COMPGI15/M052 Information Retrieval and Data Mining (IRDM)

Course Work

Dr. Jun Wang and Dr. Emine Yilmaz

Computer Science, UCL

Summary

- Objectives:
  - First hand knowledge of IR and DM algorithms.
  - Acquainted with the evaluation procedure

- One individual assignment and one group project
  - Assignment Part A and Part B (20%)
  - Group Project on an IR or DM problem (20%)
    - project topics are listed in this slides

Deadlines

- ❑ Assignment Deadline March 31st 2016
  - ❑ submission link: https://moodle.ucl.ac.uk/mod/turnitintooltwo/view.php?id=2300991
- ❑ Group Project:
  - ❑ 3 persons one group. Inform us your plan (**group member and project topic**) **by Midnight March 1st via** **https://docs.google.com/spreadsheets/d/1E0Oh5Q0uXrM_Jx5TDrMHrjwb19P1MihRvLWprVK6LuY/edit#gid=311367834**
  - ❑ Submit your poster of the group project for us to print (Stay tuned)
  - ❑ Poster Session (for the group project only) **March 18th** (Stay tuned)
  - ❑ Final report (each one group) **by Midnight April 16th**
    - – Submission link: **https://moodle.ucl.ac.uk/mod/turnitintooltwo/view.php?id=2301021**

Assignment Part A: Text Retrieval

- ❑ For the step-by-step guide and task descriptions, check moodle:
  - ❑ https://moodle.ucl.ac.uk/mod/resource/view.php?id=2294587
- ❑ What you need to do:
  - ❑ Download Clueweb data from the provided URL in the guide
  - ❑ Based on terrier retrieval engine (Java)
  - ❑ Indexing, retrieval, and evaluation of a text retrieval system

Assignment Part B: MapReduce

- It is a software framework aiming to support distributed computing on large data sets on clusters of computers. Invented by Google
  - Many real world tasks are expressible in this framework
  - Making large-scale data-driven modeling possible
- Amazon provides us with free Elastic Compute Cloud (Amazon EC2) accounts
  - Each of you will get an AWS account (worth100$)
  - You need to work alone

The Group Project

- It is a programming assignment
  - Mainly use Java, but other programming languages are possible
  - 3 persons per group
- Assessment of the report
  - 20% weights
  - Clear description of your systems/algorithms
  - Motivations of your design choices
  - Sensible manipulation of the data, and *new* findings/conclusions from your analysis
  - Appropriate choice of evaluation measures

The Group Project

- ❑ Choose one of the projects shown next
- ❑ Poster presentation at the last lecture of the course

  - briefly summarize the goals, initial methods and preliminary experiments/results of your project.

- ❑ Report the results and study in the final report
  - ❑ Clear description of your systems/algorithms
  - ❑ Motivations of your design choices
  - ❑ Results of your data analysis and experiment
  - ❑ Appropriate choice of evaluation measures and conclusion from the study
  - ❑ You are allowed to base on any online materials, but need to cite the source and specify any new findings (discover, insights, conclusions) from your results

Group Project Option 1

Search Engine for the UCL website

- build a live search engine that indexs and searchs content within domain ucl.ac.uk
- compare it with
  - http://search2.ucl.ac.uk/s/search.html?query=computer%20science&collection=website-meta&profile=_degrees&tab=degrees
  - https://www.google.co.uk/#q=computer+science+site:ucl.ac.uk
- Use any open source IR package whenever it is needed
- Implement a PageRank algorithm for ranking
- Analyse and report your results:
  - 1) generate your own test set with queries and judgements
  - 2) compare the results with the above two with right metrics
  - 3) write a clear manual, report, and upload the code to github with a clear indication it is for IRDM 2016 group project at UCL

Group Project Option 2

Sentiment analysis and other kinds of analysis i.e. trending topics, common phrases etc over Twitter data

  - Use data mining techniques to explore the twitter data (ask our TAs for the dataset)
  - Develop statistical models to predict sentiments, trends, events etc
- Analyse and report your results:
    - 1) review the related work and provide your solutions
    - 2) properly evaluate using suitable metrics
    - 3) write a clear manual, report, and upload the code to github with a clear indication for IRDM 2016 group project at UCL

