keep the articles that are in the subject we target.
3. We make http request to obtain the user activity of the filtered articles (Sometimes no activity is available for a page, for example if the page has been recently created. In this case we discard the article).
4. We iteratively filter the the article and the hyperlinks Dataframes so that every article has at least one hyperlinks, and that hyperlinks source and target are in article Dataframe.
5. We import the graph in Networkx (articles are nodes, hyperlinks are edges).
6. We make a community detection, and assign a color for each node in function of its community (we use greedy_modularity_communities algorithm from Networkx).
7. We compute positions of nodes using Force Atlas 2 layout algorithm.
8. We make two more http request. The first is to obtain the title of a page in an other language (we made it for French and Spanish). The second to get the activity of this article in the Wikipedia project corresponding to the good language (fr.wikipedia.org or es.wikipedia.org).
9. We normalize the user activities in both language, and compute the mean and standard deviation for each article.
10. We export the graph as a json with all the characteristics that we previously computed.

Output:
- A json file with nodes and edges of the graph.

The whole process takes about 15 minutes.

3.2.2 Visualization

The json is then imported using Javascript library Sigma.js to give the following visualization.

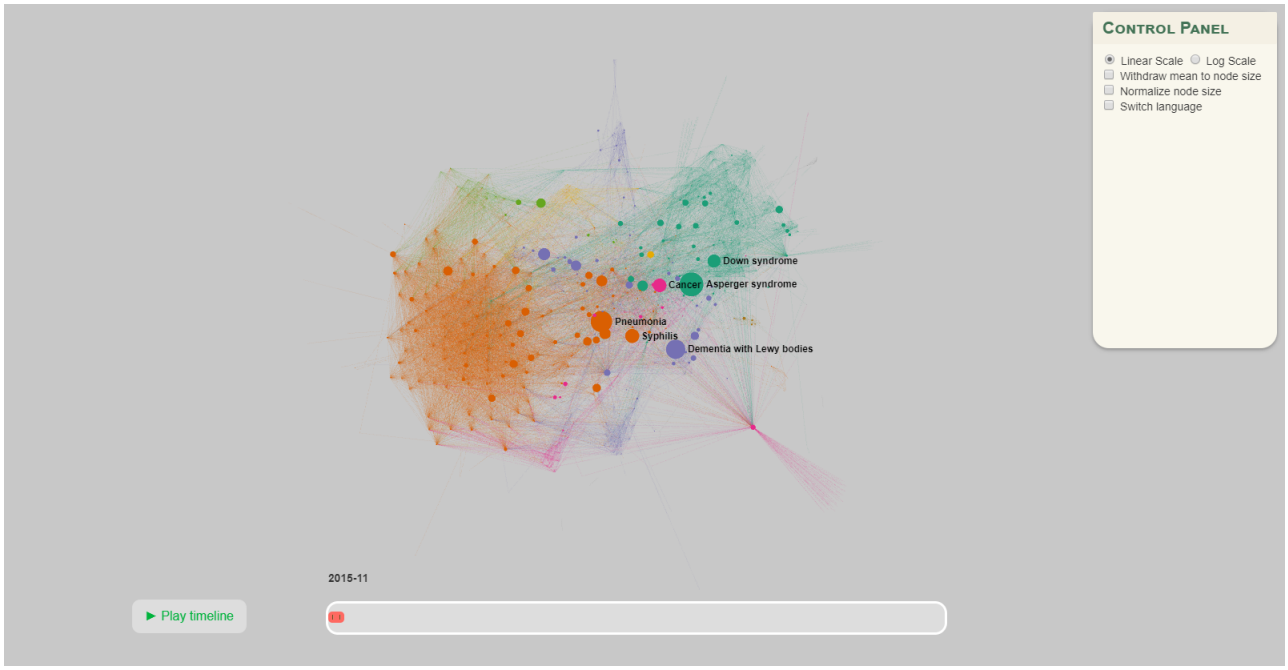

Figure 1: Overview of the visualization

The visualization permit further manipulation of the graph:

- Selection of a node to display only its neighbors and show its views at given date (in Wikispedia visualization we can also see node category)
- Manipulation of node size in Control Panel to give to the analyst different points of view to observe the graph.
- Possibility of changing language by clicking a check box.
- Timeline that is clickable and draggable to observe user activity at a given month, and a button to play the timeline passing month every two seconds.

