Information Retrieval 1

Summary

Andrew Yates
a.c.yates@uva.nl

University of Amsterdam

Outline

1. Organization
2. Recap
3. Conclusion

Last Q&A session

- Briefly, RecSys Q&A and discussion questions
- Opportunity to ask questions on any content
- Remainder: I will take requests for example exam/exercise problems to solve (or choose myself if there are no requests)

Exam

- “Open book”: you may bring a sheet of paper with notes on both sides (prepared any way you like)
- In-person, on paper
- Off-site (see DataNose)

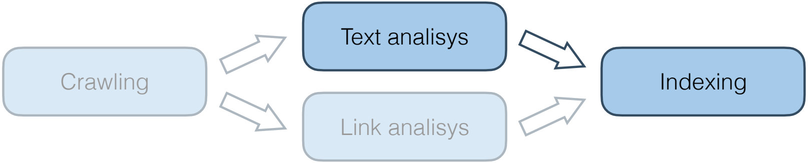
Outline

1. Organization
2. Recap
3. Conclusion

Information Retrieval 1

1. Basic techniques (IR0 recap)
2. Four pillars of IR
    - Evaluation
    - Document representation and matching
    - Learning to rank
    - IR-user interaction

Basic techniques



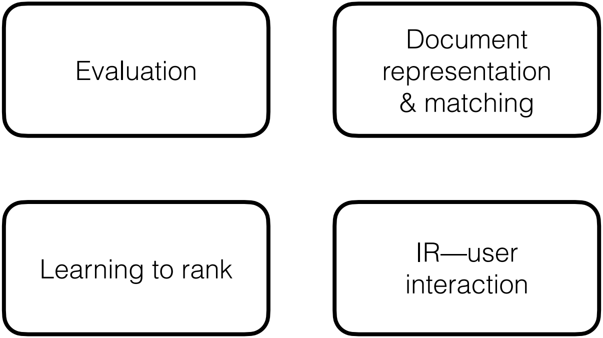
Text analysis

1. Statistical properties of written text
    - Zipf's law
    - Heaps' law
2. Text analysis pipeline
    - Stop-word removal
    - Stemming
    - Phrases

Indexing

1. Inverted index
    - Vocabulary
    - Inverted lists
2. Constructing an index
    - In-memory problem
    - Distributed indexing
3. Updating an index

Four pillars of IR



Evaluation

Document representation & matching

Learning to rank

IR—user interaction

(Offline) Evaluation

1. Test collections
   - Test documents
   - Test queries
   - Relevance judgements
2. Offline evaluation metrics
   - Unranked: precision, recall, F1
   - Ranked: RR, AP
   - User-based: DCG, RBP, ERR

Document representation and matching

1. Term-based retrieval
    - VSM+TF-IDF
    - QLM
    - BM25
2. Semantic retrieval
    - LSI
    - LDA
    - AWE & Doc2vec
    - KNRM & Transformer-based neural methods

Document representation and matching

1. Vector-based
   - Documents and queries as vectors
   - Match using cosine similarity
   - Methods: VSM, LSI, AWE, Doc2vec
2. Distribution-based
   - Documents and queries as distributions
   - Match using QLM or Kullback-Leibler divergence
   - Methods: QLM, LDA
3. Transformer-based
   - Don't fit neatly into other categories
   - Methods: CEDR, SentenceBERT, ColBERT

Learning to rank

1. Point-wise (standard ML)
2. Pair-wise
    - Point-wise model $f(d_i)$, pair-wise loss $\mathcal{L}(d_i, d_j)$
    - Method: RankNet
3. List-wise
    - Replace cost with $|\Delta evaluation\_metric|$
    - Method: LambdaRank

IR-user interactions

1. Interactions
   - Ambiguous and biased
2. Click models
   - Attempt to distinguish between bias and relevance
   - Methods: PBM, cascade model
3. Counterfactual and online LTR and evaluation
   - Debias logged data for learning and evaluation
   - Or learn/evaluate from online interactions

Scenarios

1. Conversational search
   - Document representation & matching
   - IR-user interactions
2. Recommender systems
   - Content-based
   - Collaborative filtering
   - Neural methods

Outline

1. Organization
2. Recap
3. Conclusion

Information Retrieval 2

- If you enjoyed this class, consider joining IR2 next semester
- Project-based course
- Guest lectures on advanced topics

Recruiting TAs for next year

- Responsibilities include helping with questions and assignments, possibly running LCs, and help with grading
- I will send an announcement after the exam has been graded

Conclusion

Thanks for following the course,
and good luck on the exam.