Name:	**Studiengang: ☐ B.A. | ☐ MA.**
Vorname:	**In FlexNow angemeldet: ☐ Ja | ☐ Nein**
Matrikelnummer:	
Studienfächer:	**Fachsemester Informationswissenschaft:**

Allgemeine Hinweise:

1. Überprüfen Sie bitte, ob Sie alle Seiten der Klausurangabe vollständig erhalten haben (Gesamtzahl: 9)

2. **Bearbeitungszeit: 90 Minuten**, maximal erreichbare **Punktezahl: 55**. Die jeweils erreichbare Punktezahl 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.

Aufgabe 1) Indexing and Boolean Retrieval

apple -> D1 -> D4 ->D10

aardvark -> D2

banana -> D1 -> D10

fruit -> D1 -> D9 -> D11

fun -> D3 -> D7 -> D13 -> D14 -> D15

honey -> D4 -> D9 -> D13

person -> D7 -> D9 -> D14 ->D16

railway -> D7 -> D13 -> D14 -> D15

a) For the document collection above which of the documents should be returned for the following queries:

Gegeben ist die obige Dokumentensammlung. Welche Dokumente sollten die folgenden Anfragen liefern?

1. person AND railway (3 marks)

2. fruit OR (apple AND NOT banana) (3 marks)

b) Which of these boolean query syntax examples does the following pseudo code function calculate? (3 Marks)

Welchen der unten genannten boolischen Ausdrücke rechnet die folgende Pseudocodefunktion aus?

OR ☐ AND NOT ☐

MERGE (x, y)

```
1.  answer <- ()
2.  while x!=NIL and y!=NIL
3.  do if docID(x)=docID(y)
4.       then x<- next (x)
5.       y <- next (y)
6.  else if docID (x) < docID (y)
7.      then ADD (answer, docID(x))
8.      x <- next (x)
9.  else y <- next (y)
10. return (answer)
```

c) Provide similar pseudo code for an AND merge algorithm for two terms x and y (4 marks)

Schreiben Sie Pseudocode für einen "AND merge" Algorithmus für zwei Wörter x und y (4 Punkte)

d) Which of the following are explicit methods for dealing with phrasal queries e.g. "Informationswissenschaft Regensburg"? Tick all that are appropriate (5 marks)

Welche der folgenden Techniken sind ausdrücklich für Mehrwortanfragen geeignet? Bitte kreuzen Sie alle zutreffenden Antworten an (5 Punkte)

Cosine Normalisation	☐
Bi-word index	☐
Pagerank	☐
Inverted index	☐
Positional index	☐

e) Of those that you ticked, which would be least suitable for longer phrasal queries e.g. "Peter picked a pickled pepper" (1 mark). Explain why (3 Marks).

Welche Ihrer ausgewählten Antworten wäre am wenigsten für längere Mehrwortanfragen (wie z.B "Peter picked a pickled pepper") geeignet? (1 Punkt). Erklären Sie warum (3 Punkte)

Aufgabe 2) Retrieval Models

a) In the Vector Space Model (VSM) documents and queries are represented in an n-dimensional space. What is n? (3 marks)

Im Vektorraummodell werden Dokumente und Anfragen in einem Raum mit n Dimensionen dargestellt. Welche Größe hat n? (3 Punkte)

- the number of documents in the collection ☐
  der Anzahl der Dokumente in der Sammelung
- the number of words in the vocabulary ☐
  die Grösse des Wortschatzes
- the length of the query ☐
  die Länge der Anfrage
- the average length of documents ☐
  die schnittliche Länge der Dokumente

b) Explain why the Euclidean distance is by itself an inappropriate measure for comparing documents and queries in the vector space model (3 marks)

Erklären Sie warum der Euklidischer Abstand alleine nicht geeignet ist für den Vergleich von Dokumenten und Anfragen in einem Vektorraummodell (3 Punkte)

c) What is the Inverse Document Frequency (IDF) of a term that occurs in every document (3 marks)

Wie groß ist die Inversdokumentfrequenz (IDF) für ein Wort, das in jedem Dokument in der Sammlung steht? (3 Punkte)

- *1* ☐
- *1.5* ☐
- *0* ☐
- *0.5* ☐

d) Consider the table of term frequencies for 2 documents denoted Doc1 and Doc2 below. Compute the idf and tf-idf weights for each term for each documents 1 and 2. There are 1 million documents in the collection. (6 Marks)

Rechnen Sie die idf und tf-idf Gewichte für jedes Wort in Dokumenten 1 und 2 aus. Es gibt 1 Million Dokumente in der Sammlung (6 Punkte)

term	TF (Doc 1)	TF (Doc 2)	DF
train	27	24	20,018
travel	3	0	18,078
fun	0	29	7,000
best	14	17	11,213

e) Which document (1 or 2) would rank higher for the query **best travel** using a Euclidean normalized tfidf weighting scheme. (4 Marks)

*Welches Dokument würder höher eingeordnet werden für die Anfrage **best travel** wenn eine euklidischnormalisierte Gewichtung benutzt würde? (4 Punkte)*

A reminder of the cosine function:

Die Cosinusfunktion lautet:

$$\cos(\vec{q},\vec{d}) = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

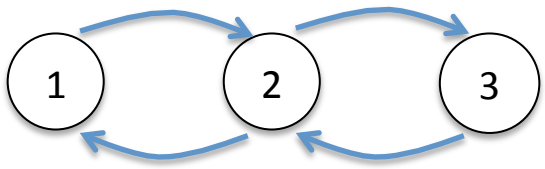
3. Web search and Evaluation

a) Name 3 important differences between a traditional IR system and a web search engine (3 marks)
Nennen Sie 3 wichtige Unterschiede zwischen einem traditionellen IR-System und einer Websuchmaschine (3 Punkte)

b) Explain why it is important for a crawler to detect whether two pages that it has downloaded are "near duplicates". (2 marks)
Erklären Sie, warum es für einen Crawler wichtig wäre, erkennen zu können, ob zwei Seiten fast identisch sind.(3 Punkte)

c) For the following web graph write down the transition probability matrices for the surfer's walk with teleporting probability 1) $\alpha = 0$ and 2) $\alpha = 0.5$ (4 marks)

Schreiben Sie die Übergangsmatrix für den folgenden Webgraphen mit Teleportingwahrscheinlichkeit 1)) $\alpha = 0$ and 2) $\alpha = 0.5$ auf (4 Punkte)



d) The table below shows the relevance judgements for the top 10 results returned by two systems for a given information need.

Die Tabelle zeigt die Relevanzurteile für die 10 bestgerankten Dokumente für 2 Systeme

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

Calculate Precision@10 for each system (3 Marks)

Rechnen Sie Precision@10 für beide Systeme aus (3 Punkte)

e) From a user modelling perspective what does the metric precision@10 model? (3 Marks)

Welche Idee liegt dem Maß precision@10 zugrunde? Warum wird es benutzt? (3 Punkte)