Capstone 2 Proposal
Correlating web traffic to external events

Jonas Cuadrado

1. **Problem to solve:**

Knowing how different event affect the behavior of a company's client can allow a company to improve its service, wither by offering specific products or by adapting to the clients needs, as well as allowing the company to predict trends based on seasonal trends.

In this example, the goal is to correlate external events to Wikipedia searches spikes. The external events will be tracked from news articles written by real journalists from different media. I have chosen Wikipedia due to its publicly available data, its popularity, and its ability to attract plenty of users from diverse backgrounds.

2. **Client:**

In this specific project, Wikipedia could use that information to predict higher demand of traffic and open more servers to reduce delays. This can be generalized to any company.

In general, every company should be interested in how their users behave. Being able to track their interests and observe correlations between events is of upmost importance to determine how to improve and where to invest in the corporate model.

3. **Datasets:**

There is a readily available dataset of traffic to Wikipedia per hour on https://dumps.wikimedia.org/other/pagecounts-ez/projectcounts/. The data is massively compressed using a simple scheme:

In every line there is a project code (namely, language and few other properties, the title, the monthly total views and hourly counts compressed as follows:

- Hour: from 0 to 23, written as 0 = A, 1 = B ... 22 = W, 23 = X
- Day: from 1 to 31, written as 1 = A, 2 = B ... 25 = Y, 26 = Z, 27 = [, 28 = \, 29 =], 30 = ^, 31 = _
- Example: 33 views on day 2, hour 4, and 155 views on day 3, hour 7 are coded as 'BE33,CH155'

This allows a whole month of data to be stored in a single 4 or 5 GB file.

For the news, Kaggle has a public dataset containing articles from different sources for a period of time of over 1 year: https://www.kaggle.com/snapcrack/all-the-news

To better extract the data, some NLP will be required to obtain the topic, at least, for most of the articles. The date is immediately available, which facilitates the time correlation between events.

4. Approach:

An initial analysis of the data will determine a threshold to what is a significant spike. A 400% increase may not be meaningful if it goes from 2 to 10 views in one hour. To look at the spikes I will use some distributed platform like TensorFlow or Pyspark, and then use it to export it to a pandas dataframe when the data is filtered and reduced.

The spikes will have time information, and title. For every spike, I will look at articles of dates near the spike and determine what came first, where, and the topic. It is likely that social media trends may see a reflection on the press later than other topics, for example.

Finally, I will look at time correlations between the events, try to find delays or other interesting data properties.

5. Deliverables:

A milestone report with initial data analysis and quantification, probably with the spikes filtered.

A final report with all the details, codes, and a slideshow that describes the project.