What can you learn from Wikipedia without even reading it?

George Abi Younes
george.abiyouens@epfl.ch

Amaury Combes
amaury.combes@epfl.ch

Patryk Oleniuk
patryk.oleniuk@epfl.ch

Abstract

In this project, we are using data and statistics skills learned in ”Applied Data Analysis” class. The project topic conforms to the ”data science for social good theme. This papers shows a correlation-based analysis on Wikipedia and GDELT as well as UNData data sets in order analyse countries properties and its relations to its Wikipedia page. We used Python tools to perform the analysis and to display the results. Additionally, we aggregate the results and show world statistics on the countries Wikipedia usage.

1 Introduction

In the light of recent events, we noticed that after a substantial event, be it international, local, or even specifically related to a famous person, the Wikipedia page of that entity gets updated quite fast and abundantly in a relatively short period of time. Following that, we decided to attempt to combine Wikipedia datasets with an events dataset comprising of anything substantial happening in the world (political conflict, natural disaster or important news). Following are the research question that we attempted to answer throughout the project:

1. Is the number of changes in country-clustered Wikipedia information a good indicator of stability of the country? Could we try do estimate the stability of a country by this factor ?

2. How often do citizens contribute to Wikipedia and does it depend on the countries' wealth?

3. Which type of events cause the most changes in Wikipedia?

4. Which countries react very frequent and on-time with putting events in their Wikipedia page?

2 Data Collection

The first step was to choose the proper sources providing reliable and robust data. Following are the data sets we use for our data collection:

2.1 Wikipedia:

In order to grab the edits history for any Wikipedia National page, we used Wikipedia's API (2) and proxied it through our HistoryFetcher class included in our source code.

2.2 GDELT:

We initially selected the UCDP GED dataset in order to get the ”main events”. Unfortunately, as this dataset contains only armed events, we were not able able to find an adequate preliminary visual argument for our intuition and thus sought an alternative. Hence we decided to switch to a richer dataset: GDELT (1), which is a free cloud-based service that offers a variety of tools and services to allow you to explore, and export the GDELT Event database containing different event types. We enclosed it in our source code and fetched daily event for all available countries between 2011 and 2017. We also included the column *Quadclass* 1. To handle the data, we used a library called gdelt-PyR. Since it is very memory-intensive process (Dataframe of more than 50GB in our case), we performed this task on the remote virtual machine with reconfigurable amount of memory (Google Cloud) and by separating data in batches me managed to aggregate the data. Hence, at the end of this operation we have a number of events per QuadClass per country per day. This is available to store in a normal PCs memory (50MB).

2.3 UNData:

(3) This dataset was used to obtain basic informations about the countries in the world, including population, GDP and % of Internet users, as well as country codes.

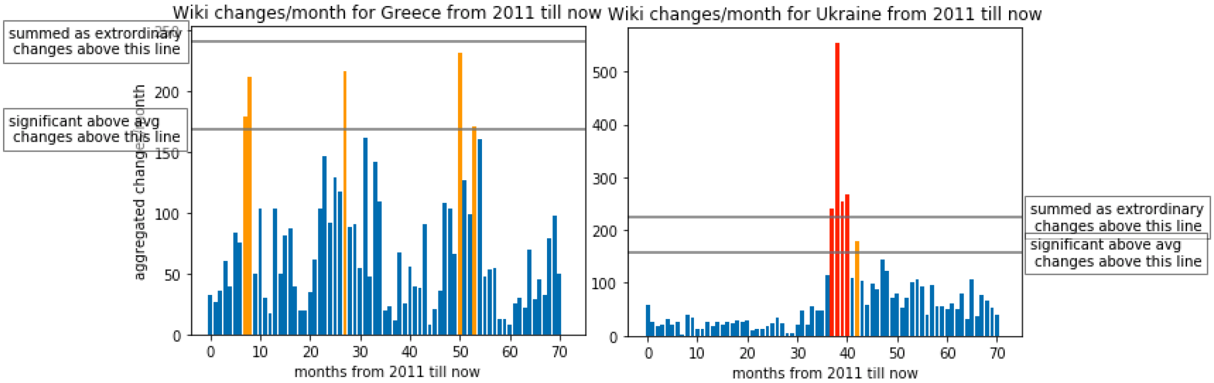

Figure 1: Wikipedia Edits for Greece and Ukraine Analysed by the Algorithm. Results of instability-Greece:0.04, Ukraine:0.29

3 Prototype Wikipedia Changes Index

This index is rather intuition-based and simple. We have found that it does lead some righteous results. Nevertheless, the index still suffered from certain flaws explained below.

