Natural Language Processing and Search Course Exercises

João Magalhães

Universidade NOVA de Lisboa

December 10, 2022

Text Parsing and Tokenization

1) Are the following statements true or false?
   a) In a Boolean retrieval system, stemming never lowers precision.
   b) In a Boolean retrieval system, stemming never lowers recall.
   c) Stemming increases the size of the vocabulary.
   d) Stemming should be invoked at indexing time but not while processing a query.

2) Consider the tf-idf term weighting.
   a) What is the *idf* of a term that occurs in every document? Compare this with the use of stop word lists.
   b) Can the *tf-idf* weight of a term in a document exceed 1?

3) Assume a biword index. Give an example of a document which will be returned for a query of *New York University* but is actually a false positive which should not be returned.

4) Consider the Vector Space Model and classic Language Models:
   a) How is the tokenização done in the VSM?
   b) How is a word token represented as a vector?
   c) Is it possible to compute the similarity between words in the VSM? Explain why.
   d) How can you represent a word that has never been seen in the vector space model?
   e) How is the word sequence guaranteed in the VSM/LM? Explain.

5) Transformer Language Models take a data-driven approach to text tokenization. Please explain:
   a) How is BPE tokenization achieved?
   b) Stemming is a common element of whitespace tokenizers. Should it be also included in the BPE tokenizers?
   c) How to represent a word that was never seen with BPE?
   d) How to represent a word that was never seen with the whitespace tokenizer?

Evaluation

6) Consider an information need for which there are 4 relevant documents in the collection. Contrast two systems run on this collection. Their top 15 results are judged for relevance as follows (the leftmost item is the top ranked search result):

   System 1 R N R N N N N N R N N N N N R

   System 2 N R N N R N N N R N N R N N N

   a) What is the MAP of each system? Which has a higher MAP?
   b) Does this result intuitively make sense? What does it say about what is important in getting a good MAP score?
   c) What is the Recall after 10 retrieved documents of each system?
   d) Plot the precision-recall curve for both systems. Interpret the curve.

7) Below is a table showing how two human judges rated the relevance of a set of 12 documents to a particular information need (0 = nonrelevant, 1 = relevant). Let us assume that you've written an IR system that for this query returns the set of documents {4, 5, 6, 7, 8}.

docId	Judge 1	Judge 2
1	0	0
2	0	0
3	1	1
4	1	1
5	1	0
6	1	0
7	1	1
8	1	0
9	0	1
10	0	1

   a) Calculate precision, recall, and F1 of your system if a document is considered relevant only if the two judges agree.

8) Consider two ranking algorithms that for the same query produced the two following ranks:

S1: d4 d3 d5 d8 d2 *and* *S2: d3 d9 d5 d6 d1*

a) Assuming that the relevant documents are d9, d1, d3 and d4, compute the precision and recall values of each system.

b) Assuming that the multi-value relevance judgments of documents are d9=1, d1=1, d3=3 and d4=2, assess and compare the two ranks with the appropriate metric.

c) Assume no relevance judgments and compare the two systems.

Language Models and Retrieval models

9) Consider the table of term frequencies for 3 documents denoted Doc1, Doc2, Doc3. The collection contains 750,000 documents in total.

	docFrequency	**Doc1**	**Doc2**	**Doc3**
car	18,165	27	4	24
auto	6,723	3	33	0
insurance	19,241	0	33	29
best	25,235	14	0	17
TOTAL	-	345	430	370

a) Compute the tf-idf weights for the terms car, auto, insurance, best, for each document.

b) Compute the rank of the three documents for the query “auto insurance” on the vector space model.

10) Consider the two following documents:

d1 : Jackson was one of the most talented entertainers of all time

d2: Michael Jackson anointed himself King of Pop

a) Using a BM25 retrieval model determine which document is more relevant to the query q= “Michael Jackson” (consider b = 0.75 and k = 1.5).

b) Using a language model with Jelinek-Mercer smoothing determine which document is more relevant to the query q= “Michael Jackson” (consider $\lambda = ½$)

c) Using a language model with Dirichlet smoothing determine which document is more relevant to the query q= “Michael Jackson” (consider $\mu = 100$)

11) Show that models resulting from Dirichlet smoothing can be treated as probability distributions. That is, show that $\sum_t M_d^u(t) = 1$.

12) Consider the Language Model with Jelineck-Mercer smoothing:

    a) What is the role of the lambda factor?

    b) Why is smoothing necessary in Language Models?