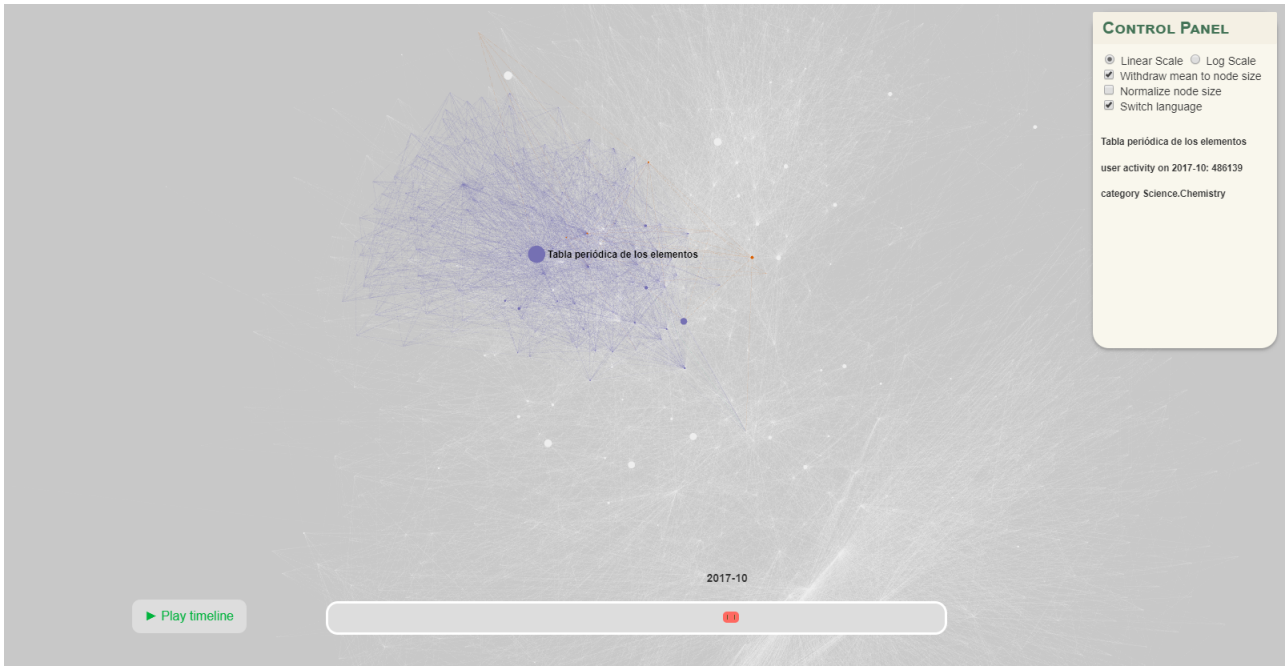

Figure 2: Visualization activating different visualization options

4 Results

The result we obtain is a powerful tool to analyze user activity over Wikipedia, and we will see some examples that make this tool really interesting.

4.1 Sanfillipo disease

We begin by observing the diseases graph on September 2018 in French and we observe the following pattern.

