HACL Project Status Week 3

COVID-Twitter

Rufeng Ma

26 July 2020

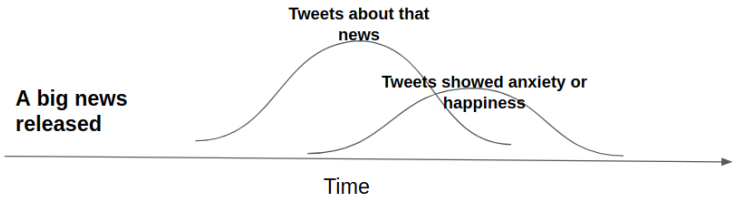
Contents

Weekly Work Summary

- Summary of work since last week **Team scope:** We clarified the purpose of the whole project during the meeting last week. For COVID-Twitter project, we could have multiple aims but this project should be a team project. As a team each person will contribute to different parts. We initially defined short term and long term aims. Short term aim: to get the sentiment analysis about masks first. Long term aim: do some time series analysis. We would like to find the correlation between some policy announcements and the twitter users' attitude. **Personal scope:** I had the brainstorm to determine my project "A time series analysis about twitter users' mental health during COVID-19". However, as a new NLP learner this time series analysis is not the one we could immediately get done in 6 weeks. I decided to help Albraham to solve the open issues on Github. In this way, I think I can have a deep understanding of his code. The backend engineering is the part I need to catch up. I think my advantage is that I have creative ideas. I know the process to get things done. However, the knowledge of R coding, NLP and data analysis is the obstacle I have. So I did things in the following 'Personal contribution' part.
- Summary of github commits
  - include branch name :hacl-mar6
  - include files: silhouette_mar6 (not push to the master yet, local folder)
- List of presentations, papers, or other outputs (with links) Idea about the "time series". https://docs.google.com/document/d/18Zwb1pSSit9663FUn1j4NZRxEDkgQYvkOd3MzRAmmdg/edit?usp=sharing
- List of references (if necessary) Blog about NLP and examples: https://sanjayasubedi.com.np/nlp/nlp-intro/ Paper: predicting the political alignment of twitter users https://cnets.indiana.edu/wp-content/uploads/conover_prediction_socialcom_pdfexpress_ok_version.pdf Blog about twitter sentiment analysis: https://medium.com/@r.ratan/tweepy-textblob-and-sentiment-analysis-python-47cc613a4e51
- List of location(s) of all work submitted to github
- Indicate use of group shared code base /home/mar6/COVID-Twitter/analysis/covid-twitter-hacl-template.Rmd. /home/mar6/COVID-Twitter/analysis/Elasticsearch.R
- Indicate which parts were done by you or as part of joint efforts The new method implementation is done by me.

Personal Contribution

- Clearly defined, unique contribution done by you (code, ideas, writing) **-Reading:** Blog about NLP and examples (link in references) Paper: predicting the political alignment of twitter users (link in references) Blog about twitter sentiment analysis (link in references) **-Thinking (Writing):** Idea about the **time series**. Refer to the 'List of presentations' part.



-Coding: Text clustering using Python and the bbc dataset. For choosing the optimal k value, I tried both the elbow method and the silhouette score method. Text clustering using R and covid-twitter dataset. Implementing the **silhouette score** method to replace the **elbow method**.

Discussion of Primary Findings

- Discuss primary findings:
  - What did you want to know? I am wondering if the silhouette score is a good way to replace the elbow method.
  - How did you go about finding it? I implemented the silhouette score method. I chose k from 2 to 10 based on my experience when I changed the k. Then I give a list of the silhouette score. Just choose the highest score to redo the clustering. Then I compared it with the original cluster result. It looks much better.
  - What did you find? I found this method works well in Python. I decided to add it to Abraham's code with R in his Rmarkdown file.
- Provide illustrating figures and or tables

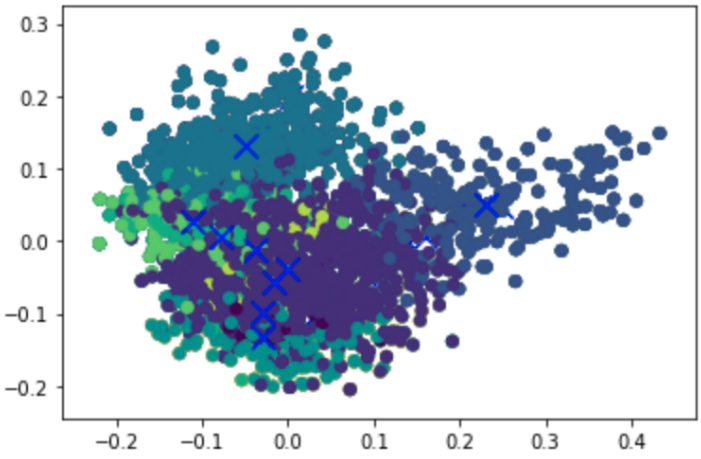


Figure 1: score

- Make sure any source code for your figures and tables are embedded in notebook or provide github location

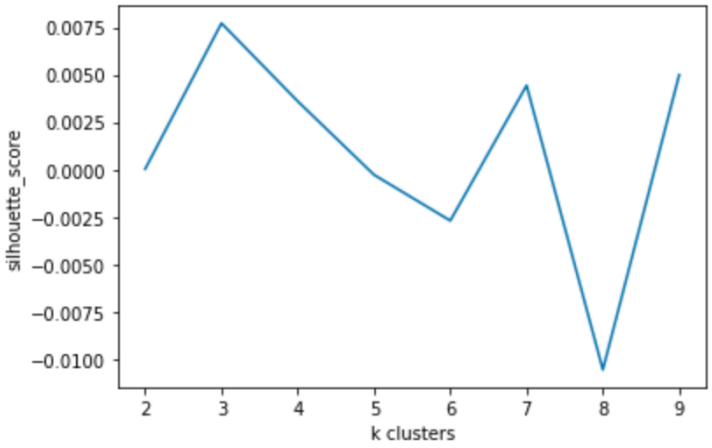


Figure 2: silhouette score

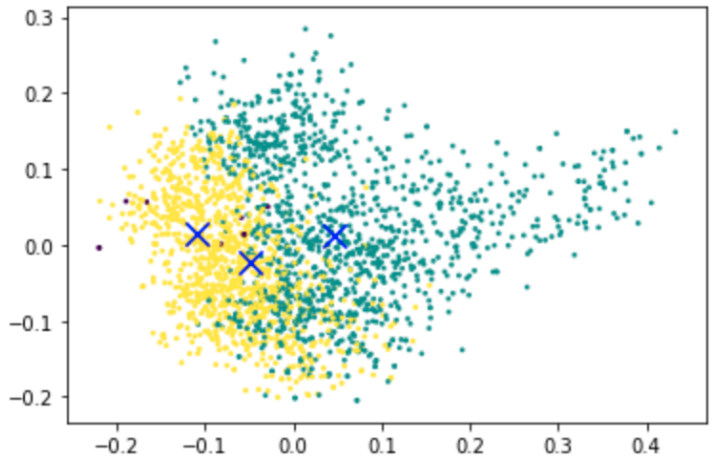


Figure 3: score