Universität Regensburg

Prof. Dr. U. Kruschwitz 26.7.2023

Written Exam

" Introduction to Information Retrieval"
SS 2023

You have 90 minutes to work on this exam. You should answer all questions. The total number of marks is 100. Make sure your submission includes your **name and registration number**.

Klausur
" Einführung in das Information Retrieval"
SS 2023

*Sie haben 90 Minuten Zeit. Beantworten Sie bitte alle Fragen. In der Klausur sind insgesamt 100 Punkte zu erreichen. Bitte geben Sie in Ihrer Arbeit **Namen und Matrikelnummer** mit an.*

Nachname, Vorname	
Matrikelnummer	

Question 1 / Aufgabe 1: **(40 marks / 40 Punkte)**

Basics / ***Grundlagen*****.**

Question 1.1 / Aufgabe 1.1: (10 marks / 10 Punkte)

Explain the importance of the factor *IDF* in the ranking function *BM25*.

Erklären Sie die Bedeutung des Faktors IDF im BM25-Algorithmus.

Question 1.2 / Aufgabe 1.2: (10 marks / 10 Punkte)

Words in textual documents and in query corpora tend to follow a Zipf distribution. Briefly discuss the implications of this on algorithms for *auto-completion*.

Wörter in Textdokumenten und Suchmaschinenanfragen folgen üblicherweise einer Zipf-Verteilung. Diskutieren Sie kurz, welche Auswirkungen das auf 'auto-completion'-Algorithmen hat.

Question 1.3 / Aufgabe 1.3: (10 marks / 10 Punkte)

Neural approaches have become very popular in Information Retrieval (IR). Briefly outline a BERT-based architecture that illustrates the benefits of such an approach.

Auf dem Gebiet des Information Retrieval (IR) sind neuronale Ansätze sehr populär geworden. Skizzieren Sie eine BERT-basierte Suchmaschinenarchitektur, die die Vorteile eines solchen Ansatzes illustriert.

Question 1.4 / Aufgabe 1.4: (10 marks / 10 Punkte)

Unlike neural approaches, the Boolean model of information retrieval has become much less popular these days. Briefly describe two use case scenarios in which key properties of the Boolean model might nevertheless be desirable.

Im Gegensatz zu neuronalen Ansätzen ist das Boolsche Model im Information Retrieval deutlich weniger populär geworden. Skizzieren Sie zwei Anwendungsbeispiele, bei denen Kerneigenschaften dieses Modells aber trotzdem wünschenswert sind.

Question 2 / Aufgabe 2: **(40 marks / 40 Punkte)**

Applications and Evaluation / ***Anwendungen und Evaluierung*****.**

Question 2.1 / Aufgabe 2.1: (10 marks / 10 Punkte)

Outline the typical steps that need to be performed by an enterprise search engine to match a user request against the documents stored in the system's database. Discuss how enterprise search might differ from Web search.

Skizzieren Sie die typischen Schritte einer Enterprise-Suchmaschine, um eine Nutzeranfrage mit der Dokumentensammlung zu vergleichen. Wie unterscheidet sich Enterprise-Suche von Websuche?

Question 2.2 / Aufgabe 2.2: (10 marks / 10 Punkte)

Discuss the applicability of the PageRank algorithm in an enterprise search setting. Start by sketching the algorithm for calculating the PageRank value of a page using a simple example.

Diskutieren Sie den Nutzen des PageRank-Algorithmus in der Enterprise-Suche. Skizzieren Sie dazu zunächst anhand eines Beispiels, wie der PageRank einer Seite berechnet wird.

Question 2.3 / Aufgabe 2.3: (10 marks / 10 Punkte)

Outline a search scenario in which you would prefer *A/B testing* over alternative evaluation approaches. Justify your answer.

Beschreiben Sie ein Szenario, bei dem Sie ‚A/B-Testing' alternativen Evaluationsansätzen vorziehen würden. Begründen Sie Ihre Antwort.

Question 2.4 / Aufgabe 2.4: (10 marks / 10 Punkte)

Several evaluation metrics have been developed to assess the quality of results returned by search engines. Two such measures are *precision* and *recall*. What can you say about precision and recall for queries for which no relevant documents exist in the collection? Discuss whether *discounted cumulative gain* or *mean reciprocal rank* might or might not be suitable alternative measures for the given scenario.

Etliche Evaluationsmetriken wurden entwickelt, um die Qualität der von einer Suchmaschine ermittelten Ergebnisse zu bewerten. Zwei davon sind ‚Precision' und ‚Recall'. Wie berechnet man diese, wenn es zu einer Anfrage gar keine passenden Ergebnisse in der Sammlung gibt? Diskutieren Sie, ob ‚Discounted Cumulative Gain' oder ‚Mean Reciprocal Rank' in so einem Fall geeignete Alternativen wären oder nicht.

Question 3 / Aufgabe 3: **(20 marks / 20 Punkte)**

Advanced Concepts ***/ Weiterführende Konzepte.***

Question 3.1 / Aufgabe 3.1: (10 marks / 10 Punkte)

Identifying misinformation is one of the major search engine challenges that have emerged in recent years. One step in that direction is automated fact-checking. Outline an IR-based approach that aims at confirming or rejecting a claim. Discuss important design decisions.

Falschinformationen zu erkennen hat sich als eine der wichtigsten Herausforderungen moderner Suchmaschinen herausgestellt. Die Überprüfung von Behauptungen auf ihren Wahrheitsgehalt ist dabei ein wichtiger Aspekt. Skizzieren Sie eine IR-basierte Architektur, deren Ziel es ist, eine Behauptung zu bestätigen oder zu widerlegen. Diskutieren Sie dabei wichtige Designentscheidungen.

Question 3.2 / Aufgabe 3.2: (10 marks / 10 Punkte)

Query log analysis has been used in recent years to improve search engine results. Outline how Google might incorporate such log data to re-rank results. What are the benefits and potential limitations of using query logs in this case?

Die Analyse von Logdaten, welche das Verhalten von Suchmaschinennutzern dokumentieren, hat zur Verbesserung von Suchmaschinenergebnissen geführt. Erläutern Sie, wie Google solche Logdaten beim ‚Re-Ranking' von Ergebnissen einbeziehen könnte. Was sind die Vor- und Nachteile der Auswertung von Logdaten in diesem Fall?