Final Assignment - Social Web

Arthur-Ervin Avramiea
2517642
a.e.avramiea@student.vu.nl

Mihnea Dobrescu-Balaur
2549278
mihnea@linux.com

Zilvinas Kucinskas
2547940
zil.kucinskas@gmail.com

1. INTRODUCTION

With the rise in popularity of the Web, everybody around the world produces content, from local bloggers to global news agencies. This means that the online medium is now full of content, and that is great. The Web embodies diversity and pluralism in opinions, making sure that anybody can find useful content. However, with this much content available, it can be difficult for people to keep track of the items that interest them the most. And because of the decentralised nature of the Web, it can be difficult to find relevant articles.

In order to solve these problems, websites started providing RSS[1] feeds and more advanced users have started following them. There are even online aggregators, like the recently closed Google Reader[2], in which an user could add multiple RSS feeds that he or she is interested in, and then the website will keep track of new articles and display them in managed lists.

The problem with RSS and aggregators like Google Reader is that they are not so user friendly, having a dense interface, similar to a full webmail inbox. Also, they display the articles in a plain way, with lots of text. This all adds up to an overload of information to the user. More recent applications, like Feedly[3] and Flipboard[4], have taken a fresh approach of the problem. They create rich, magazine-like layouts for the articles, and they also help with content discovery, having predefined and curated âĂIJfeedsâĂİ that the users can subscribe to.

[1]http://en.wikipedia.org/wiki/RSS
[2]http://www.google.com/reader/about/
[3]http://feedly.com/
[4]https://flipboard.com/

We believe that we can take this idea a step further, and create a rich visual timeline, in which the media content (images and videos) take precedence over the text, removing everything but the headline. This timeline helps the users quickly get up to speed with the latest events and news from around the world, and when they find an article that is interesting, they can read it from the original source, with the original layout, in just one click.

Since there is no such thing as one size fits all, our application allows users to select what regions they are interested in, as well as what domains. So, for example, one user might be interested in Politics around the US, while another user might be interested in Tech news from Asia. Personalising a timeline is easy, and a user can store multiple timelines on his or her account.

2. STRUCTURED DATA

Our application has to include articles from all over the world, on all possible topics, and to cover this demand we decided to use, first and foremost, the RSS feeds of the main news agencies in the world. Wikipedia provides a list[5] of them, grouped by country. Besides news agencies, we also include information from prominent newsletters like the New York Times[6] and online publications like The Verge[7].

Besides XML data (via RSS), we also use the Twitter API in order to get JSON data of the tweets related to any given article.

To enrich the userâĂŹs visual experience, besides the media from the original article, we use the Bing search API to find relevant images and videos, that we then later embed in the rendered story.

All the mentioned data sources and others (detailed in the Analysis section) get mixed in a pipeline that builds a JSON object representing the visual summary of the story that we want to render for the user. Then, using

[5]http://en.wikipedia.org/wiki/List_of_news_agencies/
[6]http://www.nytimes.com/
[7]http://www.theverge.com

Web technologies we fetch the corresponding JSON files in the frontend application and render them, building the timeline.

3. DATA ANALYSIS

In order to give the user an approximation about the impact of the story they are skimming, we use Sentiment140[8] to perform sentiment analysis on the tweets that we found for that story. Since location is important for our application and for our users, we cluster the sentiment results by region. Figuring out where do the tweet authors live exactly is not trivial, since the majority of tweets does not contain location information. To solve this problem, we rely on the information that users share on their profile page. However, since that data is not structured at all, we have to reason about its text value and decide what region it represents. We do this using the Google Geocoding API[9].

4. ZILVINAS KUCINSKAS INDIVIDUAL PART - SENTIMENT ANALYSIS AND VISUALIZATION

4.1 Rationale

Our application not only provides information about hot topics, but also analyzes Twitter posts with each topic. It uses Twitter API to get tweets to each topic. It analyses those tweets using Sentiment140[10] analysis. Some other services use simple keyword based approach to analyse the tweets, but this one uses classifiers built from machine learning algorithms. It can provide both the aggregated sentiments or assesment of an individual tweet. Aggregated information is provided directly on the site using pie diagram, and it is possible to query the service to get data in JSON format. Our application splits tweets by country and provides visualisation using jVectorMap[11]. This API provides capabilities to generate custom maps or use existing ones. We use world map to provide visualisation of sentiments.

