NATIONAL UNIVERSITY OF SCIENCE AND TECHNOLOGY

FACULTY OF COMMUNICATION AND INFORMATION SCIENCE

DEPARTMENT OF RECORDS AND ARCHIVES MANAGEMENT

TECHNIQUES AND APPLICATIONS IN INFORMATION RETRIEVAL

IIM 1203

Main Examination Paper

February 2025

This examination paper consists of 3 pages.

Time Allowed : **3 hours**

Total Marks : **100**

Examiner's Name : **Mr D. T. Sigauke**

INSTRUCTIONS

1. Candidates must answer any **four (4)** questions.
2. Each question carries a total of **25 marks**.

MARK ALLOCATION

QUESTION	MARKS
1.	25
2.	25
3.	25
4.	25
5.	25
6.	25

1. a) Give the classical definition for information retrieval as proposed by Manning et al., (2008). **[5 marks]**

   b) Describe four (4) document pre-processing tasks that prepare a text document for an information system. **[8 marks]**

   c) What is meant by the "semantic gap" in information retrieval? **[3 marks]**

   d) Explain Boolean, Vector Space and Probabilistic retrieval as classical approaches to information retrieval. **[9 marks]**

[Total 25 marks]

2. a) Explain the *relationship* between precision and recall? **[6 marks]**

   b) Refer to the following tables and answer the questions that follow.

X	Relevant	Not relevant
Retrieved	40	360
Not retrieved	10	1090

Y	Relevant	Not relevant
Retrieved	15	15
Not retrieved	35	1435

   i) Use a truth table to help calculate the precisions and recalls of System X and System Y respectively, based on their respective formulas. **[8 marks]**

   ii) Based on the values you have discovered which system is most effective? **[2 marks]**

   c) Both precision and recall need to be taken into account when evaluating retrieval systems. Why is it not sufficient to pick one and use only that? **[6 marks]**

   d) Give an alternative formula which can be used to evaluate information retrieval. **[3 marks]**

[Total 25 marks]

3. Discuss the ethical, safety and privacy concerns that have arisen with the use of modern search engines and question answering (QA) systems. **[25 marks]**

4. a) Distinguish data retrieval from information retrieval based on five (5) facets associated with the type of data being retrieved. **[15 marks]**

   b) Use the following inverted index to answer the subsequent questions.

i. What two (2) components make up an inverted index. **[2 marks]**

ii. Find the query result for "BRUTUS AND CAESAR". **[2 marks]**

iii. Display the query result of "BRUTUS AND CALPURNIA NOT CAESAR". **[2 marks]**

iv. List any (3) sorting techniques that can be applied to an inverted index. **[3 marks]**

[Total 25 marks]

5. a) What is a large language model (LLM)? **[3 marks]**

b) Compare sparse retrieval with dense retrieval referring to the type of indexes that they develop. **[8 marks]**

c) Explain the advantages and limitations of self-retrieval by large language models (LLMs) as an innovation to generative retrieval. **[14 marks]**

[Total 25 marks]

6. Read the following extract of a paragraph from a book entitled *Invisible Search and Online Search Engines by Haider and Sundin (2019:11)* and answer the questions that follow.

The elephant in the room: Google

There is no way around acknowledging that, what Ken Hillis and his colleagues (2013, p. 1) aptly describe as an "astonishing naturalization of the process of search in everyday life" and what we choose to call the concurrent *mundane-ification* of search, happened in tandem with the rise of one specific commercial search engine, Google. Each minute more than three million queries are submitted to Google Search. Globally in the second decade of the 2000s the market share of Google Search constantly lies between 85% and 90% (StatCounter n.d. a). Notable exceptions are Russia and especially China (StatCounter n.d.b). In both countries other large commercial search engines exist – with their own indexes and search tools: Baidu in China and Yandex in Russia. This has, of course, to do with censorship and related issues of information control, especially in China, but also with character sets used to write in these country's languages. However, both Yandex' and Baidu's interfaces and ways of displaying results are noticeably inspired by Google. All available statistics point in one direction and that is that Google dominates all aspects of the search engine market, largest share of searches on all devices, most visitors to its site, most users of its browser, Chrome, which in effect is a search engine, and so on. This data is, of course, time sensitive, but, in order to understand the role of search engines as a social and cultural phenomenon Google has to be taken seriously and has to be situated in its specific historic context.

a) Define a search engine. **[3 marks]**

b) Explain why Google is described as the "*elephant in the room*"? **[8 marks]**

c) Give any four (4) alternative search engines besides Google. **[4 marks]**

d) In your opinion, do you agree that Google has to be taken seriously as a social and cultural phenomenon? Give examples to help support your answer. **[10 marks]**

[Total 25 marks]

END OF PAPER