3.1 Algorithm

We designed a simple country classifier function, which would assign a numeric value on a country based on its Wikipedia changes. We expect it to give results $r \in (0,1)$ and the higher the results, the more *unstable* the country's Wikipedia. The Algorithm to compute the instability:

1. We fetch the Wikipedia page changes for a selected country and aggregate it monthly.
2. Afterwards, we use our own tunable outlier detector to detect the time periods with extraordinary edit rates and label those as outliers. The extraordinary monthly edits $edits_{eo}$ and significant monthly edits $edits_{sig}$ are classified as follows:

$$edits_{eo} = edits \geq c_{oef} \cdot avg(edits_{tot})$$

$$edits_{sig} = edits \geq 0.7 \cdot c_{oef} \cdot avg(edits_{tot})$$

and

$$edits_{sig} \notin edits_{eo}$$

3. Finally, the instability index r is calculated:

$$r = \frac{\sum edits_{eo} + 0.2 \cdot \sum edits_{sig}}{\sum edits_{tot}}$$

where *eo* subscript stands for extraordinary daily edits, *sig* for significant edits and *tot* for all edits. The $coeff$ has been chosen experimentally as $c_{oef} = 3.4$.

3.2 Results & Main Findings

The algorithm works as shown in Figure 1. The e_{eo} events in Ukraine's Wikipedia page have been highlighted in red and the e_{sig} in orange. We then generalized the application of this algorithm to measure its robustness on several countries.

When analyzing more results in Figure 2 we can observe how well our prototype stability index performs. As expected, the general trend of how the prototype classifies the countries is correct to some extent but with some exceptions. For example, developed countries such as France and Norway obtain a low instability score reflecting good stability of the country. Whereas, countries suffering from conflicts such as Ukraine and Niger obtain a very high instability index by our algorithm. A few exceptions that can be seen are Germany, Ireland or Sweden whereby our index labeled those with a high instability index. This is probably due a high activity that is not necessarily due to conflict on their Wikipedia page. Also, peak Wikipedia changes do not always cause country instability, e.g. political changes. Additionally, we could see that Greece obtained very low score of instability even if we could observe lots of peaks, possibly related to the Euro crisis. This is because if the country has big variations of the Wikipedia stability very frequently, this classifier could be wrong. This is why other solutions were applied.

4 MLE Classifier

4.1 Algorithm

This classifier is based on the Maximum Likelihood Estimation equations. In particular, for each events $event_m$ in the month m:

$$results = ???/avg(events_{tot})$$

Some other comments

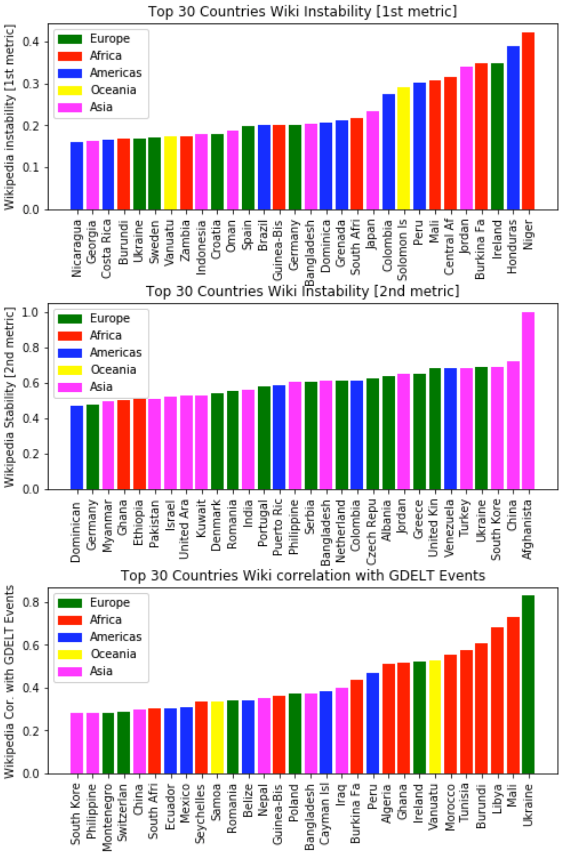

Figure 2: Comparison between 3 used algorithms.

