67-300 SEARCH ENGINES

RECAP

LECTURER: JOAO PALOTTI (JPALOTTI@ANDREW.CMU.EDU)

24TH APRIL 2017

BASICS - LECTURE 1

- Preprocessing steps:
  - Case fold
  - Tokenize
  - Stopwords
  - Stemming
- Tokens Vs Types
- Bag of Words Vs N-Grams

DOCUMENT REPRESENTATION - LECTURE 2

- What are documents?
- What are fields?
- Is there an unique way to represent a document?
- What is an (inverted) index?
- What kind of information do we add to the index?
- What is Boolean Search?

VECTOR SPACE MODEL - LECTURE 3

- How a document is represented in the vector space model?
- Three main components of most ranking algorithms:
  - What is TF?
  - What is IDF?
  - What is document normalization?
- Can you explain the meaning of TF-IDF?
- Which measure do we use usually to calculate distance in the vector space?

PROBABILISTIC MODELS - LECTURE 4

- What is the Probability Ranking Principle (PRP)?
- What is the Binary Independence Model (BIM)?
- What are the meanings of *binary* and in *independence* BIM?
- How BM25 deals with TF, IDF and document length normalization?

$$RSV_d = \sum_{t \in q} \boxed{\log\left[\frac{N}{df_t}\right]} \cdot \boxed{\frac{(k_1+1)tf_{td}}{k_1((1-b)+b \times (L_d/L_{ave}))+tf_{td}}} \boxed{\frac{(k_3+1)tf_{tq}}{k_3+tf_{tq}}}$$

LANGUAGE MODELS - LECTURE 5

- Unigram Language Models are simple to implement, yet highly effective. Can you manually run it given a set of toy examples?
- Why LMs are considered generative models?
- Would you be able to describe the how the next word suggestion of your favorite chat app in your mobile works?
- What is smoothing? Why do we need it?
- In special, smoothing with collection counts is good because it add an IDF factor to the already TF and doc length normalization factors. Can you explain why?

LM with Jelink-Mercer Smoothing: $P(w|d) = (1-\lambda)\frac{c(w,d)}{|d|} + (1-\lambda)p(w|Collection)$

QUERY REFORMULATION AND RELEVANCE FEEDBACK - LECTURE 6

- Two family of methods: global and local.
- Global:
  - Do not look at the results
  - Expand query using external resources or playing with relationship of words in the collection. Why is it challenging?
- Local:
  - Iterative process: do look at the results
  - Uses mostly positive feedback, why negative feedback might be problematic?
  - What is Pseudo-Relevance Feedback? How does it work? What are the parameters?

IR EVALUATION - LECTURE 7

- What is user happiness and what is the relationship of it with all proposed metrics?
- What is the Cranfield Paradigm? What are the key components of it?
- Can you formally define Precision and Recall?
- Can you tell me in which situation a metric is more adequate? (MRR, Precision@K, NDCG)
- What is AB-Test? How does it work?

IR EVALUATION 2 – LECTURE 8

- What is pooling?
- Why is pooling necessary?
- What is the main problem with pooling? (Think of someone testing him/her new super ninja retrieval method in a collection created 3 years ago - What kind of problem he/she might face?)

IN-CLASS-EXERCISE - LECTURE 9

- Already updated to Github and Piazza.

LINK ANALYSIS - LECTURE 10

- How is the anchor text used in retrieval? What is a Google Bomb and how can you make one at home?
- What is the relationship between citation network and quality?
- Can you describe the main concepts of Page Rank? What is teleport in the context of Page Rank?
- What is the meaning of Page Rank of a page?

  Probability of being at Page X after an infinite number of walks in the Web Graph
- How can we incorporate Page Rank in our models?

LEARNING TO RANK - LECTURE 11

- What are the 3 main tasks in ML?
- Why do we need ML in IR?
    - Small intermediary steps and feature engineering
        - Clustering users, classifying pages, filtering spam…
    - Learn how to weight different retrieval models
- Default Learning to Rank schema uses features from the query and the document. What are the problems of including features of single users? How can we deal with it?

SEARCH LOG ANALYSIS - LECTURE 12

- Main use of search logs is providing valuable information regarding the users. Sometimes disturbing and sensitive information.
- What kind of task can we perform examining the query logs of a search engine?
  - Again: Why do we want to cluster users?
  - Why is it important to correctly segment sessions?
  - Cite a way we could do query suggestion?
- What are pseudo-documents?

THANKS FOR YOUR PARTICIPATION

- Do not forget to review this course!
- Hope you have learned how the 10 blue links appear to you and what is the technology behind the search box.
- Hope you can go to your Linked in page now and understand that you are a document in that context and you want to be among the top documents retrieved. You should know how to do it well!
- Most of all, hope you have had a good time here!