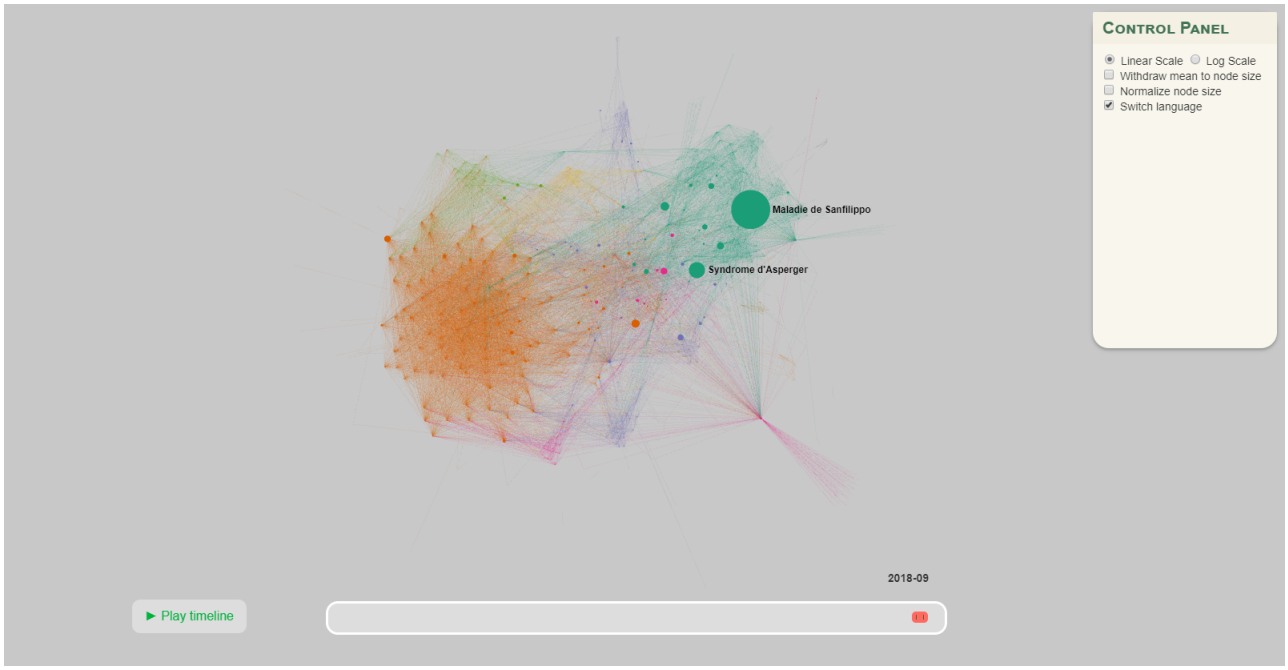

Figure 3: Diseases graph in French on September 2018

We then produce a plot of the time serie of user activity on Sanfillipo Wikipedia page to confirm an abnormal activity [5].

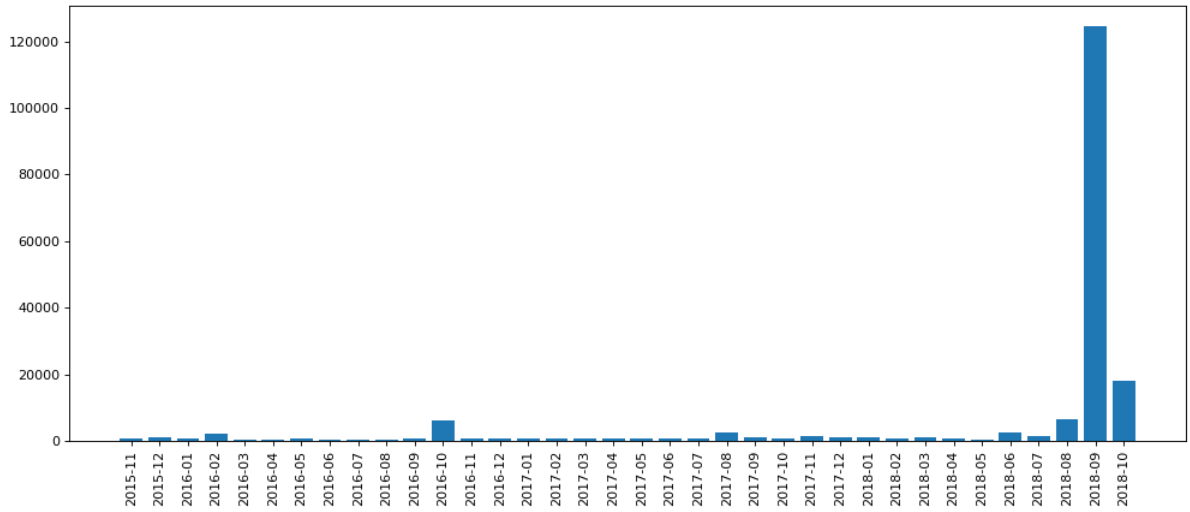

Figure 4: Sanfillipo user activity time serie

We can see on this temporal analysis that the activity of Sanfillipo Wikipedia page is 24 times higher on September 2018 than in average.
The explanation behind this very high sudden activity can be found directly on Sanfillipo Wikipedia article. The reason is that on September 17, a TV film "Tu vivras ma fille" was broadcasted by one of the main French TV broadcaster TF1, and that the main character had this disease.

