Network Tour of Data Science

Khalil Mrini, Khalid Omari, Ana Stanojevic

November 28, 2017

1 Project Description

In the 2011 Arab Spring, the two most significant revolutions happened in Tunisia and Egypt. Seven years later, we aim to revisit the most recurrent demands expressed by the people on social media, as well as how the news media covered the events. We want to show the topics most discussed on the web, and the links between topics, people, events and places. Through this project, we aim to give a new graph-oriented insight into the revolutions.

2 Data Acquisition

We want to use the Spinn3r dataset that corresponds to the 2011 ICWSM Data Challenge. According to the online description, the dataset extends from January 13th to February 14th 2011, roughly covering the time period in between the Tunisian presidential resignation and the Egyptian one. The data size is gigantic (3TB) and we expect to use it from the cluster. The sources include social media websites such as Facebook and Reddit, and news media such as the BBC and the New York Times.

The dataset was the subject of a Data Challenge and we read the only paper that participated in that challenge [PAM+11]. We found that there is more room to improve on their analysis, as they did not tackle the popular demands, the non-English-language data, and sentiment analysis did not bring useful insights.

3 Data Exploration

The data includes a variety elements from the web, and we will need to filter it to relevant posts. That means we will have to keep posts related to the Tunisian and Egyptian revolutions. To do so, we will scrape the corresponding Wikipedia articles for keyphrases. In addition to English, we plan to use Arabic- and French-language data.

After this filtering, we want to use Text Mining techniques, such as LDA and TF-IDF, to model topics as nodes with edge weights representing their co-occurrence. We also want to use Named-Entity Recognition to recognise organisations, locations and names and model them as nodes.

4 Data Exploitation

There are many useful ways to exploit described data.

Network Science: It would be interesting to understand and model news spreading through this network, with a focus on viral news.

Spectral Graph Theory: Using eigen-decomposition we can find clusters (communities), consisting of topics, names, etc. and possibly visualize them.

Graph Signal Processing: We can use techniques from this field to remove noise from the graph (denoising).

Disclaimer

The dataset is provided by the Applied Data Analysis (ADA) course in an IC cluster. Khalid and Khalil are in the ADA course and work on this dataset, but there will be no common work, as the ADA course will focus on NLP and Data Analysis techniques, whereas this one will focus on Graph and Network Science.

References

[PAM+11] Jaehyuk Park, Beunguk Ahn, Rohjoon Myung, Kyuree Lim, Wonjae Lee, and Meeyoung Cha. Revolution 2.0 in tunisia and egypt: Reactions and sentiments in the online world. In *Proceedings of the fifth international AAAI conference on weblogs and social media*, volume 1, 2011.