4.2 Results & Findings

As shown in the Figure 2, the MLE method performs a bit better than the 1st one, because Ukraine is the most unstable European country, followed by the UK and Greece. The presence of UK or Germany in TOP 30 is probably caused by the irrelevant Wikipedia peaks. Hence, since we cannot easily detect the event types on Wikipedia, we are to combine it with the GDELT events n the next section.

5 Wikipedia and Event Correlation

In this section we are correlating the GDELT event data (4 event classes, per day) with the Wikipedia changes. Before analysis, the daily data is aggregated by month to counteract the inter-event peak delays. The correlation coefficient is calculated based on the Person Coefficient (4) between to data series, which returns $r \in (-1, 1)$ - correlation coefficient and $p \in (0, 1)$ - the probability that the data is uncorrelated even when having the correlation r. For each country we are saving the value $p \cdot r$. This is to counteract the fact that there are no events in the country and therefore stable Wikipedia will be highly correlated with stable events. This highly increased our accuracy and precision, since if the return values is high the Wikipedia is very probable to be correlated. You can observe example correlation in Figure 3.

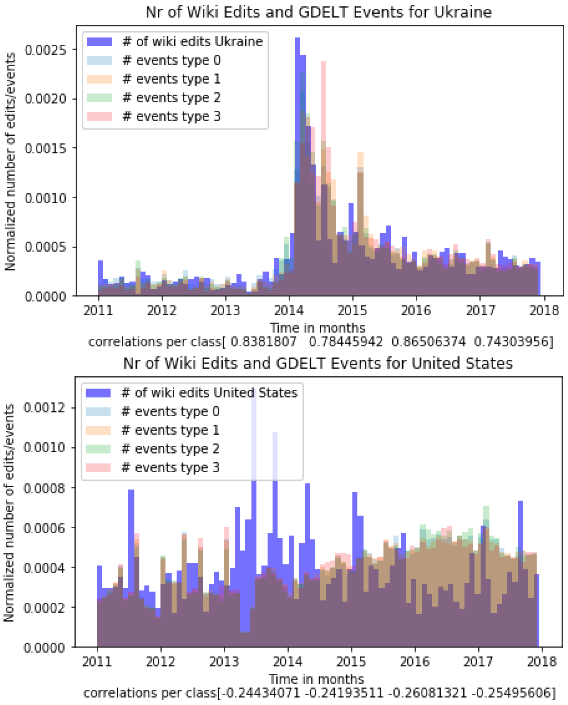

Figure 3: Example of Wikipedia Changes and GDELT Events correlation for 4 event types for Ukraine and US.

5.1 Results & Findings

Results from correlation are shown in Figure 2 and more specifically 3. It is very interesting to see Ukraine in the first page. This is probably because many of people use Internet and were highly affected by Donbass war. We could also see that the North African countries are clustered with very high correlations, probably because of the ”Arab Spring” happening 2010-2015. High correlations can mean bad (wars) but can also mean very good, it mean that the Wikipedia is reacting to the events. Some countries that we believe should be highly correlated are not (France, USA), probably because of the number of unimportant events reported to GDELT which were note relevant to Wikipedia. The negative correlations, especially the higher ones, are unexpected and should be investigated further.

