

Prof. Dr. U. Kruschwitz

29.7.2020

Klausur "Einführung in das Information Retrieval" SS 2020

<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)

Briefly explain the motivation for using the factors *idf* and *tf* in the weighting formula *tf.idf*.

Geben Sie eine kurze Motivation für die Faktoren idf und tf in der Formel tf.idf.

Question 1.2 / Aufgabe 1.2:

(10 marks / 10 Punkte)

Briefly discuss three different reasons that might explain the popularity of *Elasticsearch* over alternative search engines when applying it to a local Web site.

Diskutieren Sie drei Gründe, die die Popularität von Elasticsearch bei lokaler Suche im Vergleich zu anderen Suchmaschinen erklären könnten.

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.

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

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)

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 2.3 / Aufgabe 2.3:

(10 marks / 10 Punkte)

Discuss the applicability of the PageRank algorithm in an enterprise search setting.

Diskutieren Sie den Nutzen des PageRank-Algorithmus in der Enterprise-Suche.

Question 2.4 / Aufgabe 2.4:

(10 marks / 10 Punkte)

Outline a search scenario in which you would apply *A/B testing* to evaluate a search system within an enterprise search setting. Justify your answer.

Beschreiben Sie ein Szenario, bei dem Sie ‚A/B-Testing‘ in der Enterprise-Suche anwenden würden. 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)

Web search engines have moved away from simply returning ten blue links for a user query. The output tends to be much more structured with vertical results and detailed pieces of knowledge being returned or an exact answer for a question. If I submit the query “*Where is regensburg uni?*” to Google, I get as a result the postal address of the University of Regensburg (ahead of any matching documents). Outline a possible architecture that might be in place to answer questions such as the one above.

Websuchmaschinen liefern heutzutage mehr als nur die zehn besten Treffer. Stattdessen ist das Ergebnis deutlich strukturierter durch das Einbinden vertikaler Suche oder die direkte Beantwortung von Nutzeranfragen. Auf meine Google-Anfrage “Where is regensburg uni?” liefert die Suchmaschine die Postadresse unserer Universität noch vor der Liste der besten Treffer. Beschreiben Sie eine mögliche Suchmaschinenarchitektur, die diesem Beispiel zugrunde liegen könnte.

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 answer questions such as the one in Question 3.1. What are the benefits and potential limitations of using query logs in this case?

Die Analyse von Logdaten, welche Nutzeranfragen dokumentieren, hat zur Verbesserung von Suchmaschinenergebnissen geführt. Erläutern Sie mit Hilfe des Beispiels in Aufgabe 3.1, wie Google solche Logdaten zur Beantwortung der Anfrage einbeziehen könnte. Was sind die Vor- und Nachteile der Auswertung von Logdaten bei Anfragen dieser Art?

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.