Information Retrieval (CS60092)

End-Semester Examination **Maximum Marks 70**

This question paper has 2 pages and 7 questions.

Note: There are no clarifications. In case of doubt, you can take a valid assumption, state that properly and continue. Please answer PART-A in the first half of the answer sheet (initial 5-6 pages) and PART-B in the later half.

PART - A

Question 1: Suppose your corpus contains the following 3 sentences:
S1: vector space model makes the same assumption
S2: cleaner state of assumption than vector space
S3: language model views document as generative model
Use a mixture language model to rank these document as per relevance to the query – 'model assumption'. Use lambda = 0.5. Do not remove the stop words. **[5]**

Question 2: Suppose you have computed embeddings of all the vocabulary terms using word2Vec. Let $\boldsymbol{w_x}$ denote the embedding for a word **x**, which is L2 normalized. Express how would you use this to incorporate query expansion while computing $P(q|M_d)$ as per the language modeling, where M_d corresponds to the document language model $P(.|M_d)$. Let V denote all the terms in the vocabulary. **[3]**

Question 3: Consider a web-corpus where the average length of the documents is 200 tokens. Suppose you find that there are 2500 different terms in the first 500 documents, and 20,000 different terms in the first 40,000 documents. Assume that the total number of webpages in the corpus is 2 billion. Use Heap's law to approximate the vocabulary size of the collection. **[3]**

Question 4: Suppose you want to index corpus from a language for ranked retrieval, for which average word size (number of characters) per token is 12, and the average word size per term is 16. Also, it is not very uncommon to have words having 28 characters in this language, so you assign 28 bytes per term for the dictionary storage.
(a) Assume that in your corpus, you have 5 million unique words. Estimate the size of dictionary, while using the standard array of fixed width entries.
(b) How much compression can you achieve on this, if you store dictionary as a (long) string, with pointer to the next word showing end of the current word? [Report the final size of the dictionary]
(c) On top of that, suppose you use blocking with 8 strings in a block? What would be the additional saving? [Report the size after this step] **[1+2+3 = 6]**

Question 5: Answer the following questions. The answers should be brief, to the point only. Essays will not be graded. **[6 x 3 = 18]**
(a). Use of idf in vector space model helps in assigning a higher weight to the rare terms in comparison to the common terms. How is a similar effect achieved in language model? Justify your answer.
(b). Explain briefly how query logs can be used for query expansion.
(c). Suppose you are using BSBI for index construction, and you have sorted *n* equally sized blocks. The final step is to sort all these blocks. Suppose you can fit only 2 of these blocks in the memory at a time,

while there are 20 such blocks. Would you prefer to use a binary merge or multi-way merge? Justify your choice.

(d). Suppose you are using gamma codes for compressing the posting lists. Suppose your posting list looks like: [18|55|296]. What will be the corresponding compressed list using gamma codes?

(e). Express an ordering in terms of smallest to largest reduction in size of [dictionary (A), non-positional index (B) and positional index (C)] when you perform i). case folding and ii). Removal of 100 stop words. Assume the collection is reasonably large. Justify your answer briefly. Your answer should be in form of an ordering, e.g., A < B < C, which corresponds to A undergoing the smallest reduction in size.

(f). Why is document frequency preferred for computing idf over the collection frequency? Further, idf is useful only for queries with multiple words but does not affect ranking for single word queries. Justify.

PART - B

Question 6: Consider hashtags on Twitter. The problem is to predict the popularity of hashtags where popularity is measured as the number of tweets in which the hashtag has appeared so far (i.e., the frequency of usage, say *F*, of the hashtag). In this context, you are tasked to design a feature based machine learning framework to predict hashtag popularity. Answer the following questions. **[(2+3)+2+6+6+6=25]**

a) Pose the above as a regression problem (draw a schematic). What are typically the inputs and the output of the regression model?
b) How will you typically choose the training and the test data so that there is no information leakage?
c) Identify three features that you can extract from the hashtag name itself (i.e., hashtag content features). Justify each of your choice in one line.
d) Identify three features that you can extract from the collection of tweets in which the hashtag appears (i.e., tweet content features). Justify each of your choice in one line.
e) Suppose you know that your collection of hashtags whose popularity you wish to predict are typically of the form #10ThingsAboutMe, #20ThingsAboutMe, #FlashbackMonday, #FlashbackFriday, #5ThingsIHateAboutYou etc. These types of hashtags are called *idioms*. Can you identify three special hashtag content features for idioms? Justify in one line each of your choice.

Question 7: Let us consider the following reference summary and the system-generated summary (candidate summary). For the candidate summary, compute the ROUGE-1 and ROUGE-L recall values. **[4+6 = 10]**

Reference summary:

Skyfall is the twenty-third spy film in the James Bond series, produced by Eon Productions for MGM, Columbia Pictures and Sony Pictures Entertainment. Directed by Sam Mendes, it features Daniel Craig's third performance as James Bond and Javier Bardem as Raoul Silva, the film's villain. Skyfall will also be the first James Bond film to be released in IMAX venues.

Candidate summary 1:

film to James Bond also be Skyfall will the IMAX venues first be released in . Skyfall is the twenty-third spy film in the James Bond series, produced by Eon Productions for MGM, Columbia Pictures and Sony Pictures Entertainment. Directed by film's villain Sam Mendes, it features Javier Bardem third performance and as Raoul Silva, the Daniel Craig's as James Bond.