The main rationale behind this feature is providing users with approximation of the impact article gives user's globally. There is three variants of different opinions - positive, negative and neutral. Everytime application is refreshed - new tweets are gathered for each separate article and sentiment analysis is provided.

According to Wikipedia[12], the main purpose of data visualization is to communicate information. Based on article "**Why is data visualization so hot?**" [13], humans are able to interpret information way better if they see it visually, because huge amount of data can be transmitted to the brains through the optic nerve. Usually it's hard to understand the data looking only to numbers, but using data visualisation it is possible to see patterns and trends in the data way easier and faster [**?**].

In my opinion, every text human reads could provide some impact on him, he may feel angry, scared, astonished, excited, neutral and so on. Usually, the more followers user has, the more it is influential to others. Sentiment analysis and visualisation feature provides our application users the opportunity to compare their sentiments to the ones in the Twitter social network. By splitting sentiments by country, we provide them the opportunity to find some patterns in the sentiment world map. Because each country has separate traditions and culture, people from different countriens feel differently about same topics. For example, occupation in Crimea is a really hot topic right now. By analysing sentiments, we were able to see that Russia has a positive sentiment regarding Crimea and for example France has more negative sentiment about it. So by having this information it is possible to make assumptions that Russia wants to occupy Crimea, and russian media is supporting it and that media from EU, France is against Crimea occupation. This can be shown in the Appendix figure 5.

4.2 Motivation

Our application is oriented to people, who is interested in reading news articles, for users, who is keen to know what is happening around the world or in different domains of knowledge. By providing only pictures, videos and headline of an article, it is far more easier to search content to read. People can skim pages really fast and select which articles they want to read. They can also personalize their content by specifying which domain or regions he is interested in, also he has an opportunity to have several timelines. By clicking on specific article user gets redirected straight to the article in some online media source.

4.3 Scoping

Besides nice features our application also has some limitations. For example pictures and videos extracted from Bing could not always reflect the article, but we believe at least images and videos extracted directly from each online media source article is accurate. For example, there is a screenshot of an article in our application in this report Appendix figure 6, which demonstrates this limitation. Also sentiment analysis could not be completely accurate. For example, tweets containing sarcasm can be evaluated completely opposite from it's initial sentiment. And almost always more than 50 percent tweets are evaluated as neutral. This is also a limitation, because variance between positive and negative sentiments is not so obvious.

[8] http://www.sentiment140.com/
[9] https://developers.google.com/maps/documentation/geocoding/
[10] http://www.sentiment140.com/
[11] http://jvectormap.com/
[12] http://en.wikipedia.org/wiki/Data_visualization
[13] http://blog.visual.ly/why-is-data-visualization-so-hot/

4.4 Evaluation

How would you evaluate the success of your application (e.g. what metrics would you use, what would be an appropriate approach for evaluation and validation)?

4.5 Future work

How would you improve and further develop your application design if time and efforts would permit?

