CSE 435/535: INFORMATION RETRIEVAL

PROJECT 4: DISSECTING TWITTER DATA TO ANALYZE GOVERNMENT AND PUBLIC ATTITUDE TOWARDS COVID GOVERNANCE

Deadline: 11th 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 40K tweets
- 500 tweets/POI for 3 POIs/country, where country being USA, India and Italy
- The language of the tweets also ranges in these country specific languages (English, Hindi and Italian)
- Tweets posted in 5 consecutive days focused on reactions of general public to government’s policies on COVID.
- Thus, you have the dataset good enough to create a multi-lingual IR system.

Project Goal

- To build a solution to analyze the government and public attitude towards COVID governance.
- To analyze influencer score and use it to improve tweets ranking
- To enhance knowledge of building end-to-end IR system including relevancy model and analytics.
- To build a search engine and analytic web UI to present useful insights

Groups and Dataset Sharing

- You need to form your own groups of 2-3 members.
- Sign-up your team on the Google Form posted on Piazza before 20th Nov, 5 PM
- You are allowed to share your data within the group.
  - Based on Project 1, each student should have at least 40K tweets. So, the total dataset size among each group would be 80K – 120K tweets
- You are free to collect more data.

Requirement 1 – Social Network Analysis

- **Calculate Influencer Score:** The idea is to generate scores for each tweet, based on their potential of influence. This can be achieved in a direct and an indirect way:
  - **Direct approach:** Use the number of retweets or likes for each tweet as a proxy for the unnormalised influence score.
  - **Indirect approach:** You can collect the number of followers that a person has, and assign equal weights to all his/her tweets, as a proxy for the influence score.

- **Suggestions on using Influencer Score**
  - You can weigh your KPIs based on the normalised influence score.
  - You can create a social network amongst all the actors (people) that you have in your dataset, treat the influencer score as the starting score, and calculate page rank score for each actor.
  - You can use the normalised influence score or page rank score to reorder the documents that are retrieved by your search engine in the UI.

Requirement 2 – 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?

- For each country, perform topic analysis all the tweets to extract main topics people are concerned about

- Use your own creativity to come up with more high-level analyses

Requirement 3 – Insights/Analytics

- Main purpose is to show insights based on the outputs of requirements 1 and 2
- You can do additional processing such as sentiment analysis, location analysis, keyword analysis, 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.
- Decide on appropriate visualizations (charts, graphs, maps)

Requirement 4 – 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.
- In order to show facets, you may need to do named entity tagging or topic generation
- You may also implement ranking based on Influencer Score
- You are encouraged to implement more search-based functionality and demo various interesting searches

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.

Grading

- Grading is based on relevancy, language spread of served results, ranking techniques and impact measures.
- Points distribution:
  - Requirement 1 – **5 points**
  - Requirement 2 – **7 points**
  - Requirement 3 – **10 points**
  - Requirement 4 – **5 points**
  - Report – **3 points**
- We also plan to select best performing groups to present their work in the class
  - 8 groups will be selected to present their work in 7 minutes with additional 2 minutes for Q&A
  - The selected groups will get bonus points
  - More details will be released 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

- 18th November: Project released
- 7th December, before 5 PM: Interested groups submit videos for class presentations
  - Sign-up sheet will be released 3 days before
- 9th December: In-class presentation for selected groups (at-most 2 bonus points)
- 11th December: Final submissions due

Demo

- https://youtu.be/GoXhy6SKhxg
- Note that this demo only involves Requirement (3) and (4). You are also required to show approaches to Requirements (1) and (2).

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