

Name:	Studiengang: <input type="checkbox"/> B.A. <input type="checkbox"/> MA.
Vorname:	In FlexNow angemeldet: <input type="checkbox"/> Ja <input type="checkbox"/> Nein
Matrikelnummer:	
Studienfächer:	Fachsemester Informationswissenschaft:

Allgemeine Hinweise:

1. Überprüfen Sie bitte, ob Sie alle Seiten der Klausurangabe vollständig erhalten haben (Gesamtzahl: 6)
2. **Bearbeitungszeit: 60 Minuten**, maximal erreichbare **Punktezah: 55**. Die jeweils erreichbare Punktezah ist bei jeder Frage angegeben. Bitte teilen Sie Ihre Arbeitszeit entsprechend ein.
3. Denken Sie daran, die Daten oben einzutragen, **bevor** Sie mit der Bearbeitung beginnen.
4. Treffen Sie bitte die Auswahl Ihrer Antworten bei Multiple-Choice-Fragen **direkt** auf dieser Klausurangabe.
5. Verwenden Sie für die Beantwortung der Freitext-Fragen ebenfalls diese Klausurangabe. Sie können jederzeit auch die Rückseiten beschreiben, falls der Platz auf der Vorderseite nicht ausreichen sollte. Bitte geben Sie in jedem Fall an, auf welche Frage sich die Lösung jeweils bezieht.
6. Benutzen Sie keine Bleistifte, keine rot schreibenden Stifte und kein TippEx, o.ä.
7. Keine Hilfsmittel sind zugelassen d.h. keine Foliensätze oder selbstgeschriebene Notizen.
8. Ein Taschenrechner dürfen Sie gerne benutzen.
9. Mobiltelefone sowie Computer am Arbeitsplatz - auch ausgeschaltet - sind **nicht zugelassen**.
10. Geben Sie keine mehrdeutigen (oder mehrere) Lösungen an. In solchen Fällen wird stets die Lösung mit der geringeren Punktzahl gewertet. Eine richtige und eine falsche Lösung ergeben also null Punkte.
11. Wenden Sie sich bei Unklarheiten in den Aufgabenstellungen immer an den Aufsichtsführenden. Hinweise und Hilfestellungen werden dann, falls erforderlich, offiziell für alle Teilnehmer durchgegeben.
12. Bei falschen Antworten in Multiple-Choice Fragen werden Punkte Abgesogen.

1a) Name two particular problems associated with performing IR on Japanese documents
Nennen Sie zwei Probleme, die man lösen muss, wenn man IR auf japanische Dokumente anwendet. (3 Punkte)

b) Is a positional index useful for the following queries? Explain each answer (hint: you may want to refer to bi-word indexes in some of your answers)

Wäre ein 'Positional Index' für die folgende Anfragen hilfreich? Begründen Sie jede Antwort (Hinweis: ein Vergleich mit 'bi-word Indexes' könnte für manche Anfragen hilfreich sein)(3 Punkte)

- football

- "Thomas Müller"

- "Dzsenifer Marozsán"

d) Are the following statements true or false? *Welche der folgenden Aussagen stimmen zu? (8 Punkte)*
Stimmt zu *Stimmt nicht zu*

The jaccard coefficient relates to a probabilistic mathematical framework
Jaccard ist auf probabilistischer Basis realisiert

The performance of jaccard in IR is limited as it has no length normalisation component
ein Grund, dass jaccard für IR weniger geeignet ist, ist dass die Länge der Dokumente nicht berücksichtigt wird

The performance of jaccard in IR is limited as it fails to account for the rareness of terms
ein Grund, dass jaccard für IR weniger geeignet ist, ist dass selten vorkommende Wörter nicht anders gewichtet werden.

The performance of jaccard in IR is limited as it has no concept of term-frequency
ein Grund, dass jaccard für IR weniger geeignet ist, ist dass Term-Frequenz nicht beachtet wird

2a) Explain what is meant by the "feast or famine" problem in Boolean retrieval. Relate each to a query operator.

