CS 4960/6550: Introduction to Information Retrieval

Assignment 2

Instructor: Qingyao Ai *Teaching Assistant: Tao Yang*

Question 1 (10 points)

What is the fundamental difference between language modeling approaches and classical probabilistic models in IR? Discuss.

Question 2 (18 points)

(a) What is the unigram language model probability of a term in a document? (proof required)
(b) Using maximum likelihood estimation, compute the probability of "Information", "resources" and "system" (with normalization, no stopping, and no stemming) in the following sentence:

Information retrieval is the process of obtaining information system resources that are relevant to an information need from a collection of information resources..

Question 3 (10 points)

What are the reasons for language model smoothing? Discuss.

Question 4 (14 points)

Consider a corpus $\boldsymbol{C}$ with 200 documents, and the average length of documents in $\boldsymbol{C}$ is 200. and terms "information", "retrieval", "company" that have appeared 100, 40, and 60 times in $\boldsymbol{C}$, respectively. Suppose that there is a query "information retrieval company" and a document that only contains one sentence:

According to Wikipedia, Information Retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources

Compute the query log-likelihood score for this query-document pair with and without JM smoothing ($\lambda = 0.5$). In this question, only consider normalization; no stopping or stemming is required. Discuss what you can learn from this example.

Question 5 (12 points)

Among additive smoothing, JM smoothing, and Dirichlet prior smoothing, which one is expected to perform best and which one is expected to perform worst? Why? (assume the parameters of all models are well tuned)

Question 6 (26 points)

Suppose that query q has 6 subtopics (i.e., T_1, T_2, T_3, T_4, T_5, and T_6), and we have 8 candidate documents for this query. Their relevance to each subtopic is shown in Table 1:

Consider the following rank lists:

- List 1: d_1, d_5, d_3, d_4, d_2, d_6, d_7, d_8
- List 2: d_1, d_5, d_8, d_4, d_7, d_6, d_2, d_3

Table 1: The relevance of different documents with respect to the subtopics of query q. In the table, relevant topic-document pairs are indicated with symbol "X".

Document	T_1	T_2	T_3	T_4	T_5	T_6	Total number of relevant topics
d_1				X			1
d_2							0
d_3		X	X		X		3
d_4							0
d_5	X					X	2
d_6	X						1
d_7							0
d_8	X	X		X		X	4

Compute the α-nDCG ($\alpha = 0.5$) of those lists (8 points for each). Let the total number of relevant topics in each document indicate the relevance label of that document, then compute nDCG for each list. Discuss what you observe from the results of α-nDCG and nDCG, and explain why (4 points).

Question 7 (10 points)

It has been observed that improving diversity often leads to improvement in relevance. What is the reason?