Assignment 7 (20 ptos)

Due Date: March 28th, 23:59pm, BR time.

Evaluate a set of query-document pairs provided to you. Each one of these query-document pairs was produced by a search engine developed in this course. Each pair query-document should be evaluated through a score assigned to it, a score varying in the range [1-5] where 1 indicates that the document is non-relevant to the query and 5 indicates that the document is most relevant to the query. Consider the query q given by

q : hotel na regiao da pampulha

Then, the meaning of each eval score associated with the query-document pair [q, d_j] is as follows:

1. Document d_j is *not* relevant to query q and has no relation to it
    a. This is the case, for instance, in which d_j is a hospital
2. Document d_j is *not* relevant to query q but I understand its relation to the query
    a. This is the case, for instance, in which d_j is a hotel located in a town other than Belo Horizonte
3. Document d_j is *weakly* relevant to query q
    a. This is the case, for instance, in which d_j is a hotel located in Belo Horizonte but distant from Pampulha, a hotel located in downtown
4. Document d_j is relevant to query q
    a. This is the case of a hotel in Pampulha or nearby Pampulha such as, for instance, Quality Hotel Pampulha
5. Document d_j is *most* relevant to query q
    a. This is the case of a hotel in Pampulha, located by the lake and considered by many to be one of the best in the region, such as San Diego Suites Pampulha

This evaluation process should be completed by next friday,

March 19th, 23:59pm.

Once this evaluation process is completed, you will be provided with a summary of the results, i.e., a set of relevance judgements for the 20 example queries we selected in TP 6. For each query-document pair [q, d_j] these relevance judgements will look as

- 5 : document is most relevant to query
- 4 : document is relevant to query
- 2 : document is not relevant to query

We will consider that 2 relevant judgements signal "relevance". Otherwise, we will assume that the document is *not* relevant to the query.

Using these relevance judgements, compute recall/precision curves for these 20 queries, using the top 5 results returned by the query processor you have implemented. Draw two separate precision/recall curves:

1. one using just the Vector Model and
2. a second one using the Vector Model combined with Page Rank.

Compute also a bar diagram indicating the number of relevant documents for each of our 20 queries.

Write a report analyzing and detailing your findings.