CSE 435/535: INFORMATION RETRIEVAL

PROJECT 4

ANALYZING THE IMPACT OF POLITICAL RHETORIC IN TRADITIONAL AND SOCIAL MEDIA

DUE: 5TH DECEMBER; 23:59

University at Buffalo The State University of New York

Overview of previous projects

- The first 3 projects dealt with:
  - **Project 1: Content ingestion and Indexing**
    - How do you gather data relevant to a topic, person? How do you retrieve the replies?
    - How do you effectively index this data using Solr?
  - **Project 2: Query Processing and Scoring**
    - How to effectively process queries and score documents?
  - **Project 3: Tuning Performance of IR Models**
    - How do you tune relevance for specific information needs?
- Project 4: Focused on problem solving resulting in an end-to-end IR solution

Dataset

- At the end of project 1, you had at least 1,000 tweets/POI for 15 POIs
- You selected POIs such that there were 5 POI/country, where country was one of 3 different countries (USA, India and Brazil)
- The language of the tweets also ranges in these country specific languages (English, Hindi and Portuguese)
- Tweets posted in 5 consecutive days have at least 20 replies each.
- Thus, you have a dataset good enough to develop the required solution

Project Goal

- **Framing a problem**: To analyze the impact of political rhetoric of influential actors by monitoring both social and traditional media
  - Impact by analyzing response to tweets in social media (sentiment, volume)
  - Impact by analyzing relevant news articles
    - Societal impact: events related to topics mentioned by actors
    - Does the rhetoric incite social unrest or violence?

- **Complete IR Solution**: Experience in building an end-to-end IR solution involving content ingestion, search, topic categorization, analytics and visualization
  - Focus on user experience

Groups and Dataset Sharing

- You need to form your own groups of 3-4 members.
- You are allowed to share your data within the group.
  - Based on Project 1, each student should have at least 33K tweets. So, the total dataset size among each group would be 99K – 132K tweets.
  - There is no minimum or maximum size requirements for the dataset. You need to have enough data to meaningfully draw insights and visualizations.
- You are free to collect more data.

Requirements

- Create an end-to-end IR based website (hosted on AWS)
- Your website should at a minimum include the following pages:
  - Analytics and Visualization: Include Charts and Map based visualizations that provide insights on the POIs' tweets and the impact of their rhetoric in their country.
  - Search: A separate web page allowing keyword search on your dataset
    - Faceted search

Analysis and Visualization - Topics

- For each country, perform topic analysis on POIs' tweets to extract main topics that POIs are talking about.
- It is your discretion whether you want to analyze each POI individually or merge all POI tweets together while detecting topics.
  - Eg: Tweets belonging to POIs from USA can be merged together for topic analysis.
- Make sure you have segregated the tweets by country so that you can show some interesting analysis on the impact of POIs in their respective country.

Analysis and Visualization - Impact

- **Social Media:** Different type of analyses, such as sentiment analysis, keyword analysis etc., on the responses/replies to the POI on each topic. You are free to show additional social media insights based on the engagements POI's tweets got for a particular topic. Use your creativity!

- **Societal Impact:** To measure if there was any impact of POI's tweets, you are required to extract mainstream news articles which talk about any incidents that could be related to the POI's tweets.
  - Local news articles are the preferred sources but you can use any other sources that provide details on the incidents.
  - Use any measure that you can come up with for analysis, such as number of relevant articles that talked about POI's tweets, map based analysis on the cities the articles were reported from, etc.
  - These articles should be multilingual based on the POI's country.

Search

- A webpage to perform search operations on your indexed data
- Ideally, left side of the web page should render a 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 keywords.
- You are encouraged to implement any further search-based functionalities that you would like.

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.

End Goal and Grading

- Your system should enable the user to get wide-range of knowledge and analysis of POI's rhetoric.
- Grading is based on relevancy, language spread of served results and impact measures.
- Points distribution:
  - Analysis and Visualization – 5 points
  - Search – 4 points
  - Report – 1 point
- 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 5-8 minutes slots
  - 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.
- The 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

- 6th November: Project released
- 2nd December, before 5 PM: Interested groups submit videos for class presentations
  - Sign-up sheet will be released 3 days before
- 4th December: In-class presentation for selected groups (at-most 2 bonus points)
- 5th December: Final submissions due

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

- Python Web frameworks :
  - Flask : https://www.fullstackpython.com/flask.html
  - Django : https://www.djangoproject.com/
- Java Web frameworks :
  - Spring MVC : https://docs.spring.io/spring/docs/current/spring-framework-reference/web.html
  - Google Web Toolkit : https://www.gwtproject.org/
- Java Script Web frameworks :
  - React.js : https://reactjs.org/
  - Angular : https://angular.io/

Resources

- News Articles Scraping:
  - beautifulsoup : https://github.com/waylan/beautifulsoup
  - News-Please: https://github.com/fhamborg/news-please

- Other useful data sources:
  - Reddit search API : https://www.reddit.com/dev/api/
  - Google search API : https://developers.google.com/custom-search/v1/overview
  - Common Crawl: https://commoncrawl.org/

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