Information Retrieval Project

Prof. Dr. Goran Glavaš
Fabian David Schmidt
Benedikt Ebing

Center for AI and Data Science (CAIDAS)
Fakultät für Mathematik und Informatik
Universität Würzburg

Team project

- Schedule
  - Topics published: May 24
  - Topics selected and confirmed: June 04

  - **Project coaching:**
    - Ask questions to Fabian/Benedikt in case you are stuck

  - **Project presentations:** July 18
    - Present what you did: methods/models, implementation, evaluation
    - 10 minutes per team + 5 min Q&A
    - All team members should present and clearly state what their contribution was

Team Project

- **Purpose:** „hands-on" experience implementing and evaluating information retrieval model(s) and performing IR tasks
  - Best way to understand something is to (try to) implement it

- Other goals:
  - Experiencing teamwork
  - Exercising how to clearly present results of your work

Topics

1. Learning to Rank (Supervised Retrieval)
2. Cross-Lingual Sentence Retrieval
3. Efficient Vector Space Retrieval
4. Sentiment Analysis with Text Similarity and Link Analysis Algorithms
5. Retrieval with Pretrained Neural Language Models

Topic 1: Learning to Rank (Supervised Retrieval)

- In some settings, we have enough relevance judgements to train supervised retrieval models
- **Learning to rank** (L2R, LETOR): training supervised machine models for IR
- **Task:**
  - Implement and evaluate two L2R models
    - One *point-wise* L2R model
    - One *pair-wise* L2R model
  - Design good, informative features for both models
    - Different unsupervised ranking functions can be used as features
  - Evaluate the performance of the models on test collections

Topic 1: Learning to Rank (Supervised Retrieval)

- Point-wise L2R model
  - One training instance is a query-document pair (q, d)
  - You are predicting whether the document is relevant for the query
  - Ranking: order documents by the classifier's confidence
- Pair-wise L2R model
  - One training instance is a triple (q, d1, d2) consisting of a query and two documents
  - You are predicting which of the two documents (first or second) is more relevant for the query
  - Ranking: merging pairwise decisions into consistent ordering
- Datasets:
  - Medical Information Retrieval dataset:
    - https://tinyurl.com/nfcorpus

Topic 1: Learning to Rank (Supervised Retrieval)

- **Task:**
  - **Own** implementations of features, but you may use
  - Existing implementations of L2R algorithms
  - Existing implementations of evaluation metrics (MAP, MRR, NDCG)
  - RankLib – a L2R library you may use
    - https://sourceforge.net/p/lemur/wiki/RankLib

Topic 2: Cross-Lingual Information Retrieval (CLIR)

- **Cross-lingual retrieval**: query is in a different language from document collection
- Creating a retrieval system, that can, given a sentence in one language recognize its translation from a large collection of sentences in another language

Topic 2: Cross-Lingual Sentence Retrieval

- Inducing a multilingual embedding space from monolingual word embeddings
  - Many ways to do it:
    - https://arxiv.org/pdf/1902.00508.pdf
    - https://arxiv.org/pdf/1710.04087.pdf
    - http://aclweb.org/anthology/P18-1073
  - Simplest way to do it:

$$\begin{array}{r} \\ \textbf{bird} \\ \textbf{pretty} \\ \textbf{...} \\ \textbf{eat} \end{array} \begin{array}{c} \mathbf{S} \\ \begin{bmatrix} -1.18 & 0.21 & ... & 0.11 \\ 0.23 & -0.53 & ... & 0.34 \\ ... & ... & ... & ... \\ 0.78 & 1.33 & ... & -0.47 \end{bmatrix} \end{array} \cdot \mathbf{M} = \begin{array}{c} \mathbf{T} \\ \begin{bmatrix} 0.59 & 1.01 & ... & 0.37 \\ -0.34 & -0.27 & ... & 0.41 \\ ... & ... & ... & ... \\ 0.81 & -0.31 & ... & 0.29 \end{bmatrix} \end{array} \begin{array}{l} \\ \textbf{Vogel} \\ \textbf{schön} \\ \textbf{...} \\ \textbf{essen} \end{array}$$

Topic 2: Cross-Lingual Sentence Retrieval

- **Task:**
  - You may use some pre-trained multilingual word embeddings
    - E.g., https://github.com/Babylonpartners/fastText_multilingual
      https://github.com/facebookresearch/MUSE
  - Implementation of the **supervised classification** for recognizing sentence translation pairs using multilingual word embeddings
  - Evaluation of the models on EuroParl datasets
- **Datasets**:
  - EuroParl parallel corpora: http://opus.nlpl.eu/Europarl.php (sentence-level CLIR)