Group Project Option 3

Learning to rank for information retrieval

- Learning to rank is widely used for search engines
- This project aims to implement and test various techniques and compare them. Possible topics:
  - *Freshness and Relevance*
  - Predicting Search Satisfaction Metrics
  - Active learning
- Suggested dataset: LETOR, Package: Lerot
- Analyse and report your results:
  - 1) review related work and provide a solution(existing or new)
  - 2) properly evaluate using suitable metrics
  - 3) write a clear manual, report, and upload the code to github with a clear indication for IRDM 2016 group project at UCL

Group Project Option 4

Intent mining from Query Logs

- With billion queries for search engines, it is worthwhile to mine these logs to understand or facilitate a new query. Given a query that is ambiguous or underspecified (e.g. ‘harry potter’, ‘apple’, ‘jaguar’), one can mine search logs to suggest more queries (e.g. ‘harry potter philosopher's stone movie’, ‘apple phone’ or ‘jaguar latest model’) or just expanding the current query to improve search results.
- The aim of this project is to extract query subtopics or intents from a large corpus of search logs. A sample query log and test topics will be provided to the teams. Participating teams are encouraged to try different sources for identifying subtopics, such as document collection, wikis etc.
- Dataset: ***AOL*** ***and MSN LOG, Trec tasks track and IMine task*** *(ask TAs)*
- Analyse and report your results:
  - 1) review related work and provide a solution(existing or new)
  - 2) properly evaluate using suitable metrics
  - 3) write a clear manual, report, and upload the code to github with a clear indication for IRDM 2016 group project at UCL

Group Project Option 5

The HomeDepot Kaggle Challenge

- The HomeDepot Kaggle challenge requires participants to produce a model ranking products sold by HomeDepot by relevance to user queries. This requires development of a Learning to Rank model, but the specific domain (of DIY and home-related products) offers some additional interest to the general problem of ranking text documents relative to a query. Implement a ranking model of your choice and report your results on the test dataset provided. Extra credit available if your solution appears on the Kaggle leaderboard!

- Dataset: Homedepot Kaggle Data
- Analyse and report your results:
  - 1) review related work and provide a solution(existing or new)
  - 2) properly evaluate using suitable metrics
  - 3) write a clear manual, report, and upload the code to github with a clear indication for IRDM 2016 group project at UCL

Group Project Option 6

Collaborative Filtering and Recommendation

- Collaborative filtering is a popular for e-commerce
- This project aims to implement and test various techniques and new ideas. Possible topics:
  - Address coldstart - *recommendation for new users/items*
  - Using deep learning
  - Predict when to recommend
- Datasets: shopping transactions; moviesLens etc
- Analyse and report your results:
  - 1) review related work and provide a solution (existing or new)
  - 2) properly evaluate using suitable metrics
  - 3) write a clear manual, report, and upload the code to github with a clear indication for IRDM 2016 group project at UCL

Group Project Option 7

Time series forecasting

- How to identify patterns in time series is critical for many businesses including finance and e-commerce
- This project aims to implement and test various time series techniques for forecasting. Possible topics:
  - Using deep learning & neural networks reading1, reading2, reading3
  - Using regression. reading1, reading2
- Datasets: Energy dataset; Climate data; UCI
- Analyse and report your results:
  - 1) review related work and provide a solution (existing or new)
  - 2) properly evaluate using suitable metrics
  - 3) write a clear manual, report, and upload the code to github with a clear indication for IRDM 2016 group project at UCL

Group Project Option 8

Mining fine-art paintings for creativity understanding

- Can we mine fine-art painting data to assess creativity?
- This project aims to implement classification techniques for fine-art painting understanding. Possible topics:
  - Quantifying creativity
  - Classification of paintings; additionally, fashion classification
- Datasets: wikiart; artchive; wga
- Analyse and report your results:
  - 1) review related work and provide a solution (existing or new)
  - 2) properly evaluate using suitable metrics
  - 3) write a clear manual, report, and upload the code to github with a clear indication for IRDM 2016 group project at UCL