Whatever-OZS
Shixin Li - sl3368
Zewei Liuz - zl1446
Osvaldo Bulos - obr214

M11 Realtime dashboard for Wikipedia streaming data

Project description:
The end goal is to use the steam of Wikipedia page views and edits to identify global events in realtime. Whenever a major event takes place (a plane crash, the results of an election, a terrorist attack, Nobel prize awarded, etc) takes place the relevant Wikipedia pages are updated and visited by interested people, resulting in peaks on both the number of page views and edits. Whenever such a peak is detected we will looks for similar peaks in the corresponding page in other Wikipedia editions. It peaks are detected in multiple (3+) language editions at the same time, we know we are in the presence of a global event. By analyzing the way in which attention is paid to events in different editions we will model how cultures mutually influence each other, how relevant one culture is to another and to cluster editions based on their mutual attention. Finally a predictive model will be developed to predict the effect of an event occurring in one location might have on others.

Data sets:
Wikipedia edits and page views

Goal:
Online visualization of the global events detected on historical data and the predictive model.

Preliminary Idea:
The first step of the project will be to identify the major news from Wikipedia.
We are taking in consideration 3 approaches to accomplish this task:

1. Obtaining the news from the Wikipedia's Current Events portal.
2. Identifying the newest created articles related to events.
3. Identifying peaks from articles that already exist.

From these events, the information we are interested will be primarily page's views and their categories given by Wikipedia.
This task will be replicated for the different Wikipedia Editions.

A dashboard will be implemented so the user can select a date, obtain the top events for that day, and compare the number of views for that event across the different Wikipedia editions.

Once this information is collected and cleaned, a predictive model will be implemented.
The main task of this predictive model will be, given a series of categories, determine the effect of the same article/event on the different Wikipedia Editions (different languages).

Technical Issues:
We are planning to build a web scraper using Python and scrape news from wikipedia as our dataset. This scraper will be implemented so it can be run periodically. It needs to be run at least once a day.

Whatever-OZS
Shixin Li - sl3368
Zewei Liuz - zl1446
Osvaldo Bulos - obr214

After the scraper has been built and tested, the wikipedia crawling will be performed. At this point, a good data infrastructure will be needed to store in the cleanest way possible, to facilitate their manipulation for the next steps

The implementation of the dashboard will be done over HTML using different visualization tools such as Javascript, Google Maps, Google Charts and D3.

In the predictive model part, we could apply many technology such as machine learning,deep learning and graphical model to predict the major events' global effects.

Finally, we will incorporate our predictive model to our system.

Time Line

Week	Tasks and Deliverables
October 3rd - October 9th	Definition of the problem and pre-research. Selection of strategy and requirements.
October 10th - October 16th	Design crawling system. Build prototype for crawling system
October 17th - October 23th	Team Update Research on predictive model First Tests for Crawling system.
October 24th - October 30th	First Version of Crawling System.
October 31st - November 6th	Coding scripts for data cleaning.
November 7th - November 13th	Second Version of Crawling System. Data collection.
November 14th - November 20th	Predictive Model Testings and Model Selection.
November 21st - November 27th	Incorporate our predictive model to our system
November 28th - November 6th	Third version of our system
December 5th - December 11th	Testing and bug fixing.
December 12th - December 18th	Final Presentation. Final Version of functional Dashboard. Final Version of predictive model.