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: 90**. 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.

Documents (D1:D4). These are used throughout the exam. *Diese Dokumente werden für bestimmte Fragen gebraucht.*

D1: Google's Project Glass is a wearable computer that records video and shows contextual information in real time.
D2: Wearing glasses is so annoying. Google should invent an artificial eye.
D3: Making glass structures like this requires skill.
D4: The simple interface made Google popular

Aufgabe 1) Indexing and Boolean Retrieval

a) Explain the concept of stemming and why it might be used in an IR System (3 Punkte).

Erklären Sie was "Stemming" ist und warum es im Kontext eines IR-Systems vielleicht hilfreich sein könnte.

b) Stem documents D1:D4 marking any changes on the print out below. (4 Punkte)

Wenden Sie "Stemming" auf die folgenden Dokumente an. Sie sollten Änderungen direkt in den Dokumente markieren

D1: Google's Project Glass is a wearable computer that records video and shows contextual information in real time.
D2: Wearing glasses is so annoying. Google should invent an artificial eye.
D3: Making glass structures like this requires skill.
D4: The simple interface made Google popular

c) Are the following statements true or false? (6 Punkte)

Sind die folgenden Aussagen wahr oder falsch?

Wahr		**Falsch**
☐	In a Boolean retrieval system, stemming never lowers precision. *In einem boolischen Retrieval System, senkt Stemming niemals die "precision".*	☐
☐	In a Boolean retrieval system, stemming never lowers recall. *In einem boolischen Retrieval System, senkt Stemming nie Recall.*	☐
☐	Stemming increases the size of the vocabulary. *Stemming vergrössert die Grösse der Wortschatz.*	☐
☐	Stemming should be invoked at indexing time but not while processing a query. *Stemming sollte nur zur Indizierungszeit angewendet werden.*	☐

d) A data structure we talked about regularly in the class was a term-document incidence matrix (shown below). This is a conceptually useful structure, but in practice it is not used. Why? (3 Punkte)

Eine der Datenstrukturen, die wir im Kurs analysiert haben, war die "term-document incidence matrix" (abgebildet im Diagram unten). Vom Konzept her ist diese Datenstruktur sehr nützlich, aber in der Praxis wird sie nicht benutzt. Warum?

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

e) Name a data structure that is used in practice (3 Punkte)
Nennen Sie eine Datenstruktur, die in der Praxis häufig genutzt wird.

f) Explain how the data structure you named in (e) works by creating an index for the Shakespeare documents from part d) (6 Punkte)

Erklären Sie, durch die Herstellung eines Indexes für Shakespeare-Dokumente in d), wie die Datenstruktur, die Sie für e) genannt haben, funktioniert.

Aufgabe 2) Retrieval Models

a) Why are term frequency (TF) and inverse document frequency (IDF) used so often in document scoring functions? (5 Punkte)
Warum werden Termfrequenz (TF) und Inversdokumentfrequenz (IDF) so häufig in Scoring-Funktionen benutzt?

b) Why is the logarithmic function often used in the calculation of TF and IDF? (2 Punkte)
Wofür spielt die Logarithmierung bei der Berechnug TF und IDF eine Rolle?

c)

> "Gangnam Style" is a K-pop single by the South Korean musician PSY. The song was released in July [2012] as the lead single of his sixth studio album and debuted at number one on South Korea['s] Gaon Chart. In December, [2012], "Gangnam Style" became the first video in the history of the Internet to be viewed more than a billion times. As of January [1], [2013], the music video has been viewed over [1.10] billion times on YouTube, and it is the site['s] most watched video after surpassing Justin Bieber['s] single "Baby".

Under a MLE-estimated unigram probability model, what are **P(the)** and **P(video)** for the above text about ''Gangnam Style''? Treat terms in quotes as a single term. Endings to be stemmed are marked using [] and you can ignore the numbers marked in [] (6 Punkte)

*Mit einem Maximum-Likelihood-Schätzung (MLE) Unigram -Modell schätzen Sie **P(the)** und **P(video)** für den oben genannten Text über ''Gangnam Style''. **Annahmen:** Sie sollten Wörter in Anführungszeichen als ein Wort behandeln. Endungen, die gestemmt werden sollten, sind mit [] bezeichnet. Zahlen in [] können auch ignoriert werden.*

d) Under a MLE-estimated bigram model, what are **P(lead|the)** and **P(times|billion)** for the same text? (6 Punkte)