13) Suppose we have a collection that consists of the 4 documents given in the below table.

docID	Document text
1	click go the shears boys click click click
2	click click
3	metal here
4	metal shears click here

    Build a query likelihood language model for this document collection. Assume a mixture model (LMJM) between the documents and the collection, with both weighted at 0.5. Maximum likelihood estimation (mle) is used to estimate both as unigram models.

    a. Work out the model probabilities of the queries “click”, “shears”, and hence “click shears” for each document, and use those probabilities to rank the documents returned by each query.

    b. What is the final ranking of the documents for the query click shears?

14) You have discovered that documents in a certain collection have a “half-life” of 30 days. After any 30-day period a document’s prior probability of relevance p(r|D) is half of what it was at the start of the period. Incorporate this information into LMJM. Simplify the equation into a rank-equivalent form, making any assumptions you believe reasonable.

15) Write one sentence each describing the treatment that the LM with Jelinek-Mercer smoothing gives to each of the following quantities. Include whether it is present in the model or not and whether the effect is raw or scaled.

    c. Term frequency in a document

    d. Collection frequency of a term

    e. Document frequency of a term

    f. Length normalization of a term

Learning to Rank

16) The learning-to-rank approach aims to learn a ranking function that best ranks documents for each query.

    a) What is the input to a learning-to-rank algorithm?

    b) What is the role of the coefficients of the learning-to-rank model? What do they say about the role of each input feature?

    c) Learning to rank training data per label is highly skewed. In which ways can you compensate for the data unbalanced?

    d) The pointwise approach to learning-to-rank aims to rank documents by their importance to the input query in which way?

17) Rank fusion methods combine ranks in different manners. Compute the fused ranks for the following three lists with the CombSUM, CombMNZ, BordaFuse and RRF.

Rank 1 (id/score)	**Rank 2 (id/score)**	**Rank 3 (id/score)**
D3 / 0.5	D3 / 0.8	D9 / 0.9
D4 / 0.2	D8 / 0.8	D3 / 0.8
D2 / 0.19	D2 / 0.8	D1 / 0.7
D5 / 0.18	D1 / 0.5	D8 / 0.6
D6 / 0.07	D5 / 0.4	D2 / 0.5
D1 / 0.05	D6 / 0.32	D5 / 0.4
D7 / 0.01	D9 / 0.31	D6 / 0.3
D9 / 0.01	vzD7 / 0.30	D7 / 0.2

Contextual Embeddings and Self-Attention

18) Consider the self-attention mechanism introduced by the Transformer.

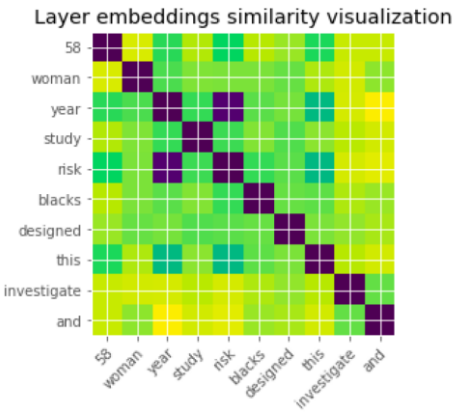
    a) Explain what is self-attention?

    b) How is the attention between two words computed?

    c) The attention value between two tokens is used in which way?

    d) How are the output embeddings of the self-attention layer computed?

19) Consider the contextual embeddings computed by the Transformer encoder.



a) In the Transformer architecture, what is it that the embeddings of layer 0 represent?

b) In the Transformer architecture, what is it that the embeddings of layer 12 represent?

c) The similarity matrix depicted next illustrates the similarity between layer 0 and layer 12. Why is the diagonal close to zero?

d) How is the text tokenization done in the Transformer architecture? Is it possible to represent a word that has never been seen before?

e) What is the role of the CLS token?

20) Consider the contextual embeddings as computed by the Transformer.

a) Does the BERT Transformer maintains sequence information? If yes, how?

b) What is the neighborhood of a word embedding vector in layer 0?

c) How can you compute the similarity between words in the Transformer?

d) How to interpret the token embeddings visualization?

Question Answering

21) Consider the common QA processing pipeline.

a) In the QA architecture that was discussed in the course, what type of data is required? How should the data be pre-processed?

b) Identify the two stages of a QA pipeline and explain their function.

c) How does the Transformer solves each part of the QA pipeline?

Live Systems Development

22) Suppose that your boss asks you to develop a test collection to replace an existing corporate search engine. The company wants the test collection to be useful for the next 2-3 years.

a) Describe how you would build or acquire the different test collection components, and how much data is required for each component.

b) Detail the process of selecting queries and acquiring the corresponding relevance judgments. Your answer needs to be practical, i.e., no magic, and your budget isn't infinite.