First Review

CS5154/6054

Yizong Cheng

10/4/2022

8/23: Information Retrieval

- Collection, documents, terms, Boolean query, Assignment 1, IR1A.py
- Ad hoc retrieval and filtering
- Information need, relevance
- Concordance, inverted index, Quiz 1, Assignment 1, IR1B.py
  - vocabulary, postings, tokenization
  - Data structures for inverted index: dictionary of sets
- Term-document incidence matrix, Quiz 1, Quiz 2 1-5
  - Bipartite graph
- Pros and cons with Boolean retrieval
  - ranking

8/25: Regular Expressions

- Import re
  - re.split('\s', doc), re.findall('\w+', doc), re.sub(), re.search()
- Metacharacters []\.^$*+{}|()
- Special sequences \s, \w, \b
- Sets [….] with special characters ^ and -
- Counter, Assignment 2, IR2A.py
- WordCloud, Assignment 2
- Quiz 2 6-9

8/30: Ranking by Set Similarity

- Query and documents as sets of terms
- Four intersections between sets A, not A and B, not B.
  - 2 x 2 contingency count matrix
- Intersection computation with an inverted index
  - The Counter, Assignment 3, IR3C.py
- Jaccard coefficient, IR3A.py, IR3B.py
  - Hamming distance
- Stopwords
  - Postings list size (document frequency of a term)
- Quiz 3: how to make an inverted index

9/1: Using Jaccard Coefficient

- K-gram inverted index
- Spelling correction, IR4A.py
- Near duplicates, IR4B.py
- Jaccard coefficient as probability
- Assignment 4
- Quiz 4

9/6: Tf-Idf Weighting

- Context dependent similarity and ranking
- Document frequency and inverse document frequency (idf)
  - Idf is only meaningful when query has more than one term.
- Term frequency (tf)
  - Count matrix
  - Bag of words model for document
  - Sublinear tf scaling
- SMART notation of variants of tfidf
  - ddd.qqq triple for tf (n, l, b), df (n, t), and normalization (n, c), Assignment 5
- Tf-Idf weight matrix, Quiz 5
  - Tfidf ranking with the inverted index, Assignment 5, IR5A.py, IR5B.py

9/8: The Vector Space Model

- From sets to (binary) vectors
  - Dot-product = intersection
  - Cosine similarity of vectors = normalized intersection, IR6A.py
- Sklearn's CountVectorizer and TfidfVectorizer
  - Tokenization included with default re '\b\w\w+\b'.
  - encoding, ngram_range, max_df, min_df, binary
  - fit(), transform(), fit_transform(), get_feature_names()
  - Assignment 6, IR6D.py
- Quiz 6: cosine similarity ranking = Euclidean distance ranking, under one condition. Exercise 6.18 of iir

9/13: Precision and Recall

- Set of relevant documents (tp + fn)
- Set of retrieved documents (tp + fp)
- Intersection → true positives (tp)
- Set differences → false positives (fp) and false negatives (fn)
- Precision = tp/(tp + fp), recall = tp/(tp + fn), Quiz 7
- True negatives (tn) and accuracy.
- F measure and balanced F measure F1 (harmonic mean)
- Kappa statistic, Quiz 7
- Assignment 7, IR7A.py

9/15: Ranking Analysis

- Similarity between rankings
  - Kendall's τ and Spearman's ρ
- Similarity between a ranking and sets
  - Cureton's rank-biserial correlation
- Rankings and the (relevant) set
  - Precision-recall graph, interpolated precision, 11-point interpolated average precision
  - Mean average precision (MAP), P-precision, Quiz 8
  - ROC curve, sensitivity and specificity
  - Assignment 8, IR8A.py

9/20: Query Expansion

- Relevance feedback (RF)
  - Pseudo RF
- Thesaurus for query expansion
  - PubMed's UMLS, WordNet
- Automatic thesaurus generation
- Co-occurrence matrix CC^T
  - Linear_kernel and cosine_similarity in sklearn, IR9A.py
  - Quiz 9
- Twice cosine_similarity for synonyms, Assignment 9, IR9B.py

9/22: Probabilistic IR

- Probability, conditional probability, Bayes rule
- prior and posterior probabilities, odds
- The probability ranking principle and classification
  - Assignment 10: IR10A.py
- BIM, the binary independence model, Quiz 10
  - No term frequencies, naïve Bayes assumption
- pt, ut, c_t = log pt(1 – ut)/(ut(1 – pt))
- RSV(d) = $\sum_{t \text{ in } d} c_t$
- (11.21) for c_t
- Okapi BM25 (11.32)

9/27: Relevance Feedback

- Probabilistic approach to pseudo relevance feedback
- Assignment 11, IR11A.py
- Quiz 11

9/29: NDCG and MRR

- Discounted cumulative gain
- Normalized
- RR
- Quiz 12