*Mit einem Maximum-Likelihood-Schätzung (MLE) Bigram-Modell schätzen Sie **P(lead|the)** and **P(times|billion)** für den ''Gangnam Style'' Text. Es sollte von den selben Annahmen wie in c) ausgegangen werden.*

e) Assuming a mixture model between the documents and the collection both weighted at 0.5 (i.e. $\lambda = 0.5$), build a query likelihood language model for document collection (D1-D4) and the query ''google glass''. Maximum likelihood estimation (MLE) should be used to estimate the collection and each of the documents as unigram models. As a reminder of what such a mixture model looks like the formula is provided below:

Angenommen eine gleichgewichtete gemischte Verteilung zwischen den Dokumenten und der Sammelung (d.h. $\lambda = 0.5$), erstellen Sie eine a query likelihood LM für D1-D4 und die Anfrage ''google glass''. *Maximum-Likelihood-Schätzung-Unigram-Modelle sollten geschätzt werden. Die Formel dafür ist abgebildet unten.*

$$P(q|d) \propto \prod_{1 \leq k \leq |q|} (\lambda P(t_k|M_d) + (1 - \lambda) P(t_k|M_c))$$

What is the final ranking of the documents (10 Punkte)

Rechnen Sie die Endfassung des Rankings aus.

f) Which of the following is used in the mixture model approach to the query likelihood model you used in e). (3 Punkte)

Welcher der folgenden Konzepte werden in der Mischverteilung, die Sie für Teil e) benutzt haben, verwendet

Used		**Not Used**
☐	Term frequency (*Termfrequenz*)	☐
☐	document frequency (*Dokumentfrequenz)*	☐
☐	collection frequency (*Sammelungfrequenz)*	☐

g) What is the effect of having a very high value of λ e.g. (0.99) in the mixture model? (4 Punkte)

Was passiert wenn ein sehr hoher Wert für λ z.B. (0.99) ausgewählt wird?

h) An extra feature of using a mixture model approach is that it offers an idf-like weighting. Explain, with the aid of a numeric example, how this works. (4 Punkte)

Eine zusätzliche Eigenschaft eines gemischten Modells ist, dass es eine IDF-ähnliche Gewichtung anbietet. Erklären Sie mit Hilfe eines fiktiven Zahlenbeispiels, wie das funktioniert.

i) Language models can be used for a variety purposes within and outwith IR. Examples we saw in class included speech recognition, spelling correction, machine translation and simulated evaluations. Choose any of the above uses and with a simple example demonstrate how language models may be used for this purpose. (4 Punkte)

Language-Modelle (LM) haben viele Anwendungsmöglichkeiten inner- und außerhalb des Bereich des Information Retrievals. Im Kurs haben wir Spracherkennung, Rechtschreibprüfung, Maschinenübersetzung und simulierte Evaluationen besprochen. Wählen Sie eines dieser Beispiele aus und erklären Sie mit Hilfe eines einfachen Beipiels wie LM für diesen Zweck genutzt werden können.

3. Evaluation

a) An IR system returns 7 relevant documents and 11 non relevant documents for a given query. There are 20 relevant documents in the collection. Calculate precision and recall for the system on this query. (5 Punkte)

Ein IR-System liefert 7 relevante Dokumente und 11 nicht-relevante Dokumente für eine gegebene Anfrage. Es gibt ingesamt 20 relevante Dokument in der Sammlung. Berechnen Sie Precision und Recall für dieses System und diese Querie (Anfrage?)

b) Consider two information needs for which there are 10 and 12 relevant documents in the collection respectively. The table below shows the relevance judgements for the top 10 results returned by two system for these needs.

Die zwei unten angegebene Informationsbedürfnisse haben 10 bzw. 12 relevante Dokumente in der Sammlung. Die Tabelle zeigt die Relevanzurteile für die 10 bestgerankten Dokumente für System 1 und 2.

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

Calculate the MAP (Mean Average Precision) for each system. (8 Punkte)
Berechnen Sie die MAP (Mean Average Precision) für beide Systeme.

$$MAP = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{Q_j} \sum_{i=1}^{Q_j} P(doc_i)$$

Q_j number of relevant documents for query j
N number of queries
$P(doc_i)$ precision at ith relevant document

c) Which of the following metrics are appropriate for the following situations. Explain your choice for each. (8 Punkte)
Welche der folgenden Maße sind für die untengenannten Suchprobleme geeignet. Erklären Sie Ihre Wahl.

Precision, Recall, precision@k, precision@1, MAP, Mean Reciprocal Rank

- Navigational Web-search

- Informational Web-search

- Legal Search (where evidence is being sourced in a company's documents for copyright infringement) *Beweis für Urheberrechtsverletzung wird innerhalb der Dokumentsammelung einer Firma gesucht*

- Known-item Search (i.e. Email Search)