4.2 1996 in Music

The same happen with the Wikipedia article 1996 in music (in English Wikipedia this time), that is in XX century visualization and the time serie is as demonstrative as the time series for Sanfillipo article.

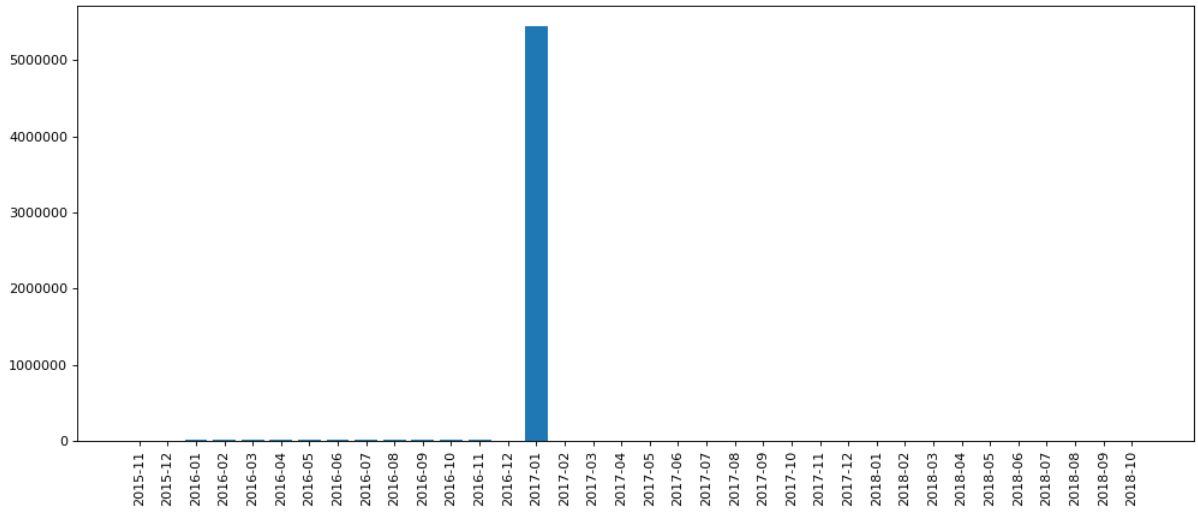

Figure 5: 1996 in Music user activity time serie

From this activity analysis, we can see the activity of 1996_in_music page on January 2017 is of 5.447.653 views (34 times the average activity), but here the explication behind this peak is unknown.
It is the most viewed page (after wikipedia main page and wikipedia search page) in mid January[5], however doesn't appear on Wikipedia trend page for mid January[8][9].
A possible reason why it doesn't appear on the trends is that the top 25 list excludes "articles that have almost no mobile views (5–6% or less) or almost all mobile views (94–95% or more) because they are very likely to be automated views based on our experience and research of the issue." And when the page had 1.131.599 desktop views on 15 January 2017, it only had 224 views from mobile-web and 44 views from mobile-app.

4.3 Winter diseases

With the visualization tool we can also observe notable trend by seeing node size increasing for a community of nodes. The classical example that we are able to see is that during winter some disease are more present, for example tonsillitis and common cold.
On French Wikipedia these trend are easily observable looking at the time series.

