exam prep

2023

question: Recall and nDCG are typically measured at a lower cutoff than MAP and MRR (you don't have to know the exact formula)

answer: False

- the MAP is a more general metric and captures the area under the precision-recall-curve → ~@100-1000
- but the MRR, DCG, nDCG measure how far-up in the search results the rel-docs are positioned → ~@5-20
- "… typically, we measure all those metrics at a certain cutoff at "k" of the top retrieved documents. And for MAP and recall this is typically done at 100 and at 1000, whereas for position MRR and nDCG we have a much lower cutoff, so at 5, at 10 or at 20 to kind of get the same experience as users would do." → see: https://github.com/sebastian-hofstaetter/teaching/blob/master/advanced-information-retrieval/Lecture%202%20-%20Closed%20Captions.md#14-ranking-list-evaluation-metrics

---

question: Judgement pairs should use pooling of many diverse system results

answer: True

- this question is referring to the pooling-process in labeling ie. with mechanical turk where you're creating a cutoff-set to reduce the labor required to label your data.
- "… if we use a diverse pool of different systems, we can then even reuse those pool candidates and this gives us confidence that we have at least some of the relevant results in those pooling results. It allows us to drastically reduce the annotation time compared to conducting millions of annotations by hand." → see: https://github.com/sebastian-hofstaetter/teaching/blob/master/advanced-information-retrieval/Lecture%203%20-%20Closed%20Captions.md#26-pooling-in-information-retrieval

2022

question: Test collections should be statistically significant

answer: False

- the systems/models we build with the test-collections but not the test-collections themselves.
- statistical significance tests are used to verify that the observed differences between systems/models are not due to chance.
- "… we test whether two systems produce different rankings that are not different just by chance […]. Our hypothesis is that those systems are the same and now we test via a statistical significance test on a per-query basis" → see: https://github.com/sebastian-hofstaetter/teaching/blob/master/advanced-information-retrieval/Lecture%202%20-%20Closed%20Captions.md#29-statistical-significance-i
- see hofstätters only work on test-collections: https://arxiv.org/pdf/2208.06936

---

question: The quality of a test collection is measured with the inter-annotator agreement

answer: False

- the degree of agreement among raters != test-collection quality
- "We can measure the label quality of annotators based on their inter-annotation agreement" → see: https://github.com/sebastian-hofstaetter/teaching/blob/master/advanced-information-retrieval/Lecture%203%20-%20Closed%20Captions.md#25-evaluate-annotation-quality
- see: https://en.wikipedia.org/wiki/Inter-rater_reliability

question: A word-1-gram that we use when training Word2Vec is also considered as a word-n-gram

answer: False

- 1-gram != n-gram
- word2vec generates a single embedding for each word by learning to either guess the word from its surroundings or the other way around.
- you do have to train it with more than a single word and pass in a window size, but the word it's being trained to reconstruct is always a 1-gram / unigram.
- don't confuse this with CNNs that generate n-gram representations (a single embedding for n words)
- "1-Word-1-Vector type of class which includes Word2Vec" → see: https://github.com/sebastian-hofstaetter/teaching/blob/master/advanced-information-retrieval/Lecture%204%20-%20Closed%20Captions.md#10-word-embeddings
- see: https://radimrehurek.com/gensim/models/word2vec.html#usage-examples

---

question: ColBERTer achieves state-of-the-art performance on the MS Marco dev set

answer: True

- "When trained on MS MARCO Passage Ranking, ColBERTv2 achieves the highest MRR@10 of any standalone retriever." → see: https://ar5iv.labs.arxiv.org/html/2112.01488

---

key words students used when discussing other questions:

- exact matching components, storage, lexical match bias and how to apply it
- stopwords, tokens ← BOW, embedding ← 2-way dimension reduction
- effect size (based on: https://dl.acm.org/doi/pdf/10.1145/3269206.3271719)
- "In ColBERTer multiple retrieval workflows are represented. Which one is employed for the effectiveness comparison in Table 5" ← Retrieve CLS then refine BOW^2
  - see: https://github.com/sebastian-hofstaetter/colberter
  - see: https://arxiv.org/pdf/2203.13088
- "Assign each retrieval model its advantage of desirable properties for a retrieval model compared to the other models."
  - TK combines Transformers with kernel-pooling
  - BERT_CAT ← Effectiveness
  - TK ← Interpretability + Effort moved to indexing
  - BERT_DOT
  - ColBERT
  - see: https://discovery.ucl.ac.uk/id/eprint/10119400/1/Mitigating_the_Position_Bias_of_Transformer_Based_Contextualization_for_Passage_Re_Ranking.pdf
  - see: https://ecai2020.eu/papers/82_paper.pdf

2019

question: What are the differences between Matchpyramid and KNRM?

answer:

- matchPyramid
  - i. compute match-matrix through cosine-similarity for all query-doc-combinations
  - ii. apply 2D convolution kernels on matrix each learning a different feature
  - iii. determine final score (as float) with a feed-forward neural net
- (c)knrm = (convolutional) kernel based neural ranking model
  - roughly as effective as matchPyramid but a lot faster
  - i. apply cnn-kernels to encode multiple words into a single embedding (n-gram embedding) → only in conv-krnm variant
  - ii. compute match-matrix through cosine-similarity for all query-doc-combinations
  - iii. apply radial-basis-function kernel on all documents and get sum

question: What would the precision-recall-curve of an ideal re-ranker look like?

answer:

- in practice precision and recall are inverses of one another: improving recall (complreteness) typically comes at the cost of reduced precision (correctness), because you're likelier to make more mistakes as you retrieve more data.
- ideally we'd like to see high precision at low recalls, gradually decreasing as recall increases. and after all relevant documents have been retrieved we usually have diminishing returns and a sharp drop in precision.
- therefore the ideal precision-recall curve would start at the top-left corner of the plot (high precision, low recall) and move diagonally downwards to the right, indicating high precision at lower recall levels and a sharp drop in precision once all relevant documents have been retrieved.

but if we'd ignore the inverse relationship between precision and recall (because we're being idealistic) then we would have perfect precision until all relevant documents have been retrieved and vertically drop to 0 precision.

---

question: Why are low-frequency words an issue for information retrieval but not so much for other tasks like information categorization?

answer: the question is unclear. but one way to argue could be:

- information retrieval = looking for similarity between query and relevant documents
- classification = looking for distinct features to maximize dissimilarity