5. REFERENCES

6. APPENDIX

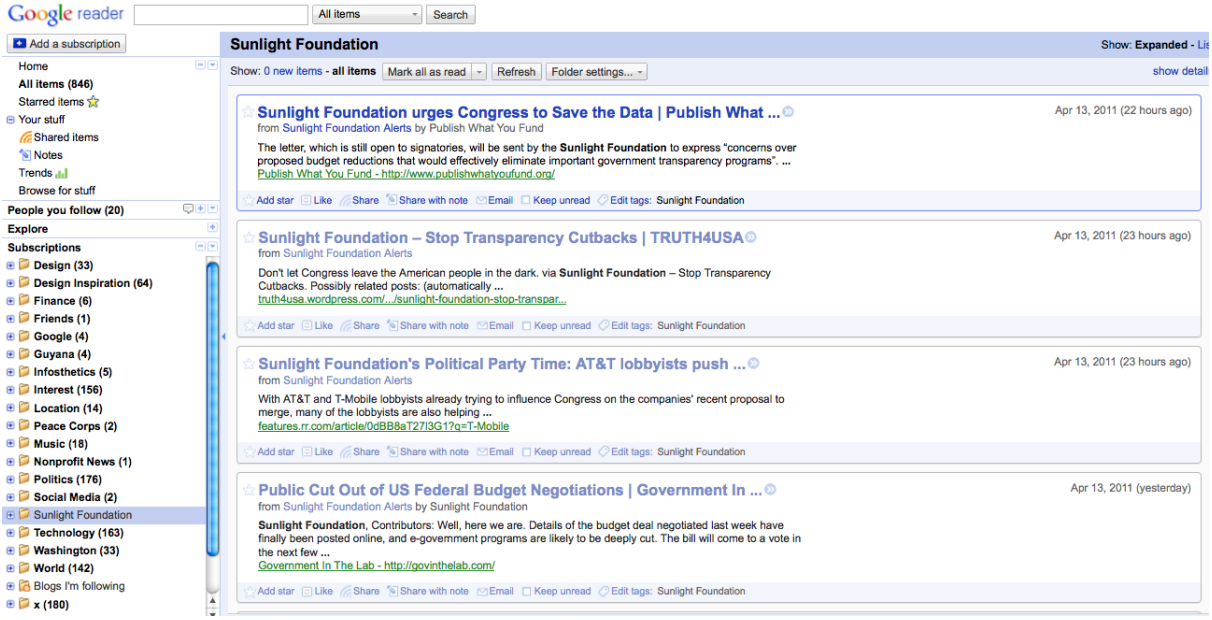

Figure 1: Google Reader screenshot

Figure 2: Flipboard screenshot

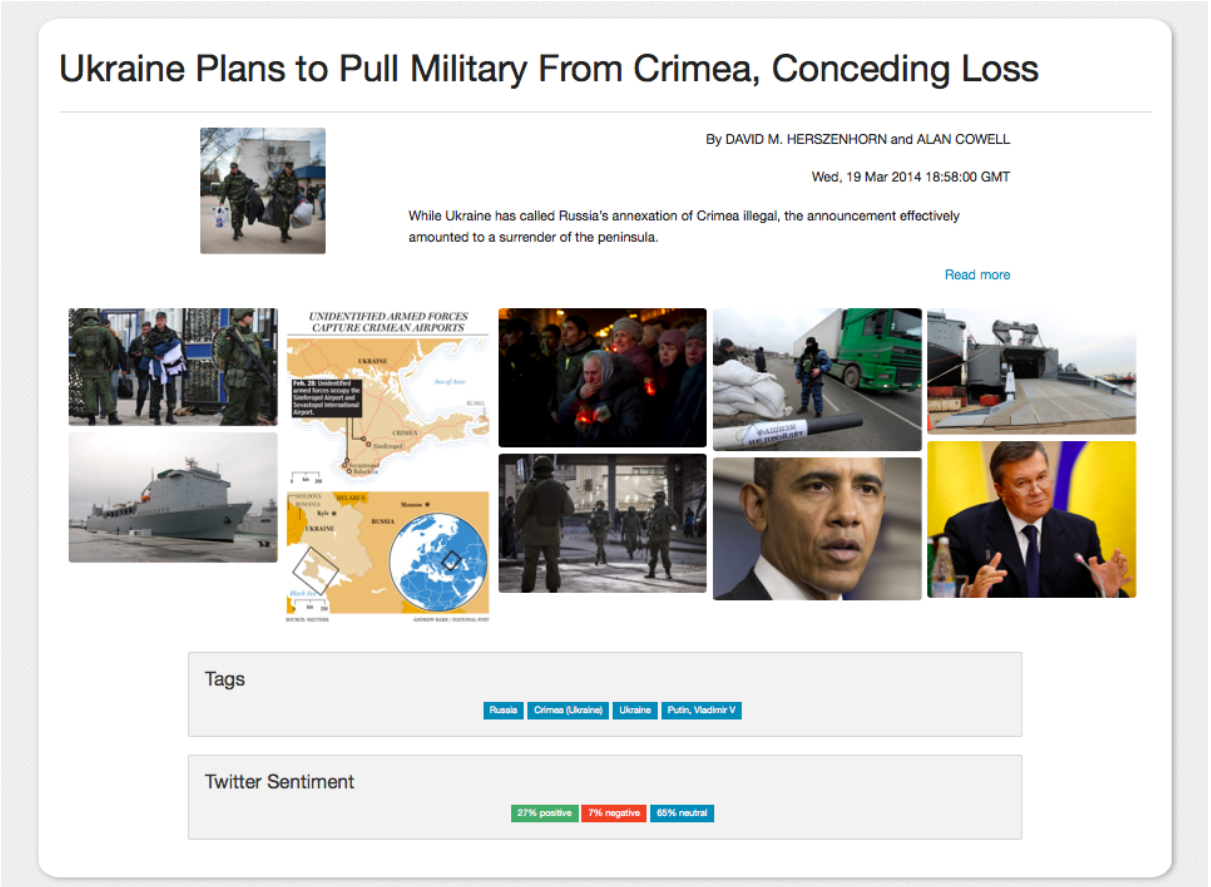

