CSE 435/535: INFORMATION RETRIEVAL

PROJECT 4: Dissecting Twitter data to analyze government & public attitude towards Covid and vaccines

Final Deadline: 10th Dec, 11:59 PM ET

University at Buffalo
School of Engineering and Applied Sciences

Overview of previous projects

- The first 3 projects dealt with:
  - Project 1: Indexing and Crawling
    - How do you gather data from a particular POI? How do you retrieve the reaction?
    - How do you effectively index this data using Solr?
  - Project 2: Scoring
    - How does query scoring work?
  - Project 3: Relevance
    - How do you tune relevance for specific information needs?
- Project 4 seeks to unify these subtasks into a single end-to-end IR system.

Dataset

- At the end of project 1, you had at least 50K tweets
- 500 tweets/POI for 5 POIs/country, where country being USA, India and Mexico
- The language of the tweets also ranges in these country specific languages (English, Hindi and Spanish)
- At least 1 reply to a minimum of 1,500 Covid vaccine related tweets.
- At least 10 replies to a minimum of 300 Covid-19 related POI tweets.
- Thus, you have a good enough dataset for a multi-lingual IR system.

Project Goal

Basic Requirements

1. Content Analysis.
   a. Analyze the attitude of the general population towards Covid vaccines.
   b. Analyze the impact of Covid related political rhetoric on the common masses.

2. Build a search engine and analytic web UI to present useful insights.
   a. Enhance knowledge of building end-to-end IR system by implementing a search engine
   b. Develop a web UI to present your content analysis

Bonus Requirements

1. Identifying excerpts of Covid vaccine related disinformation or vaccine hesitancy.
2. Identifying excerpts of persuasion against Covid vaccine related disinformation or vaccine hesitancy.

Groups and Dataset Sharing

- You need to form your own groups of 3-4 members.
- Sign-up your team using the Google Form posted on Piazza before 15th Nov, 5 PM
- You are allowed to share your data within the group.
- You are free to collect more data.

Content/Topic Analysis

- Compare number of Covid and non Covid related tweets made by the POIs of each country and correlate the Covid curve in that country with it
  - *Is there any correlation between what POIs are tweeting and the COVID curve in the country?*
- Perform sentiment analysis, topic analysis, stance detection .etc. on the Covid vaccine tweets to gauge the attitude or stance of people towards the vaccines.
- Perform sentiment analysis on the reply to Covid related POI tweets to gauge the impact of the tweet.
- Determine excerpts of vaccine hesitancy, vaccine related disinformation, persuasion for or against taking the vaccines .etc.
- Be creative and come up with more use cases!

Visualizations: Insights/Analytics

- Main purpose is to visualize the insights from the last step, into a meaningful story.
- You can do additional processing such as location analysis, keyword analysis, topic modelling, etc.
- You can ingest additional data such as news articles, youtube videos.
  - Eg: extract news articles which talk about any incidents that could be related to the POI's tweets on COVID, or incidents talking about declining vaccination rates.
- Decide on appropriate visualizations (charts, graphs, maps)

Search Engine: Faceted Search

- Create a webpage to perform search operations on your indexed data
- Ideally, left side of the web page should render faceted search functionality. There should also be a search bar at the top of the page, like Google search, where you can search your dataset based on keyword.
- You may also implement network analysis to rank the retrieved documents.
- You are encouraged to implement more search functionality and demo various interesting search results.

Final Deliverables

- A short demo video (at most 3 minutes)
- A working web application URL hosted on AWS
- A short report detailing all work done and member contributions.
  - You can use the double column ACL-IJCNLP 2021 or single column ICLR 2021 Latex template.
  - You can also use word, if you are not comfortable with Latex.
  - The report should contain the following broad sections: (i) Introduction, (ii) Methodology (iii) Sample screenshots (iv) Work breakdown by teammates (v) Conclusion
  - More details on how to submit will be shared closer to the deadline.

Grading

- Grading is based on relevancy, language spread of served results, ranking techniques and impact measures.
- Points distribution:
  - Meet basic requirements – **7 points**
  - Meet bonus requirements – **3 points**
  - Visualizations and storytelling via UI – **10 points**
  - Search Engine – **7 points**
  - Report – **3 points**

- We will select best performing groups to present their work in the class
  - 7 groups will be selected to present their work in 8 minutes with additional 2 minutes for Q&A
  - Each team member of the selected groups will receive 2 bonus points.
  - More details to be shared later

Project Summary

- The project is fairly open-ended and permits usage of any third party tools that you deem relevant
- Primary objective is to encourage students to apply IR concepts in detecting and analyzing influence of Twitter personalities in the social sphere.
- Wide latitude in evaluating your projects
  - UI, algorithms, research – several areas to innovate upon
- Don’t be afraid to be creative and stand out!

Timeline

1. 10th November: Project released.
2. 15th November: Deadline for team formation.
3. 6th December, before 5 PM: Interested groups submit videos for class presentations. Sign-up sheet will be released 3 days before.
4. 8th December: In-class presentation for selected groups (2 bonus points).
5. 10th December: Final submissions due.

Demo

- Sample demo: https://youtu.be/GoXhy6SKhxg

Resources

- Machine learning / clustering / topic modelling:
  - Python : Scikit-learn, nltk (NLP specific)
  - Java : Spark/Mahout, Weka, Mallet
  - C++ : Shogun, mlpack
- Word embeddings (pre-trained)
  - http://nlp.stanford.edu/projects/glove/
  - Pointers to download links: https://www.quora.com/Where-can-I-find-some-pre-trained-word-vectors-for-natural-language-processing-understanding
- Translation : Google and Bing APIs, several free to download dictionaries

Resources

- Mutlifaceted API libraries:
  - Microsoft Cognitive Services API : https://azure.microsoft.com/en-us/services/cognitive-services/
  - Google Cloud Natural Language API : https://cloud.google.com/natural-language/
- Sentiment Analysis:
  - NCSU tweet sentiment visualization app: https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/tweet_app/
  - Textbox: https://machinebox.io/docs/textbox?utm_source=medium&utm_medium=post&utm_campaign=fakenewspost

Resources

- Visualization / analytics examples and ideas
  - http://www.tableau.com/stories/gallery
  - https://www.census.gov/dataviz/
  - https://app.powerbi.com/visuals/
  - https://github.com/d3/d3/wiki/Gallery
  - https://developers.google.com/chart/interactive/docs/gallery
  - https://developers.google.com/chart/interactive/docs/more_charts