PaperTrends - Online News Analysis

Denis Kuliček (4409612) & Stefano Imoscopi (4407245)

Motivation/Problem addressed

We wanted to apply and design InfoVis techniques on newspaper articles to spot trends and patterns in the data, like most used words, most talked about politicians and topics, syntactic relationships between words. The goal was to provide interactive visualizations to efficiently communicate results.

Dataset

We collected a set of news articles from the guardian website using Heritrix web-crawler[1]. The topic of the articles was the political elections in the UK, focusing on the two major parties, the conservatives and the liberal democrats. We collected articles between February 2010 (the year of the elections) and January 2015.

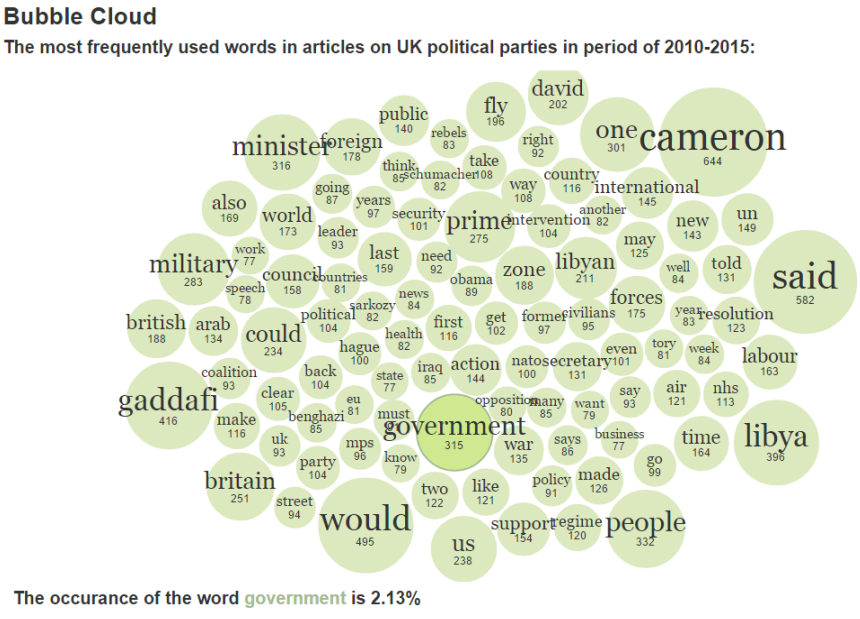
Another event we wanted to analyze was the recent attack to the Charlie Hebdo offices and the subsequent terror spree in Paris. For this we collected news articles from Time and BBC websites. Moreover we wanted to investigate the reaction of people, so we crawled comments from the popular social community and news website Reddit about the Paris attack.

We extracted word counts and other features to create labelled .csv datasets to help us create the visualizations using d3. We focused not only on the key words and the number of appearances, but also on the date of publishing, so to make a time analysis.

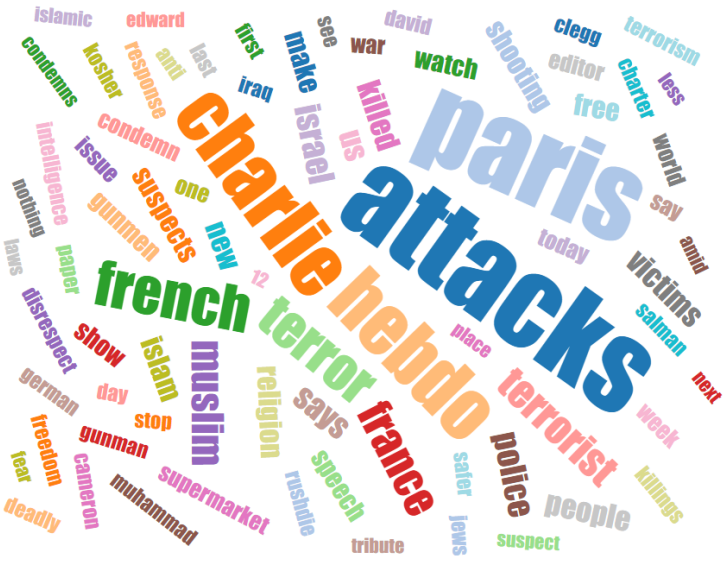
Visualizations

We decided to create four visualizations to highlight different properties of the data and let the user explore the article. We used javascript with the d3 library to create the four visualizations. They can be grouped in three groups according to the main purpose:

1. **Word clouds** and **Bubble clouds**. This two visualizations aim at highlighting the most frequent words and help the user spot important words or topics that emerge from the articles and text data we collected.



For the bubble cloud the occurrence of keywords is encoded in the area of the bubble. The keyword and number of appearances is written inside the bubble and selecting a keyword the percentage is shown. We visualize a bubble cloud from the articles taken in a single month about the two UK political leaders: James Cameron and Nick Clegg. The user can select which month to inspect and which political leader. It is clear that Cameron got more media coverage on the guardian, especially after the election he won but even before, namely from February 2010 to May 2010. Some interesting keywords can be spotted, like "*reform*", "*India*", "*Pakistan*", dealing both with internal and foreign policy.