Figure 3: Timeline item in our application

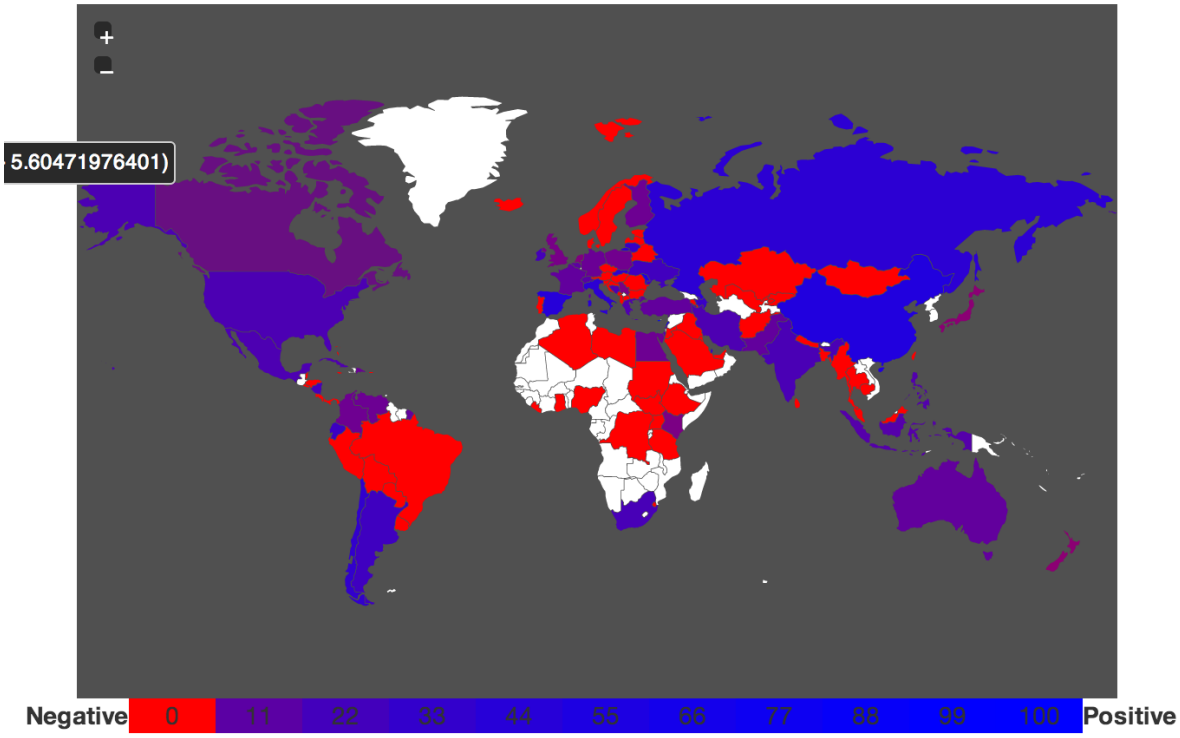

Figure 4: Media coverage visualization (polynomial)