Erklären Sie das 'Feast or Famine' Problem im Zusammenhang von boolischem Retrieval. Was hat es mit bestimmten Anfrage-Operatoren zu tun? (2 Punkte)

b) For BIM to achieve comparable performance to BM25, which of the following are required?
Welche Bedingungen sind erforderlich damit BIM ähnliche Performanz wie BM25 leisten kann? (6 Punkte)

- | | | |
|---|--|--|
| - large collection
<i>eine große Sammlung</i> | - small collection
<i>eine kleine Sammlung</i> | - collection size is unimportant
<i>Größe der Sammlung ist unwichtig</i> |
| - long documents
<i>lange Dokumente</i> | - short documents
<i>kurze Dokumente</i> | - document length is unimportant
<i>Länge der Dokumente ist unwichtig</i> |
| - appropriately set parameters
<i>richtig gesetzte Parameter</i> | - inappropriately set parameters
<i>falsch gesetzte Parameter</i> | |

Explain this last answer: *Erklären Sie Ihre Antwort auf die letzte Frage:*

c) Which properties are true for the vector space model? *Welche der folgenden Aussagen stimmen für das Vektorraummodell zu? (6 Punkte)*

- | <i>True</i> | | <i>False</i> |
|--------------------------|---|--------------------------|
| <input type="checkbox"/> | the length of the document can be accounted for in the model
<i>das Modell kann die Länge der Dokumente berücksichtigen</i> | <input type="checkbox"/> |
| <input type="checkbox"/> | the frequency with which a query term features in a document affects a document score
<i>wie oft ein Wort in einem Dokument vorkommt beeinflusst wie hoch ein Dokument bewertet wird</i> | <input type="checkbox"/> |
| <input type="checkbox"/> | rarer terms are typically weighted as less important
<i>selten vorkommende Wörter werden niedriger gewichtet.</i> | <input type="checkbox"/> |

3) The table below lists relevance judgements for the top 10 documents returned by 3 systems.
 Die Tabelle zeigt ob die ersten 10 Treffer von drei Suchmaschine relevant (R) sind oder nicht (N)

	System 1	System 2	System 3
1	N	N	R
2	N	R	N
3	R	N	N
4	N	N	N
5	N	N	N
6	N	N	N
7	N	N	N
8	N	R	N
9	R	R	N
10	R	R	N

Platz zum Rechnen

a) which of the 3 systems has the **highest** precision@10 score?
 Welches System leistet die **beste** "precision@10"? (3 Punkte)

System 1 System 2 System 3

b) which of the following 3 systems has the **lowest** average precision score?
 Welches System leistet die **schlechteste** 'Average precision'? (3 Punkte)

System 1 System 2 System 3

c) which metric best describes the comparative performance of these systems for navigational web search? Explain your answer:
 Welches Maß beschreibt am besten die Leistung der 3 Systeme? Erklären Sie Ihre Antwort. (4 Punkte)

P@1, P@5, P@10, Average precision

4a) Complete the following table for documents D1 and D2. There are 100,000 documents.
 Füllen Sie die Tabelle für Dokumente D1 und D2 aus. Es gibt 100,000 Dokumente (7 Punkte)

Term	DF	IDF	LogTF -D1	TF-IDF- D1	Euclid. Norm. D1	LogT F-D2	TF-IDF- D2	Euclid. Norm. D2
fun	500							
times	200							
regensburg	50							
tourist	100							

Q fun times regensburg
 D1 tourist fun
 D2 regensburg regensburg tourist

b) Which document ranks higher using term weights as calculated in a) and with query weighting of 1 if a term is present in the query and 0 otherwise?

Welches Dokument wird höher gerankt, falls man Wortgewichtungen wie in a) berechnet und eine Querygewichtung von 1 gibt, wenn ein Wort ein Teil der Anfrage ist und 0 wenn nicht. (5 Punkte)

5) Outline the main ideas behind Elswiler and Wilson's casual-leisure model and explain the relation to IR. (4 Punkte)

Beschreiben Sie die Hauptideen hinter Elswiler and Wilson's Casual-Leisure-Modell und erklären Sie den Bezug zu IR.

6) Consider a web graph with three nodes 1, 2 and 3. The links are as follows: *Stellen Sie sich ein Webgraph mit 3 Knoten, 1, 2 und 3 und den folgenden Verbindungen vor:*

$1 \rightarrow 3, 3 \rightarrow 2, 2 \rightarrow 1, 2 \rightarrow 3.$

a) Provide the transition probability matrix for the surfer's walk with teleporting when $\alpha = 0.5$

Geben Sie die Übergangswahrscheinlichsmatrix für den Graph, wenn Teleporting benutzt wird und $\alpha = 0.5$.
(3 Punkte)

b) If we start the random walk on node 1, what is the probabilities of being in nodes 1, 2 and 3 after 2 transitions?. *Falls wir auf Knoten 1 anfangen, mit welcher Wahrscheinlichkeit werden wir auf Knoten 1, 2 und 3 nach 2 Schritten sein?*
(3 Punkte)