CSE 435/535 : Information Retrieval

Project 4 : Dissecting Twitter data streams

Agenda

- Overview of Project 4
- Sub-project details
- Some tips and tricks
- Timelines, deliverables, logistics, etc.

Project 4 : Full scale IR systems

- The first three projects dealt with the following:
    - Project 1 : Indexing & Crawling
        - How do you gather data on a particular topic?
        - How do you effectively index this data?
    - Project 2 : Scoring
        - How does query scoring work?
    - Project 3 : Relevance
        - How do you tune relevance for specific query needs
- Project 4 seeks to unify these subtasks into a single end-to-end IR system

Digesting Twitter Streams

- Twitter data streams are dynamic, multi-faceted entities that spread across multiple dimensions:
  - Topics : A given topic often splits into smaller sub-topics and/or the main topic itself shifts
  - Languages : Tweets across languages may either be related or take a life of their own and be disjoint in content and sentiment between languages
  - Facts : A variety of entities and their relations may be embedded in a single stream thus allowing question answering o the stream

Sub-project 1 : Topic summarization

- Ingest tweets on a particular hashtag - streaming data most likely, one or more languages
- Detect subtopics and divide into subsets
- Present summaries of the sub-topics
- End goal : Enable the user to “understand” a given hashtag
- Grading : Utility in understanding the topic

Sub-project 2 : Cross-lingual IR

- Ingest tweets on a particular hashtag in multiple languages, search or streaming
- Determine ways to extract cross-lingual or semantic equivalences and index data as such
- Perform search across languages for a given query
- End goal : Serve relevant content to a user irrespective of the language of the tweet
- Grading : Relevancy and language spread of served results

Sub-project 3 : Question Answering

- Ingest tweets for a particular hashtag over an extended time period, in one or selected languages
- Extract information i.e. facts from the incoming tweets
- Support answering questions from the extracted facts
- End goal : Ability to answer questions for a given stream
- Grading : Types of questions that can be answered and the veracity of the answers

Project focus

- The project is fairly open-ended and permits usage of any third party tools that you deem relevant
  - Only restriction is use Solr for indexing purposes
- 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 on
- Don’t be afraid to be creative and stand out!

Tips and tricks

- Topic summarization
    - Think of ways to distinguish sub-topics : index time or query time?
    - What constitutes a “summary”?
    - UI could play a vital role - displaying sub-topics and summaries
- Cross lingual IR
    - You control the languages you choose
    - Think of different ways to index and query : all tweets are still related to a topic
    - How do you determine relevance?
- Question - Answering
    - How do you detect Named Entities - people, places, dates?
    - Can you extract relations?
    - Can relations be used to parse questions?
- More in recitation

Other details

- Work in teams of 4, registration form to be available today!
  - Register teams within three days
- Provide a preference between the three projects
  - FCFS allotment on a fixed number of slots
- Final deliverables
  - A short demo video (at most 3 minutes)
  - A working application URL
  - A short report detailing all work done and member contributions

Timeline

- 17th November (Today) : Project released
- 20th November : Final allotments
- 1st December : Testing hashtags announced
  - All system testing, demos, QA etc will happen on these hashtags
- 6th December : Submit videos for class presentations (optional)
- 8th December : In class presentations (bonus points)
- 9th December : Final submissions due