114726

IT550 - Introduction to Information Retrieval
End Semester Examination

25 November 2024

Total Marks: 20 **Total Time: 60 Mins**

Answer the following questions [2*5 = 10 Marks]

Limit your answers to at most 5 sentences. Answers longer than 5 sentences will not be graded

1. Figure 1 below shows the PCA for a dataset. For each of the following clustering algorithms, explain in **two sentences** whether the algorithm will perform well on this dataset, and provide a brief reason for your answer. Assume no prior knowledge about the number of clusters in the data.

   1. K-Means
   2. DBScan
   3. HDBSCan

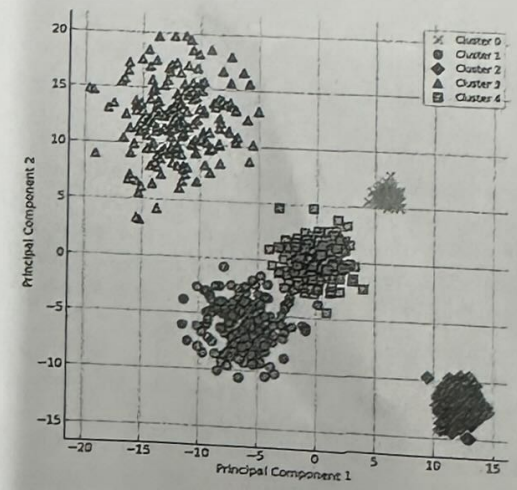

Figure 1

2. A query with limited context like *"What is a bat?"* is ambiguous because documents about the mammal bat and the bat used in cricket are both equally correct. One way to perform disambiguation is to use the queries used just before the ambiguous query a.k.a. session history.

   Imagine with the ambiguous query above you were also provided session history in the form of two more queries "What is cricket?", "How many players are in a cricket match?". Describe a non-neural, unsupervised approach in which the session history can be used to improve the query performance. Any form of pre-trained models are also not available.

3. A generalized language model can be represented by the equation below. It is used to computed the likelihood of a document given a query word w. This is indicated by the LHS of the equation. The four terms in the RHS of the equation denote the estimated probability of the given word (w) based on the given document, the estimated probability of the words similar to w (w') based on the given document and the probabilities of w and w' given the entire corpus. As evident, not only the occurrence of the word w in a document or corpus but also its semantically similar words play a role in the final probability for the document. The variables α, β and λ define the importance of each of the terms in the final score. This equation is a form of query expansion, where w is the original query and $w' \in W$ are the expansion terms.

$$P(d|w) = \lambda P(w|d) + \alpha \sum_{w'} P(w'|d) \cdot sim(w, w') + \beta \sum_{w'} P(w'|C) sim(w, w') + (1 - \lambda - \alpha - \beta) P(w|C) \quad (1)$$

Imagine you are searching for some highly technical information, where you do not know the exact technical terms, but can describe the query in simple words. For example, instead of the intended concept "Photovoltaic cell" you type in "solar panel cell".

Under this scenario if you were allowed to choose α, β and λ, what relative ordering of the values would you choose and why?

4. Explain in not more than 5 sentences how LexRank for text summarization works and what changes will you *make to* use it with pre-trained text embeddings.

5. Which of the following problems would benefit from a neural network that can learn the similarity between a pair of inputs? Give a True / False answer for each

   A. Query Auto-Completion
   B. Searching a QnA website
   C. Spell Correction
   D. Cross-Lingual search (English query, Hindi document)

Design an end-to-end system that solves the following problems

6. You have a word2vec-like embedding model that doesn't seem to handle spelling mistakes well. How will you adapt the original model, training process, input vectors and output vectors to create embeddings that are robust and less susceptible to spelling mistakes? Explain in brief the changes in one or more of these four.
   Assume you have access to training data that was used to train the original embeddings. **[3 Marks]**

7. You are designing a search engine for legal case retrieval. The user will use the platform to search for cases that are similar to a given legal situation. Sample query is below. What change will you make to the usual pre-processing, indexing and search pipeline and what evaluation metric will you use? Explain each of the four aspects in brief (At most 4-5 sentences each). **[4 Marks]**

   Sample Query: *My grandfather, who passed away two years ago in 2022, left behind a 10-acre farmland and a residential house in the village of Ranipura. He did not leave a will, though he verbally expressed his desire to keep the property within the family. My father and uncle, who have lived in Ranipura and maintained the property for over 20 years, argue they deserve the largest share due to their efforts and financial contributions. However, my two aunts, who moved to the city decades ago, insist on selling the entire property and splitting the proceeds equally among the siblings.*

   To complicate matters, distant relatives—descendants of my grandfather's brother—are claiming that the farmland belonged to their side of the family originally and was only entrusted to my grandfather in the 1970s for temporary use. They now demand its return. My aunts have threatened to file a legal suit to force a sale, while my uncle and father refuse to consider this option. I want to understand how inheritance laws apply to such disputes, whether the distant relatives have a valid claim after all these years, and whether mediation or legal intervention is the best way forward to protect the family's rights.

8. Given the Precision-Recall Curve shown in the figure with three points A, B and C. Give one typical application each where A, B and C will be the desired results. Just mention the application. You can include a 1-2 sentence explanation if *really* required. **[3 Marks]**

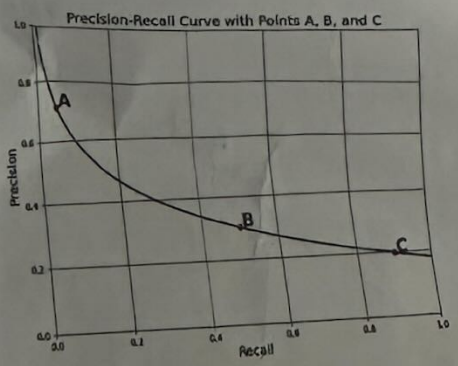

Figure 2