

Prof. Dr. U. Kruschwitz

3.8.2022

Klausur "Einführung in das Information Retrieval" SS2022

Exam "Introduction to Information Retrieval" SS2022

<i>Nachname, Vorname</i>	
<i>Abschluss (BA, MA, FKN etc.)</i>	
<i>Matrikelnummer, Semester</i>	
<i>Versuch (1/2/3)</i>	

Bitte füllen Sie zuerst den Kopf des Angabenblattes aus!

Die Klausur dauert 90 Minuten. Beantworten Sie bitte alle Fragen, also Fragen 1,2 und 3. In der Klausur sind insgesamt 100 Punkte zu erreichen.

Die Klausur besteht aus 9 Seiten.

Bitte beantworten Sie alle Fragen im freien Platz direkt nach der jeweiligen Teilfrage. Falls der Platz nicht reichen sollte, so nutzen Sie gegebenenfalls das leere Blatt im Anhang und kennzeichnen dies entsprechend. Sie können das Blatt auch als Schmierpapier benutzen. Eigene Schmierblätter sind nicht erlaubt.

Viel Erfolg!

Question 1 / Aufgabe 1:

(30 marks / 30 Punkte)

Basics / Grundlagen.

Question 1.1 / Aufgabe 1.1:

(10 marks / 10 Punkte)

Describe briefly how *Elasticsearch* combines search over structured data with search over unstructured data.

Beschreiben Sie kurz, wie Elasticsearch die Suche von strukturierten und unstrukturierten Daten miteinander verbindet.

Question 1.2 / Aufgabe 1.2:

(10 marks / 10 Punkte)

The Boolean model of information retrieval has become much less popular these days. Briefly describe two scenarios in which the Boolean model might nevertheless be the preferred option.

Das Boolesche Modell im Information Retrieval ist deutlich weniger populär geworden. Skizzieren Sie zwei Anwendungsbeispiele, bei denen dieses Modell aber trotzdem die passendste Option sein kann.

Question 1.3 / Aufgabe 1.3:

(10 marks / 10 Punkte)

Neural approaches have become very popular in Information Retrieval (IR). Briefly outline an IR architecture that illustrates the benefits of such an approach when it comes to representing textual data.

Auf dem Gebiet des Information Retrieval (IR) sind neuronale Ansätze sehr populär geworden. Skizzieren Sie eine Suchmaschinenarchitektur, die die Vorteile eines solchen Ansatzes bei der Repräsentation der Textdaten illustriert.

Question 2 / Aufgabe 2:

(40 marks / 40 Punkte)

Applications and Evaluation / *Anwendungen und Evaluierung.*

Question 2.1 / Aufgabe 2.1:

(10 marks / 10 Punkte)

Discuss how Web site search might differ from Web search.

Diskutieren Sie, wie sich die Suche auf Webportalen von Websuche unterscheidet.

Question 2.2 / Aufgabe 2.2:

(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 only a single relevant document exists in the collection? Discuss whether *discounted cumulative gain* or *mean reciprocal rank* might or might not be suitable alternative measures for the given scenario.

Eine Vielzahl an Evaluationsmetriken wurden entwickelt, um die Qualität der von einer Suchmaschine ermittelten Ergebnisse zu bewerten. Zwei davon sind ‚Precision‘ und ‚Recall‘. Welchen Einfluss hat es auf die Berechnung dieser zwei Metriken, wenn es zu einer Anfrage nur genau ein passendes Ergebnis in der Sammlung gibt? Diskutieren Sie, ob ‚Discounted Cumulative Gain‘ oder ‚Mean Reciprocal Rank‘ in so einem Fall geeignete Alternativmetriken wären oder nicht.

Question 2.3 / Aufgabe 2.3:

(10 marks / 10 Punkte)

Discuss the applicability of the PageRank algorithm in a university library catalogue setting.

Diskutieren Sie den Nutzen des PageRank-Algorithmus für die Suche in einer Universitätsbibliothek.

Question 2.4 / Aufgabe 2.4:

(10 marks / 10 Punkte)

Outline a scenario in which one might apply *A/B testing* to evaluate a Web search system. Justify your answer.

Beschreiben Sie ein Szenario, bei dem ‚A/B-Testing‘ zur Evaluierung einer Websuchmaschine angewendet werden könnte. Begründen Sie Ihre Antwort.

Question 3 / Aufgabe 3:

(30 marks / 30 Punkte)

Advanced Concepts / Weiterführende Konzepte.

Question 3.1 / Aufgabe 3.1:

(10 marks / 10 Punkte)

Separating *fake news* from *real news* is one of the major search engine challenges that have emerged in recent years. One step in that direction is automated fact-checking. Imagine you are tasked with developing a system for automated fact-checking. Assume that your system will be incorporated in a Web search engine and is called whenever a user submits a query that is classified as a claim. Outline a possible processing pipeline that could confirm or reject a claim. Discuss important design decisions.

Bewusste Falschinformation zu erkennen hat sich als eine der wichtigsten Herausforderungen moderner Suchmaschinen herausgestellt. Die Überprüfung von Behauptungen auf ihren Wahrheitsgehalt ist dabei ein wichtiger Aspekt. Stellen Sie sich vor, Sie sollen so ein System entwickeln, welches in einer Suchmaschine integriert wird, so dass jede Nutzeranfrage auf ihren Wahrheitsgehalt geprüft wird, wenn diese als Behauptung klassifiziert wird. Skizzieren Sie eine passende Architektur, die eine solche Anfrage entweder bestätigt oder widerlegt. Diskutieren Sie dabei wichtige Designentscheidungen.

Question 3.2 / Aufgabe 3.2:

(10 marks / 10 Punkte)

Log analysis can be used to personalize a search engine. Present three possible motivations for this approach. Discuss how one might integrate log analysis in the *query submission* stage of an information retrieval system.

Die Analyse von Logdaten kann zur Personalisierung von Suchmaschinen benutzt werden. Beschreiben Sie drei mögliche Motivationen dafür. Diskutieren Sie, wie Logdatenanalyse bei der Verarbeitung der Nutzeranfragen einer Suchmaschine eingebettet werden kann.

Question 3.3 / Aufgabe 3.3:

(10 marks / 10 Punkte)

Discuss how contextual information about the searcher is being incorporated in modern search engines to provide a better result ranking.

Diskutieren Sie, wie moderne Suchmaschinen den Nutzerkontext mit in die Suche einbezieht, um die Qualität der Ergebnisse zu verbessern.