Weekly Progress Report

Detecting Conversations Going Viral

Matt Chapman

Week 14, 03/04 - 07/04

1 Dataset Selection

1.1 Selection & Rationale

Selecting the correct dataset to carry out my analyses is one of the most important parts of the work I will carry out in this project. To this end, I will explain the data set that I will be working with, as well as my rationale for selecting it.

I have chosen to work exclusively with Dutch language Twitter data for the following reasons:

- The vast majority of Buzzcapture clients are Dutch companies, so it is more likely that I will be able to find relevant data by restricting gathered conversations by language.
- Limiting my queries to the Buzzcapture platform by language will provide a much less 'noisy' dataset to experiment upon.
- Twitter is *the* platform for breaking news and discussions thereof. It is likely that changes in volume that the likes of which I am looking for, will happen on Twitter before they happen on other platforms such as Facebook, Pinterest or Facebook.
- Conversation volume is much higher on Twitter than on Facebook or other competing services. As such spikes in conversation volume are likely to be higher and more easily annotated by hand for both testing and demonstrating the effectiveness of a given algorithm.

Queries will be made against the Buzzcapture Elasticsearch instance to provide both historical and real time data for conversations concerning clients that have had (recently or otherwise) considerable spikes in conversation volume.

1.2 Alternatives

The sheer amount of data collected by Buzzcapture provides a number of alternatives. Buzzcapture collect data from traditional print and radio media as well as online media platforms such as Instagram, Pinterest, Facebook and Twitter.

Data from any of these could be used with change detection algorithms to locate spikes in conversation volume, as they are all high-throughput platforms. Facebook in particular (being the largest social network in the world) would be a good source of data through sheer volume of users alone, but Twitter has the edge in 'reaction time' when it comes to events occurring in the world.

Other platforms such as Pinterest or Instagram simply don't have the volume of either users or postings to provide useful data for testing against.

1.3 Data Collection & Transformation

As the Buzzcapture back-end codebase is primarily Python, I will also be working in this language. Fortunately, Python has a number of effective libraries for data science.

Buzzcapture will provide me with scripts to query the Elasticsearch engine and provide CSV files of data that I can use. I will then use Pandas[1] to parse the data and generate some plots for reference materials.

Once the data has been obtained and plotted to ensure that it is correct and contains the correct spikes and annotations required for my analysis, I will carry out the operations necessary (also using Pandas) to make the dataset suitable for use with the change detection algorithms I will be implementing.

I am not yet sure of the format that the data export will take - so I cannot commit to a certain approach in transforming it at this point. Further work on this will be discussed in the week 15 report.

I will carry out this work using Jupyter Notebooks, to allow me to annotate and demonstrate my approaches to my project supervisor.

2 Additional Research

In addition to making some decisions regarding the experimental dataset, I have also carried out some additional research into the field. Below are a few papers that I have found. I will continue collecting papers during reading sessions each week. Papers are being stored in a Mendeley library, to which access can be provided if requested.

[3] **Anita M Pelecanos, Peter a Ryan, and Michelle L Gatton. "Outbreak detection algorithms for seasonal disease data: a case study using Ross River virus disease." In: *BMC medical informatics and decision making* 10.1 (2010), p. 74. ISSN: 1472-6947. DOI: 10.1186/1472-6947-10-74. URL: http://www.biomedcentral.com/1472-6947/10/74**

[1] http://pandas.pydata.org/

[2] **Martin Kulldorff et al. "A space-time permutation scan statistic for disease outbreak detection". In: *PLoS Medicine* 2.3 (2005), pp. 0216–0224. ISSN: 15491277. DOI: 10.1371/journal.pmed.0020059**

[1] **Jeremy Ginsberg et al. "Detecting influenza epidemics using search engine query data." In: *Nature* 457.7232 (2009), pp. 1012–4. ISSN: 1476-4687. DOI: 10.1038/nature07634. URL: http://www.ncbi.nlm.nih.gov/pubmed/19020500**

References

[1] Jeremy Ginsberg et al. "Detecting influenza epidemics using search engine query data." In: *Nature* 457.7232 (2009), pp. 1012–4. ISSN: 1476-4687. DOI: 10.1038/nature07634. URL: http://www.ncbi.nlm.nih.gov/pubmed/19020500.

[2] Martin Kulldorff et al. "A space-time permutation scan statistic for disease outbreak detection". In: *PLoS Medicine* 2.3 (2005), pp. 0216–0224. ISSN: 15491277. DOI: 10.1371/journal.pmed.0020059.

[3] Anita M Pelecanos, Peter a Ryan, and Michelle L Gatton. "Outbreak detection algorithms for seasonal disease data: a case study using Ross River virus disease." In: *BMC medical informatics and decision making* 10.1 (2010), p. 74. ISSN: 1472-6947. DOI: 10.1186/1472-6947-10-74. URL: http://www.biomedcentral.com/1472-6947/10/74.