Projects

Group Project

- Programming a system in groups of 3
- **March 6th** : Send in your project selection and partner names.
- **March 27th** : Project Poster, give an overview of your project execution and work division.
- **April 13th** : Submit the project.

Projects Continued

Carries 20% weight

70 : basic implementation

30 : Execution novelty bonus

Caveats

Projects that

- Aim to extract novel insights from data
- Propose a new method
- Solve a novel problem around provided data

Will be given higher weightage

Report

- Clear description of your system
- Motivations of your design choices
- Sensible manipulation of the data
- Appropriate choice of evaluation measures

Deliverables

- **Submit a report summarizing the goals, implemented system, experiments and results.**
- **Design and present a poster at the end of the course i.e. March 27th**
- **Poster should cover what you *plan to execute***

Opinion Detection On Twitter

Analyzing the tweets to identify the political opinions expressed by users for a candidate.

- Platforms like Twitter are very popular for Political campaigning.
- So many people tweeting about their favourite party or candidate.
- One could mine this data in interesting ways to extract and summarize political sentiments.

Dataset : http://bit.ly/eu72Lr

Readme : http://www.ayman-naaman.net/2010/11/21/twitter-sentiment-dataset-online/

Deliverables : A system that summarizes political sentiment given some tweets

Selection of Tag for Tag Clouds

Summarize a large document with a Tag Cloud.

- There are several blog posts out there without tags.
- Some phrases represent key concepts in a document
- One should be able to extract these relevant tags from these documents
- Target Relevance and Diversity of tag cloud

Dataset : http://www.markusstrohmaier.info/datasets/

Demo : http://logd.tw.rpi.edu/demo/multi-word_tag_cloud_government_dataset_titles

Deliverables : Given a set of documents, the system should be able to output a tag cloud

Interleaved evaluation

Compare and contrast existing learning to rank approaches via pairwise interleaving.

- We can train a model to rank documents with respect to a query.
- What if we have two mechanisms of ranking documents. Then, you can evaluate search via user clicks. Compare two rankers via interleaved comparisons. You can explore LambdaRank or RankNet etc. There are packages link mallet, Lerot that can help train Rankers.
- Pick and choose some L2R methods and compare them via Lerot.

Datasets: https://bitbucket.org/ilps/lerot

Useful links : http://research.microsoft.com/pubs/206529/cikm-livinglab-2013-lerot.pdf

PageRank

Implement distributed PageRank using MapReduce

- Document collections can run into terabytes, how do you score 10Mil documents in less than 1 sec ?
- Parallelize PageRank to work simultaneously on several documents.

Dataset : http://lintool.github.io/Cloud9/docs/content/wikipedia.html, bigger sample:https://www.dropbox.com/s/x60qi8rlxl77jd5/clueweb.tar.gz?dl=0

Deliverables : System that can rank large collection of documents using Pagerank with time estimates

Mining product attributes

Build a system that extracts product attributes and corresponding user sentiments from his/her review.

- For Cell Phones sentiments can be extracted for Picture Quality, Music Player Quality etc
- Aim to extract as many product features and corresponding sentiments as possible.

Data : Amazon product review data, http://users.cis.fiu.edu/~yzhan004/datadownload/amazon.rar

Deliverables : Given some reviews about a product, the system should present user's feature wise sentiments