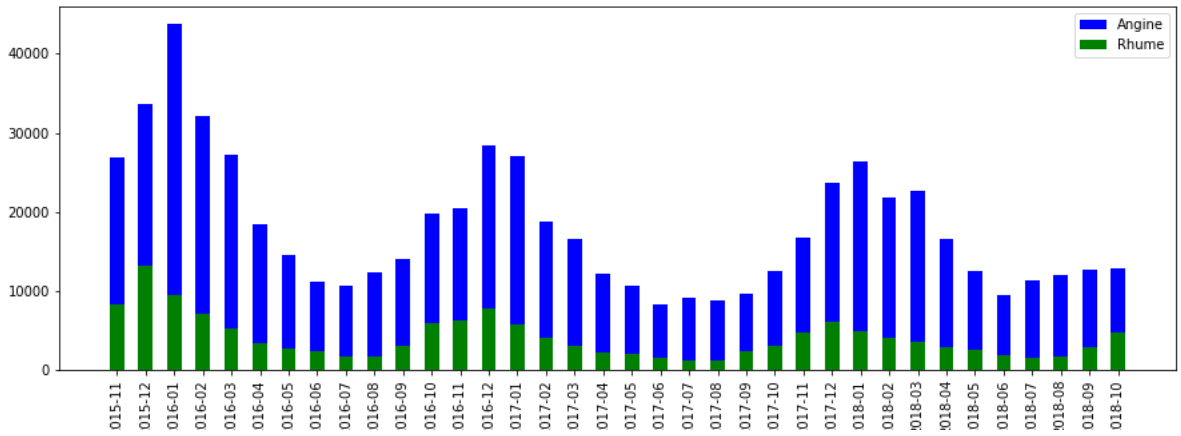

Figure 6: Angine (tonsillitis) and rhume (common cold) time series

But this phenomenon is less visible on English Wikipedia (maybe because both hemisphere look at English Wikipedia and don't have winter at the same time).

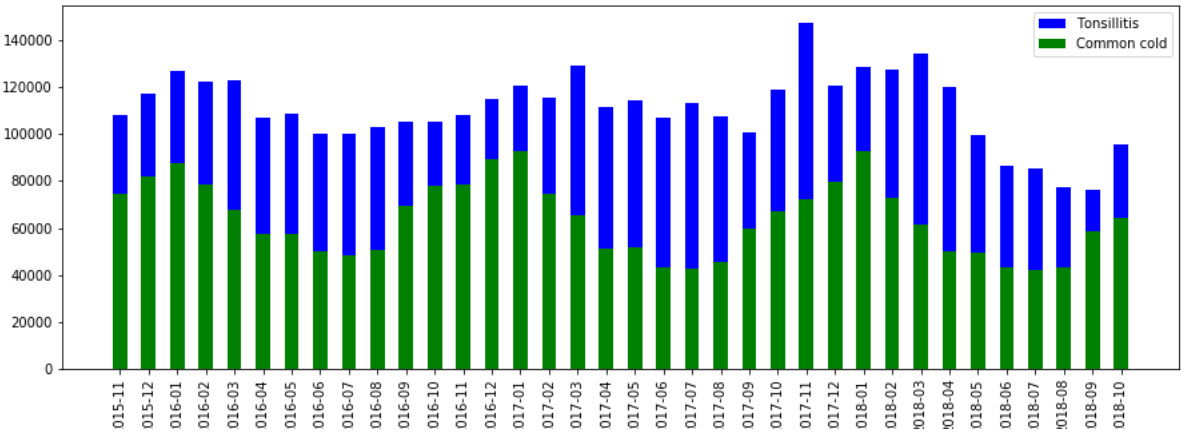

Figure 7: Tonsillitis and common cold time series

4.4 Chipko movement

We will finally analyze the Chipko movement user activity, to see that there are also interpretable peak of user activity on English Wikipedia.