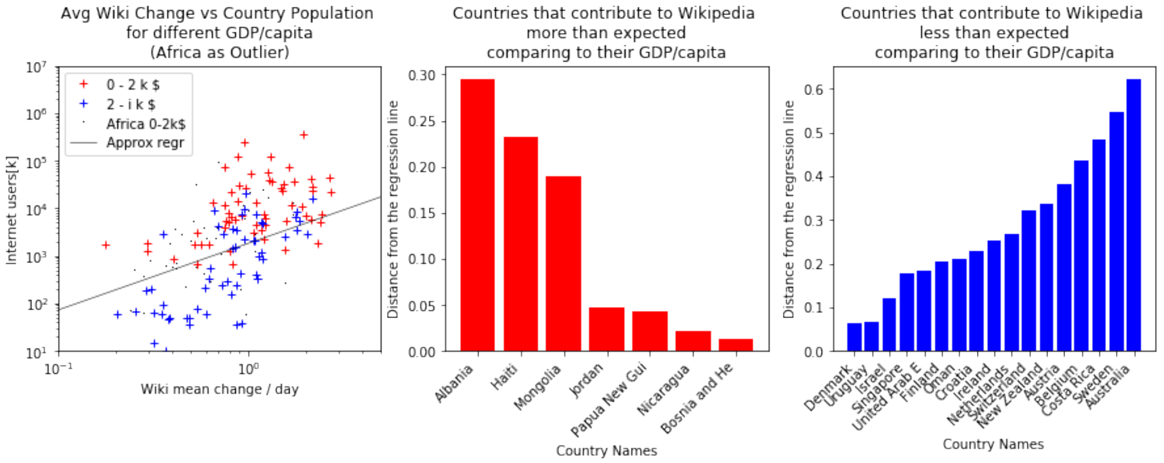

Figure 4: Mean changes in Wikipedia vs Population and GDP/capita together with regression classification and outliers.

6 General Wikipedia Activity Analysis

In this section we try to analyse citizens habits about updating their national Wikipedia page and conclude on the differences between countries types and event types.

6.1 Mean Wikipedia changes vs GDP

We tried to find correlation between the countries properties and the mean national Wikipedia changes. We found, that it is highly correlated with the wealth of the citizens and number Internet users. To prove that, we plotted in Figure 4 the 2 values in double log scale. We could observe clustering, in particular that the more wealthy the country is, the more do citizens contribute to Wikipedia. We also found many exceptions, for example African states do not follow this trend and are spread in all the area. Hence, we ran the linear SVM regression (5) line (in double log axis) without this data. Additionally, afterwards we were able to find which countries contribute less / more (blue and above the line / red and below the line) in comparison to our simple model. We could then find, which countries do NOT follow our regression model, e.g. Australia should contribute much more and Albania, even though is poor relative to others, contributes a lot.

6.2 Type of Events that cause change

Finally, we analysed which type of events cause the most change in Wikipedia n average. We computed a weighted average of correlations for all countries, where the weights are the number of the Internet users.

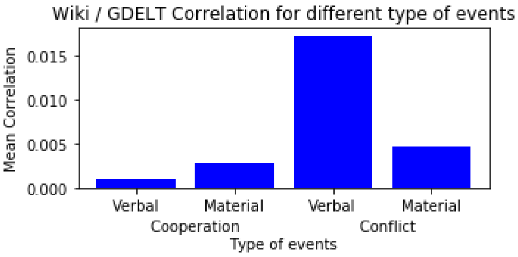

Figure 5: Wiki. Changes/GDELT corr. for diff. event types.

The Avg_{class} is shown in Figure 5. As expected the conflicts have much higher correlation, so cause much more changes, than the cooperations.

7 Conclusions

The number of changes in country-clustered Wikipedia information could be a good indicator of stability of the country, but not very reliable. Because many factors influence Wikipedia changes and not all of them are significant events. We estimated the relation between the wealth of the country and the mean Wikipedia changes. In general, each country has also different temperament and political situation which could be taken into account in further analysis.

8 References

1. $GDELT\ data\ format\ codebook$
http://data.gdeltproject.org/documentation/GDELT-Data_Format_Codebook.pdf, Page 4 about $QuadClass$ data type.

2. $Wikipedia$
www.wikipedia.org

3. *UNData*
www.data.un.org/

4. *Pearson Correlation*
https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.pearsonr.html

5. *Linear SVM*
http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html#sklearn.svm.LinearSVC