**Sample Questions (IRWS)**

**Q1. How does stemming typically affect recall in Boolean document retrieval?**

**Q2. Comment on the statement “Pseudo-feedback always increases precision and recall”**

**Q3. Why do commercial web search engines typically not provide relevance feedback functionality? Give at least 2 reasons.**

**Q4. The “Vector Space Model” is one of the most commonly used information retrieval models. Explain how it works and your answers should include how the documents and queries are represented, how the weights are calculated, and how the similarity between a query and a document is calculated.**

**Q5. Briefly explain the two main approaches in query expansion. State the main problems or**

**limitations of each approach and any solutions you may have for solving the problems.**

**Q6. TF-IDF is a typical way of weighting terms for text representation. Discuss its advantages and disadvantages.**

**Q7. Term‐document incidence matrices are most often not stored as 2D‐arrays, but as another structure. Why?**

**Q8. What is the difference between Boolean retrieval and ranked retrieval? Why are most publicly available search engines using ranked retrieval nowadays?**

**Q10. Consider a collection with the documents:**

document1: the way to the school is long and hard when walking in the rain

document2: the rain has not stopped in days and the school has closed

query: school closed rain

Answer the following questions using the documents and query above.

1. What would be the similarity score of the query with each document given above using Jaccard coefficient, if You assume there are no stop words, for this part of the question.
2. The list of stop words is (the, to, is, and, in, has, not )
3. The query is converted to a unit vector using tf-idf weighting [wt,d = tft,d \* log10(N/dft)] and Euclidean normalization. The documents are converted to unit vectors using raw tf weighting and Euclidean normalization. Compute the cosine similarity of the query with each document.

**Q11. Give pros and cons of local and global approach of query expansion.**

**Q12. What is document preprocessing?**

**Q13. Which are the parameters used for retrieval performance evaluation?**

**Q14. Explain Inverted Index file. How it can be used in Information Retrieval?**

**Q15. How DBMS is different for information retrieval system?**

**Q16. Explain any two of classical IR models with example.**

**Q17. Why are term frequency and inverse document frequency used so often in document scoring functions?**

**Q18. How do stopping and stemming reduce the size of an inverted index?**

**Q19. You have developed a new method for parsing documents that uses semantic information to decide which sentences to index and which to skip. How would you determine whether your method produces better retrieval results than indexing every sentence?**

**Q20. The “Vector Space Model” is one of the most commonly used information retrieval models. Explain how it works and your answers should include how the documents and queries are represented, how the weights are calculated, and how the similarity between a query and a document is calculated.**

**Q21. Recent research has suggested that Wikipedia is useful for query expansion and there are many different ways of integrating Wikipedia into query expansion systems. In your opinion, how would you use Wikipedia to improve the current query expansion systems?**

**Q22. Assume that in a document collection there are a total of 5 documents. Let the indexing vocabulary consist of 3 index terms. Assuming that index terms *t1*, *t2* and *t3* appear, respectively, in 3, 4 and 2 documents, determine their *idf* weights.**

**Q23. What is the difference between stemming and lemmatization? What are the advantages and disadvantages with each method?**

**Q24. The following 20 ranked results have been returned as a response to a query. Since we are experts in the area, we find results 2, 6, 8, 9, 10, 12, 15, 16, 17, 18, and 20 relevant.**

**We know that there are in total 20 relevant documents in the collection.**

**(a) Draw the precision‐recall curve for the first 20 results.**

**(b) What is *precision at 10* for this list of results?**

**Q25. Consider the query "Michael Jackson", and the two following documents:**

**1) The Wikipedia article about Michael Jackson, which begins:**

**”Michael Joseph Jackson (August 29, 1958 – June 25, 2009) was an American recording artist, dancer, singer‐songwriter, and philanthropist. Referred to as the King of Pop, Jackson is recognized as the most successful entertainer of all time by Guinness World Records. His contribution to music, dance and fashion, along with a much‐publicized personal life, made him a global figure in popular culture for over four decades. The eighth child of the Jackson family, he debuted on the professional music scene along with his brothers as a member of The Jackson 5 in the mid‐1960s, and began his solo career in 1971…".**

**2) The document whose entire contents is: "Michael Jackson"**

**If the documents are represented by their tf‐idf weights, which document would receive the higher cosine score with the query? In the light of this example, discuss the use of tf-idf as a measure for ranking search results. Which improvements would you suggest?**