CSE 435/535 Information Retrieval
Fall 2015
Project Part C: Multilingual Search System

Due Date: 23:59, Dec 11th, 2015

Overview

The goal of this project is to build a multilingual faceted search system, including a front end that allows users to search and browse multilingual data based on various criteria: topic, location, person, etc.

The following sections describe the various tasks involved, evaluation criteria and submission guideline.

Data

The data will be multilingual social media data (including but not restricted to Twitter) in multiple languages. You may reuse some of the data from the previous project, but you must include new data as well. The new data could reflect:

- New languages: French, Arabic, especially Syrian dialects are especially interesting
- New topics: extend the topic to include the Paris attacks, and relevance to Syrian refugee crisis

We enforce the following minimum requirements with regards to the data however:

- Minimum data in four languages
- At least 200 posts per language
- At least 10000 posts in total

Index

In this step, you will need to index the data as you have done in part A and B. You are free to choose whatever IR model, query processing that you wish.

Front-end User Interface

In this project, you will be required to build a user interface in order to do one of the following (i) accept queries from a user, (ii) display search results, (iii) display analytics based on indexed data. The user interface will depend on which of the project components (see below) you have implemented. In summary, you will need to create a working site where users can try out your system.

Project Components

Your project must reflect at least two of the components listed below. The demonstration of your project should clearly showcase the components/features you have implemented.

(i) Content Tagging (Monolingual)

Everyone is encouraged to attempt some level of content tagging of the data. Content tagging could include tagging of named entities (names of people, places, organizations), topics, contact information, etc. Tools such as Alchemy or the Stanford NLP Toolkit may be leveraged for this purpose. The tagged data can subsequently be used in faceted search, analytics, graph analysis, etc.

(ii) Faceted Search

This option involves leveraging the faceted search capability provided by Solr to allow various types of drill-down. This assumes some level of content tagging (see above). Facets could include people, topics, locations etc. You are encouraged to experiment with different UI options including hierarchies, graphs, etc.

(iii) Cross-Document Analytics

This option involves computing various analytics that provide insight into the data. Examples include: volume of tweets by region/topic/hashtag, sentiment analysis, analytics illustrating cultural differences, etc. The ability to identify and display trending topics (on a daily/weekly basis) would also be interesting.

Analytics should be presented using intuitive visual graphs – several widget libraries are available. Map visualization is also encouraged.

(iv) Topic Models and/or LSI

In this option, you will implement Latent Semantic Indexing on the corpus of data you have collected to demonstrate “semantic search”, rather than traditional keyword search. For those of you familiar with advanced machine learning techniques such as topic models (LDA), you are encouraged to apply such techniques to the data in order to discover and group tweets based on different topics. You may not use the in-built Carrot clustering to complete this section however.

(v) Cross-Lingual Retrieval/Analysis

In this option, you will demonstrate cross-lingual capabilities. This can take on many aspects: one example involves cross-lingual queries, and automatic translation of resulting foreign language snippets. For example, a search for a particular individual/place/organization should take place simultaneously in multiple languages – achieved by automatically tagging and normalizing entities across languages.

(vi) Ranking tweets

This option involves coming up with a novel ranking algorithm for tweets that balances recency with importance of content when presenting tweets. It could also take into account the popularity of a tweet, or the influence of a person tweeting, the location of the user, their interests etc..

(vii) Summarization

This option focuses on summarizing tweets through the use of news/Wikipedia articles. You can pick a particular hashtag or a named entity like a person, place, etc. and provide a summary based on your index. The task would involve partitioning your data into sub-topics or sub-events based on tagged information and then choosing a summary for each sub-topic. Since we do not expect language generation, you could use news headlines or extracts from WIkipedia articles as bullet points in this summary.

(viii) Graphical Analysis

This option involves inferring some graphical structure from the tweets, based on entities mentioned, topics discussed etc. Graph structures (or relationships between tweets) could also be inferred through connection of topics reflected in the tweets: wikification may be helpful in this process. Once a graph is constructed, use graph algorithms to find important tweets, entities etc.

What to submit

1. Your report in **pdf** format. File name is **report.pdf** (no other file format is allowed)
2. A video (no more than 3 minutes) that demonstrates the key features of your system. Make sure to include a voiceover and/or text that helps the viewer appreciate what they are seeing.
3. Your source file (java files) of any customized functions in src folder.

Compress these files into a tar file. File name is project_partc_[ubitname].tar (no other compressed format is allowed)

For example my ubit name is ruhan then I should use following command to submit.

submit_cse535 project_partc_ruhan.tar

Choose cse435 or cse535 based on your own course level. Although multiple submissions can be made till the deadline, we recommend that one team member make all submissions to ease the grading process.

Grading

Grading for this project will be based on (i) sophistication of techniques implemented, (ii) demonstration of features in video, (iii) project report, and (iv) functioning demo site.

Presentation

We will select projects for in-class presentation on Dec 8th. More on this forthcoming.