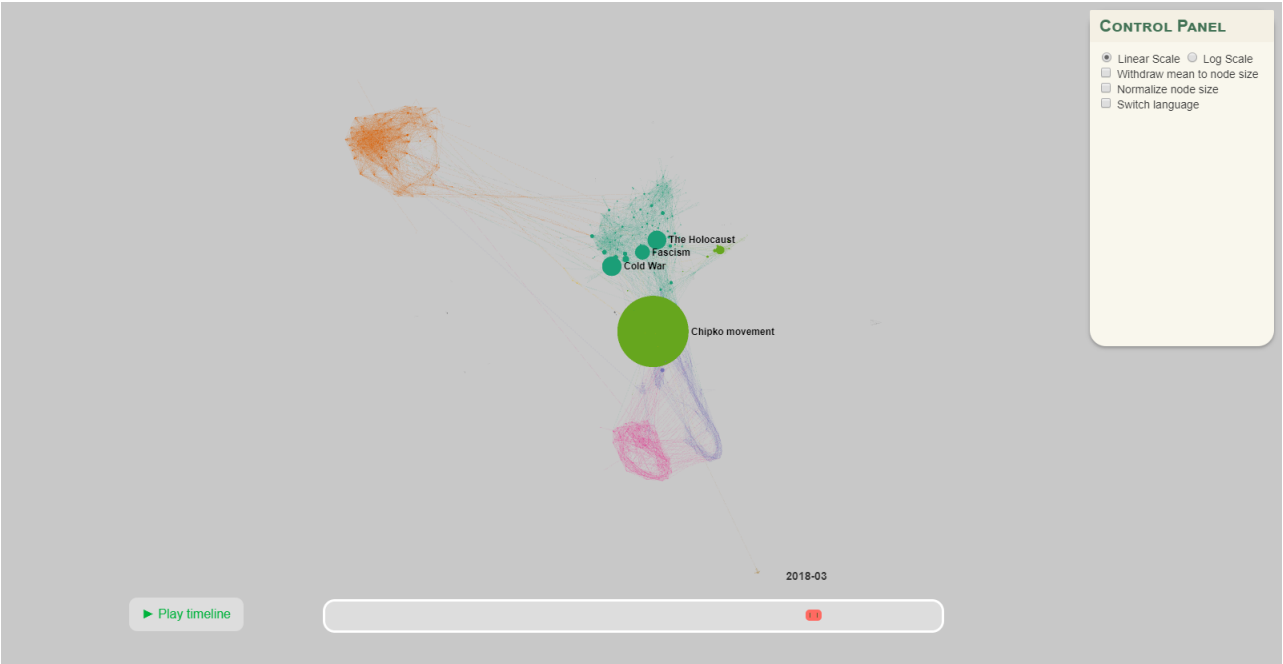

Figure 8: XX Century graph in English on March 2018

The time serie plot is also representative of this event

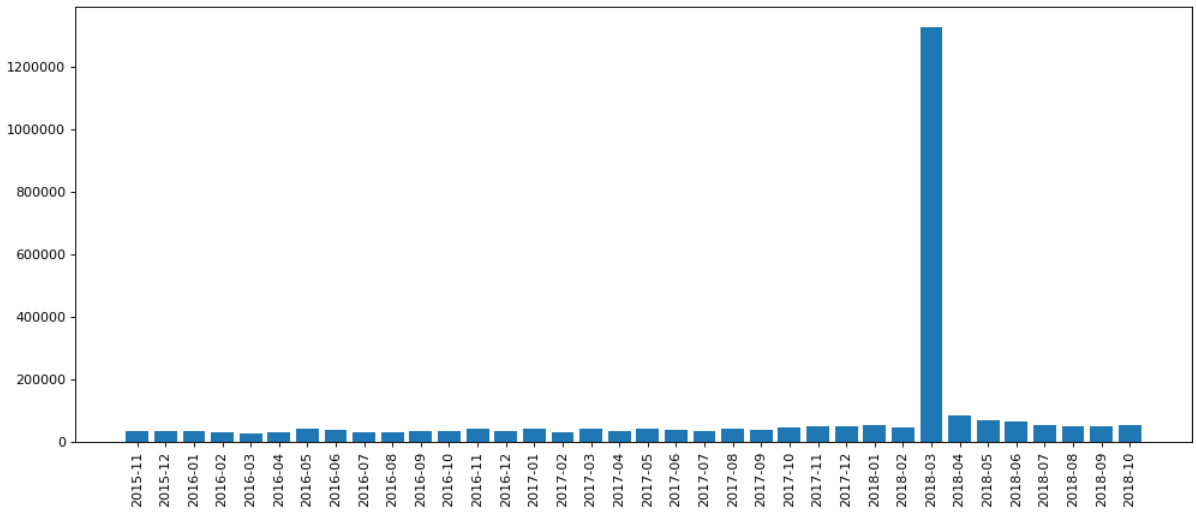

Figure 9: Chipko user activity time serie

Or this is due to the Google doodle on March 26 that was related to the 45th anniversary of Chipko movement.

5 Conclusion

As we saw with the results, some peaks of user activity are highlighted by the graph and by searching further explanation we can sometimes find the causes of this high activity. This is really interesting as it could help data scientist to analyze web phenomenon and gives clue about how we use Wikipedia in general.

This tool also gives the possibility to its users to compare the views across different Wikipedia projects as for example between French (fr.wikipedia.org) and English (en.wikipedia.org) to discover if the user activity correlates between both or not.
And finally thanks to the disposition of the graph, the users can have some insight about the organization of the encyclopedia and its clusters.

