CSE 435/535: INFORMATION RETRIEVAL

PROJECT 4: Complete Search and Analytics Solution based on dissecting twitter data

DEADLINE: DECEMBER 9, 2018; 23:59

Overview of previous projects

- The first 3 projects dealt with:
  - Project 1: Indexing and Crawling
    - How do you gather data on a particular topic?
    - How do you effectively index this data using Solr?
  - Project 2: Computing Scores
    - How does query scoring work?
  - Project 3: Ranking based on 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 were asked to continue data collection for next 3-4 weeks.
- The tweets were based on social issues such as social unrest, politics, environment, crime etc.
- Also, the geographical distribution of tweets was across 5 major cities in the world.
- The language of the tweets also ranges in these 5 city specific languages.
- Thus, you have the dataset good enough to create a multi-lingual IR system.

Project Goal

- To build a solution that provides insight related to social conversations on important societal issues
- To gain experience of building an end-to-end IR solution including data collection, search relevance, and analytics.

Requirements - IR

- Ingest tweets on
  - 5 topics: Environment, Politics, Crime, Social Unrest and Infrastructure
  - 5 cities: NYC, Delhi, Bangkok, Paris and Mexico City
  - 5 languages: English, Hindi, Spanish, French and Thai
- Detect trending phrases/hashtags from each topic/city.
- Retrieve top relevant tweets for each trending phrase/hashtag.

Requirements – Analytics and UI

- Perform analysis such as:
  - Time series – for a given city
  - Comparison across the cities – sentiment, volume etc.
  - Sentiment analysis – overall sentiment of general public for a phrase/hashtag
- Some more optional ideas:
  - Faceted search on named entity
  - Summarization – either on hashtags or topics
  - Any other analysis that you can come up with.
- UI
  - Innovative ideas on analysis and UI are encouraged.

Final Deliverables

- A short demo video (at most 3 minutes)
- A working 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 about a particular topic, including relevant tweets and analysis results.
- Grading is based on relevancy, language spread of served results and utility in understanding the topics.
- Points distribution:
  - IR – 4 points
  - Analytics and UI – 5 points
  - Report – 1 point

Project Summary

- The project is fairly open-ended and permits usage of any third party tools that you deem relevant
  - Only restriction is to use Solr for indexing.
- Primary objective is to encourage students to apply IR concepts in solving real world problems
- 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

- 16th November (Today): Project released
- 6th December, before 5 PM: Submit videos for class presentations (optional)
  - Sign-up sheet will be released 3 days before
- 7th December: In-class presentation for selected groups (at-most 2 bonus points)
- 9th 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

- 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