Name: _____ Number: _____

Information Retrieval

Departamento de Informática
NOVA FCT

25 January 2021
Duration: 1h30, ~15 mins per question.
Justify your answers.

1. Consider a search need for which there are 5 relevant documents in the collection. Two information retrieval systems returned 10 documents that were judged according to their relevance as follows:

Position:	1	2	3	4	5	6	7	8	9	10
System 1:	N	N	N	N	N	R	R	R	R	R
System 2:	N	N	N	N	N	N	N	R	N	N

   a. Compute the retrieval precision and recall for every position of the rank.

b. Draw the precision-recall curve for both systems. Compute the Average Precision of each system. Relate the AP metric to the precision-recall curves.

2. 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.

3. 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 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"?

4. Rank fusion methods combine ranks, such as the ones illustrated below, in different manners.

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

a. Considering only the top 5 documents of the three ranks, compute the fused ranks for the three lists above with CombSUM.

b. Using the Reciprocal Rank Fusion method, select a reasonable value for k and combine the top 5 documents of the three above ranks.

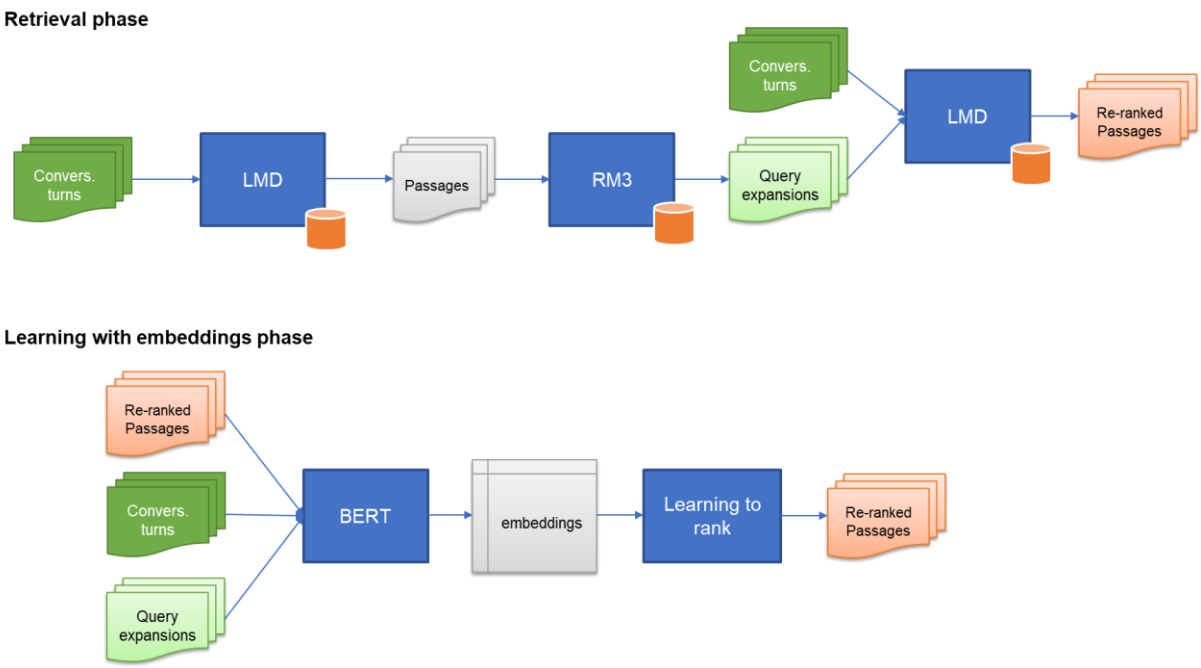
5. You have discovered that documents in a certain collection have a "half-life" of 30 days. After any 30-day period a document's prior probability of relevance $p(r|D)$ is half of what it was at the start of the period. Incorporate this information into LMD. Simplify the equation into a rank-equivalent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

6. Consider question-answering systems and the following sentence:

Painkillers that don't upset stomach

Discuss the type of documents that a Q-A system would return versus the documents a Search system would return. Explain the text processing and analysis differences.

7. Consider Conversational Search agents and the following architecture that was suggested for your second project:



a. Justify the rationale supporting each one of the components.

b. Detail the different components of the Learning with embeddings phase.