CS 221 Information Retrieval Projects

CRISTINA VIDEIRA LOPES

Project Options

- Search engine for Wikipedia
- Sentiment analysis for Twitter
- Chemistry community analysis

- Additional projects may be considered

1. Search Engine for Wikipedia

- Goal: retrieve Wikipedia pages related to the query and order them according to relevance
- UI: regular web-browser search engine interface
- Topic coverage: text processing, link analysis, feature extraction, indexing, ranking and scoring, evaluation
  - Additional optional topics: map-reduce
- Main components: lucene and/or katta
- Baseline: google over site:wikipedia.org

2. Sentiment Analysis for Twitter

- Goal: find out how the Twitterverse is feeling about certain keywords
- UI: http://twittersentiment.appspot.com/
- Topic coverage: text processing, feature extraction, classification, indexing, scoring, evaluation
  - Additional optional topics: map-reduce
- Main components: wekka or rapidminer
- Baseline: we have a test set

3. Chemistry Community Analysis

- Goal: some people believe that Chemistry is a much more closed field than all other scientific fields. Find out whether this is true or false.
- UI: none, not interactive
- Topic coverage: web crawling, text processing, feature extraction, noise reduction, network analysis
  - Additional optional topics: map-reduce
- Main components: up to you, but talk to me
- Prize: if your project is well done, and the conjecture turns out to be true, I will work on a paper with you

Project milestones

- There will be clear milestones for each project
  - E.g. by 1/30 you need to have X and Y done, etc.
- Will be disclosed by 1/18
- Will be checked and assessed by the Reader, will be given a grade
  - See them as many mini-projects
- Final project grade = weighted combination of milestone grades

Project milestones

- First milestone: form groups by 1/20
  - Groups of 1, 2 or 3, but no more
  - Use Message Board if needed, Forum “Social”
- On 1/23: I will bring a sign-up sheet to class

Gamifying the projects

- Shows the importance and feasibility of evaluation
- We may use kaggle: http://inclass.kaggle.com/
  - Example: http://inclass.kaggle.com/c/si650winter11
- The competition is not the project and vice-versa
- Your grade in the course will not depend on your competition score
  - Competition is only for fun / may be used to show off your skills to prospective employers