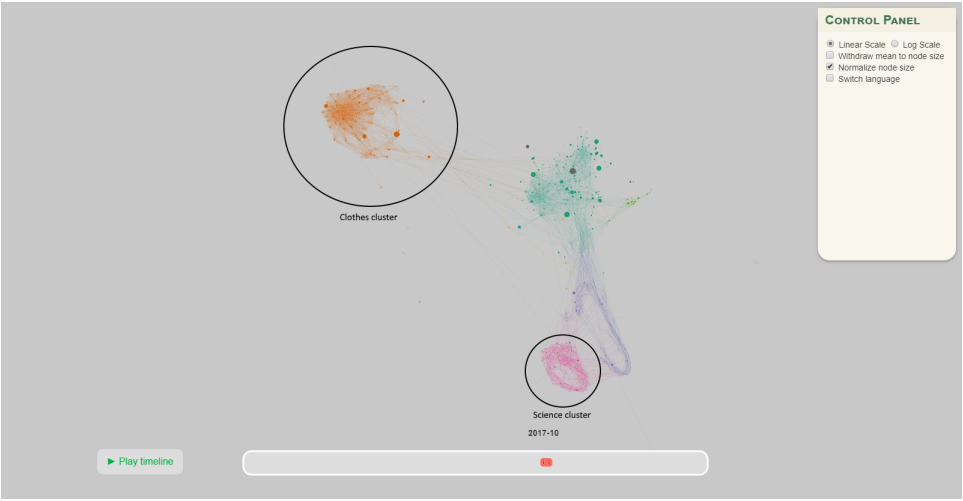

Figure 10: Example of visible clusters in XX century visualization

5.1 Limitations

However as we also saw in the results, not all events are interpretable. For example Music in 1996 isn't interpretable, we also don't see trends for winter diseases on english Wikipedia.
One limitation of this project in general is that not all events are interpretable and sometimes it's not so easy to find what is the causes of millions of views on a page that isn't viewed so mush usually.
We also remark from the results that it is harder to find interpretability in English Wikipedia as the whole world (with different interest, hobby, way of living...) is consulting it, and so trends are harder to observe. Compare to French Wikipedia where as there is less viewers and these viewers are more homogeneous, it becomes easier.

Another limitation with this work, is that we are limited in the size of the Wikipedia subset we want to show. For example during this project I tried to display diseases graph with depth three starting from disease Wikipedia page. The result was a graph with 7,616 nodes and 197,663 links, and it gave a really slow visualization.

5.2 Further work

There are still some amelioration to bring to this project and more generally a lot of work to do to understand user behaviors on Wikipedia. Possible ideas for further work on this project could be:

- Working more on the layout (disposal of the points) as it can carry more meaning, it could be a great amelioration to go deeper on the layout to highlight some internal organization and clusters of Wikipedia. Example of good layouts can for example be found in the different projects by Kirell Benzi, where the disposal of the graph gives more meaning [10].
- Give a fully automated tool to user where he can register the Wikipedia page he wants, at what depth from this article he wants the graph, and between what dates he wants it, and then produce, after few minutes or more depending on the size of the future graph, the visualization. This could be a formidable tool for data scientist and curious users to explore Wikipedia.
- Find the categories of the different Wikipedia pages as it is done in Wikispedia dataset to color nodes by category.

5.3 Availability

The whole work is available at this github and you can see the related blog here.

References

[1] Wikipedia - Description Page
https://en.wikipedia.org/wiki/Wikipedia

[2] Wikispeedia navigation paths
https://snap.stanford.edu/data/wikispeedia.html

[3] Wikipedia diseases category article
https://en.wikipedia.org/wiki/Category:Lists_of_diseases

[4] Wikipedia XX century category article
https://en.wikipedia.org/wiki/Category:20th_century

[5] Wikimedia REST API
https://wikimedia.org/api/rest_v1/#!/Pageviews_data

[6] MediaWiki Langlinks API
https://www.mediawiki.org/wiki/API:Langlinks

[7] Wikipedia graph mining: dynamic structure of collective memory
http://blog.miz.space/research/2017/08/14/wikipedia-collective-memory-dynamic-graph-analysis-graphx-spark-scala-time-series-network/

[8] Top 25 Report/January 8 to 14, 2017
https://en.wikipedia.org/wiki/Wikipedia:Top_25_Report/January_8_to_14,_2017

[9] Top 25 Report/January 15 to 21, 2017
https://en.wikipedia.org/wiki/Wikipedia:Top_25_Report/January_15_to_21,_2017

[10] The Art of Kirell Benzi
https://www.kirellbenzi.com/art/