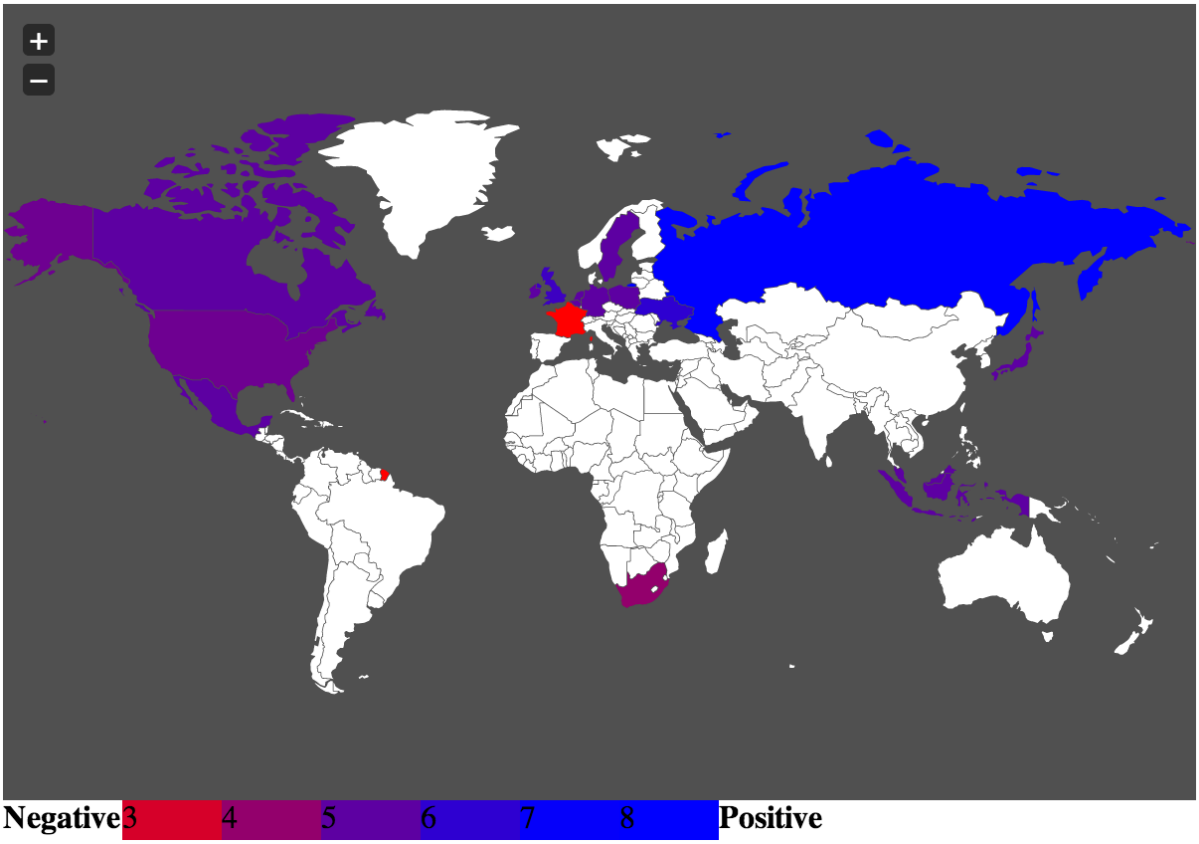

Figure 5: Sentiment analysis about crimea topics

Malaysia Turns to F.B.I. for Help in Plane Inquiry

By CHRIS BUCKLEY and MICHAEL S. SCHMIDT

Wed, 19 Mar 2014 14:45:27 GMT

Investigators were trying to recover data from a flight simulator custom-built by the pilot of the missing jet as relatives of the plane's passengers angrily criticized the Malaysian government.

Read more

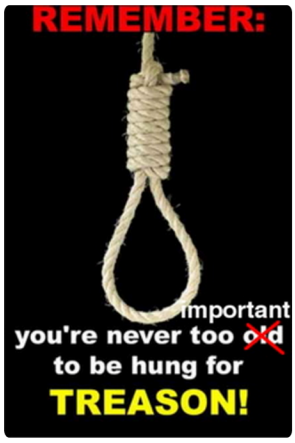

Figure 6: Screenshot demonstrating our worst scenario of extracting photos