Topic 3: Efficient Vector Space Retrieval

- **Efficient IR system** needs to be able to retrieve results, in real-time, from very large document collections
- The goal is to implement the **Vector Space Model** model with all „tricks” for efficient retrieval

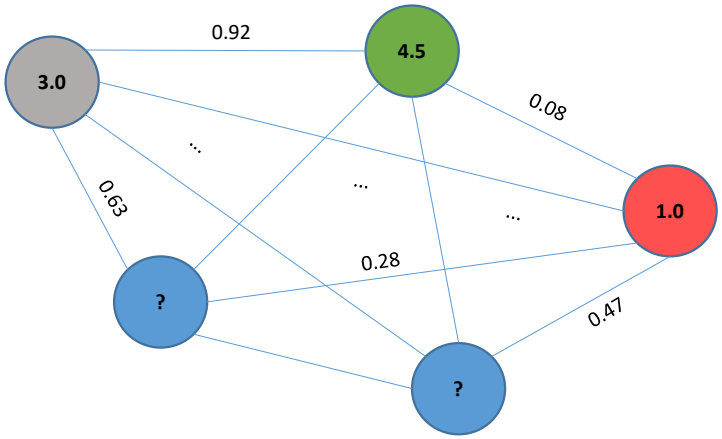
Topic 3: Efficient Vector Space Retrieval

- **Task**:
  - Own implementation of the basic VSM model (TF-IDF weighting + cosine ranking)
  - Own Implementation of speed-ups (inverted index, pre-clustering)
  - Evaluation of all VSM variants in terms of both retrieval performance and efficiency
- **Datasets**:
  - Medical Information Retrieval dataset: https://tinyurl.com/nfcorpus
- Tip:
  - FAISS: a library for fast computation of vector similarity/distance
  - https://github.com/facebookresearch/faiss

Topic 4: Sentiment Propagation with Link Analysis

- **Propagating properties** between texts based on their mutual similarity, using link analysis algorithms (PageRank)
- Task:
  - You're given some product reviews with assigned sentiment ratings, and others without the sentiment ratings
  - Implement a measure of similarity between texts
  - Induce a (fully-connected) similarity graph by computing the similarity between all pairs of documents
  - Use the induced graph and link analysis algorithms (PageRank) to learn the ratings of unannotated reviews from the ratings of the reviews for which you know the rating

Topic 4: Sentiment Propagation with Link Analysis



- Graph computation
  - For every two texts, compute some score of (semantic) similarity
    - E.g., **VSM similarity** and/or
    - **Semantic similarity using word embeddings**
  - These scores become weights of the graph edges
- Label propagation:
  - Compute the sentiment scores for uknown nodes
  - PageRank

Topic 4: Sentiment Propagation with Link Analysis

- **Task**:
  - Implement a measure of similarity between product review texts:
    1. Based on sparse text representation: cosine similarity over TF-IDF vectors
  - Induce the fully-connected graph of reviews and induce the sentiment of unlabeled reviews using:
    - PageRank (own implementation)
  - Evaluation
- **Datasets**:
  - Amazon reviews dataset: http://jmcauley.ucsd.edu/data/amazon

Topic 5: Ad-Hoc Retrieval with Pretrained Neural LMs

- **Task**:
  - Use some of the pre-trained neural language encoders (PLM), e.g., BERT or XLM, and apply them to (document and sentence) retrieval
  - **Unsupervised retrieval**
    - Use PLM to obtain text representations
    - Unsupervised retrieval based on cosine similarity
- Starting literature:
  - https://arxiv.org/pdf/1810.04805.pdf (BERT)
  - https://arxiv.org/pdf/1903.10972.pdf (BERT-based ad-hoc retrieval)

Organization

- Form groups of **3 students**
- Each group is allowed to pick a topic they like the most
- Topic selection and team forming
  - **Deadline**: **Sunday, June 04** (23:59)
  - **Send the email with:**
    - Student names and IDs (Matrikelnummer)
    - Selected topic

Organization

- **Submitting** the project results is via WueCampus
  - Upload results on WueCampus
  - Code (software) as one archive and presentation as PDF file
  - **Deadline** for submission: **???**

- **Evaluation**
  - 0 to 3 points