Web IR 2018/2019 - PROJECTS

1 RULES

1. You have to present the project either before or in the same session as the written exam
2. The project and the oral exam will contribute 50%-50% to the final evaluation
3. exception to the rule 1 is
   - june 2019 session, you can present the project by july session
4. Groups are consist of up to 3 members. Exceptions must be agreed with the instructors

2 ASSIGNMENT

1. You form groups
2. We provide you with a list of representative research papers on a number of WIR tasks and a list of possible datasets or a process to get it
3. You are free to propose a paper and a dataset
4. A paper (“assigned paper” in the following) will be assigned to each group on a FIFO basis
5. You have to design and perform experiments in the spirit of the ones in your assigned paper, using the same or a different dataset. **Note that we are more interested in the way you tackle the problem rather than in the final outcome of the experiment.** Your experiment must be agreed with us. In more detail:
   a. you select or propose a paper
   b. the selected/proposed paper is approved by us and becomes the “assigned paper”
   c. you spend a few days to understand the main ideas, experiments etc in the paper
   d. you come to us with a written proposal (half a page) of an experiment to perform on at least one of the main points of your “assigned paper”
   e. we agree or propose modifications
   f. you can propose alternative datasets
   g. now your experiment is approved and you can proceed
6. You have to discuss the findings of your study in a presentation with **at most 10 slides** organised according to the following structure:
   a. Introduction
   b. related work emphasizing the “assigned paper”
   c. dataset description.
   d. experimental setting. In this section you should point out and clarify possible simplistic assumptions or choices that you made with respect to the reference paper
   e. results. Please discuss if and why in your opinion results differ from the reference paper.

7. Dataset and code to perform the experiment must be delivered, together with the short presentation before the workshop day.
8. On the workshop day, whose schedule will be notified in advance, you will have to present your slides in 10-15 minutes + 5 minutes of Q/A

3 DATASETS

Each dataset is characterised by a unique structure. One of the purpose of the project is to properly handle such characteristics (see 4.c), some of which are briefly listed below:

1. presence of a ground truth to evaluate the quality of your algorithm
2. big data. Unfortunately, in many cases, size does not allow to handle a dataset in a convenient way. In case you chose a "big" dataset, this will be considered as a plus in the final evaluation. The attribution of "big" is assigned in point 3.f above, namely when the experiment is approved.
3. structured text, i.e. you can distinguish between title, body etc.
4. training and test set availability for classification purposes
5. linked structure, i.e either there is a "natural" link structure (e.g. hyperlinks in documents) and/or one can be inferred (e.g. co-authorship in papers)
6. etc

3.1 LINKS TO DATASETS

- UCI Machine learning repository https://archive.ics.uci.edu/ml/datasets.html
- Stanford Large Network Dataset Collection https://snap.stanford.edu/data/
- IR-Query-Processing, ground truth, big data, real data for queries TREC clue web 09 B (too big, https://lemurproject.org/clueweb09/, you have to pay)
- IR-Query-Processing, ground truth, real data: http://ir.dcs.gla.ac.uk/resources/test_collections/ (suggested Cranfield and time)
- IR-Query-Processing, real data: any collection of texts the corpus is obtained querying suitable API (bulk download)
    - Wikipedia (available in a single file, DBpedia) https://en.wikipedia.org/wiki/Wikipedia:Database_download
    - DBLP (available in a single file)
    - ArXiv
    - PubMed
    - CiteseerX
    - Twitter
- Microsoft Academic Graph/Artminer
- DBLP
- Kaggle
- Reuters

- Movie reviews pos/neg: http://www.cs.cornell.edu/people/pabo/movie-review-data/review_polarity.tar.gz
- Spam/Ham youtube comments: https://archive.ics.uci.edu/ml/datasets/YouTube+Spam+Collection
- Amazon reviews pos/neg
- https://shiring.github.io/networks/2017/05/15/got_final
- http://law.di.unimi.it/datasets.php
- https://networkit.iti.kit.edu/
- https://grouplens.org/datasets/movielens/
- https://www.reddit.com/wiki/api

4 SELECTED PAPERS

You are free to propose a paper **that must be approved by the instructors, by filling in the form at the following link:**

https://forms.gle/sM9xHcu79iPwYXfg8

In general **one paper cannot be assigned to more than one group**. The following are possible examples of the kind of papers you should consider

- Detecting Spammers on Twitter, http://homepages.dcc.ufmg.br/~fabricio/download/ceas10.pdf
- Identifying document topics using the Wikipedia category network, https://dl.acm.org/citation.cfm?id=1249180
- Assessing the value of cooperation in Wikipedia, https://arxiv.org/abs/cs/0702140
- Twitter Trending Topic Classification, https://dl.acm.org/citation.cfm?id=2119627
- Using Social Media to Enhance Emergency Situation Awareness, https://ieeexplore.ieee.org/document/6148196/
- Mining Wikipedia to Rank Rock Guitarists, http://www.mecs-press.org/ijisa/ijisa-v7-n12/IJISA-V7-N12-5.pdf
- Discovering Missing Links in Wikipedia, https://dl.acm.org/citation.cfm?id=1134284
- Measuring Article Quality in Wikipedia: Models and Evaluation, http://www.mysmu.edu/phdis2008/meiqun.hu.2008/pub/cikm07p.pdf
- Exploiting Wikipedia as External Knowledge for Document Clustering, https://dl.acm.org/citation.cfm?id=1557066
- Short Text Classification in Twitter to Improve Information Filtering, https://dl.acm.org/citation.cfm?id=1835643
- Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors, https://dl.acm.org/citation.cfm?id=1772777
- Entity Extraction, Linking, Classification, and Tagging for Social Media: A Wikipedia-Based Approach, https://dl.acm.org/citation.cfm?id=2536237

- Mapping the echo-chamber: detecting and characterizing partisan networks on Twitter. http://sbp-brims.org/2017/proceedings/papers/challenge_papers/MappingTheEcho-Chamber.pdf
- PageRank Beyond the Web, https://epubs.siam.org/doi/pdf/10.1137/140976649
- Improving recommendation lists through topic diversification, https://dl.acm.org/citation.cfm?id=1060754