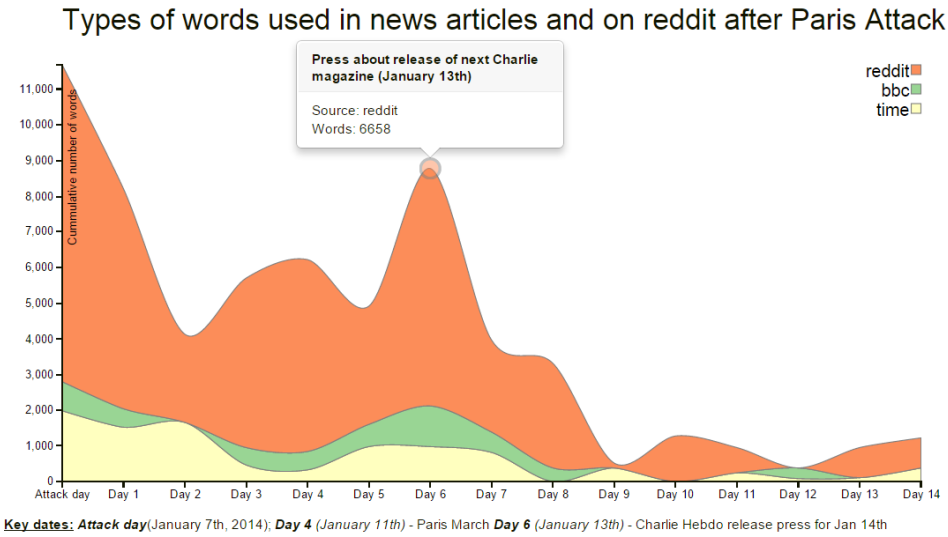


For word clouds the occurrence of the word is encoded in the size of the font. The use of different categorical colours for every word helps the user spot the single words and makes the visualization nicer and less tiring to watch. For example the obvious top keywords emerging from the titles of the Paris attacks articles are: "*attacks*, ""*charlie*", *hebdo*","*paris*", but also less obvious ones like "*terror*", "*muslim*", "*Israel*".

2. **Word tree[2]**. This visualization lets the user explore the complete text contained in an article to analyze relationships between words, most frequently occurring couples of words or expressions.

It's an interactive visualization in which the user can select the root of the tree (shift+click) and inspect the text at different levels of detail. It is a useful tool to analyze the writing style of a journalist or the style of a speech transcript.

3. **Stacked area plot over time**. We wanted to see how media coverage and people's attention evolved over time during and after the attacks in Paris. This visualization plots over time the number of words spent on the argument on articles and on reddit in comments from people. We stacked the 3 sources in the plot so the user can easily compare the amounts and see all the evolution over time of words spent on the subject. Moreover the user can filter only some particular words relating to religion, violence or freedom of speech. The time step is 1 day. Even though the data is discrete time we decided to use a continuous function to interpolate between the days cause it gives a better idea of the trend and emphasises the local maxima. We also decided not to scale the vertical axis automatically so to emphasize the different amount of words present jumping from the "*religious*" keywords to the ones about "*freedom of speech*" and "*violence*". The scaling of the y-axis can be performed clicking on the dedicated button.



Key dates: ***Attack day***(January 7th, 2014); ***Day 4*** *(January 11th)* - Paris March ***Day 6*** *(January 13th)* - Charlie Hebdo release press for Jan 14th

The amount of data generated by the people commenting on Reddit is much larger than the data from professional articles. Peaks in the interest about the Paris attacks are clearly visible, in particular the one on the 13th of January, before the release of the new Charlie Hebdo magazine and the one on day 2 during the attacks to the kosher supermarket. It is also interesting to note that the data generated from time and bbc follows almost identical distribution, while people commenting has different "peaks and valleys", with a second big peak on day 4, the day of the march in Paris for freedom of speech. Of course the global maxima is on the 7th January, the day of the attack, and after 9 days from the attack the amount of data has decreased to only 10% compared to the beginning and stays at that level in the following days.
Another interesting thing to see is that different filtering of keywords have different distributions. The order of predominance is words about "*religion*", "*violence*" and last "*freedom of speech*". It's interesting to note that the religious nature of the attack caught the attention, especially on the Reddit community.

Conclusions and future work

We are quite satisfied with the statistics and trends we could find in the text data we collected using these visualizations. We had fun working on the project and designing and tweaking the visualizations the get the best out of our data. We tried to keep a balance between usefulness and "looks". For example the area plot is not very pleasing to the eye but gives good insight, while the word cloud doesn't give so much information

and is very similar to the bubble chart, but it is very nice to see and can be a good addition to improve the looks of a web-page.
Of course this analysis of handwritten text is still rather basic and extracting more features from the data could allow us to design and explore more complex and interactive visualizations to spot more trends. The d3js[3] website is a good source of inspiration for that. We would like to visualize some bar-charts of letter distribution in different languages after collecting numerous articles in different languages.

References

[1] Heritrix web-crawler https://webarchive.jira.com/wiki/display/Heritrix/Heritrix

[2]Wattenberg, Viégas, “The Word Tree, an Interactive Visual Concordance”, IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 14, NO. 6, NOVEMBER/DECEMBER 2008 https://www.cg.tuwien.ac.at/courses/InfoVis/papers/InfoVis2008%20The%20Word%20Tree,%20an%20Interactive%20Visual%20Concordance%20[Wattenberg].pdf

Visualizations by Jason Davies http://www.jasondavies.com/

Jarke J. van Wijk, “The Value of Visualization”, Visualization, 2005. VIS 05. IEEE, 23-28 Oct. 2005, pg. 79-86 https://www.cs.ubc.ca/~tmm/courses/cpsc533c-05-fall/readings/vov.pdf

d3 js library documentation http://d3js.org/

TU Delft Data Visualization slides on InfoVis