D5: Databases and Information Systems
Information Retrieval and Data Mining, WS 2013/14
Dr. Klaus Berberich and Dr. Pauli Miettinen
Second Short Test, Duration: **45 minutes**

max planck institut informatik

Name:					
Matriculation Number:					
Tutorial Group:	A ☐	B ☐	C ☐	D ☐	E ☐

Question:	1 (5 Points)	2 (6 Points)	3 (5 Points)	4 (5 Points)	Total (21 points)
Score:					

General instructions:

- **The written test contains 4 questions and is scheduled for 45 minutes. The maximum amount of points you can earn is 21.**
- **Please verify if your exam consists of 12 pages with 4 questions printed legibly, else contact the examiner immediately.**
- **No electronic devices (calculator, notebook, tablet, PDA, cell phone) are allowed.**
- **Answers without sufficient details are void (e.g.: you can't just say "yes" or "no" as the answer).**
- **Last page consists of material that you may use to solve the questions. You may detach the last page for your convenience.**
- **You will be provided additional working sheets if necessary. Make sure to return them along with your solution sheet.**
- **Please provide your ID card when asked by the examiner.**
- **Please fill in name, matriculation number (student registration number) and tutor group in the form above and return the solution sheets into the provided box.**
- **Please sign below.**

Student's Signature _____


max planck institut
informatik

Effectiveness Measures

Problem 1.

An IR system returns 20 results for a query. The results returned at ranks 1, 2, 4, 8, 16 are relevant; the results returned at all other ranks are irrelevant. We further know that there exist 10 relevant results for this specific query, all of which have been assigned the graded label 2 : `relevant`. All other results have been assigned the graded label 0 : `irrelevant`.

(a) Compute *Precision* and *Recall*. **[1 point]**

(b) Compute *Precision@5*. **[1 point]**

(c) Compute *Mean Average Precision*. **[1 point]**

(d) Compute *DCG@5*, *IDCG@5*, and *NDCG@5* (using logarithm to base 2 according to the table below). **[2 points]**

Reminder: (a) The IDCG is the DCG of the best possible result. The best possible top-5 result in our case would have graded labels $\langle 2, 2, 2, 2, 2 \rangle$*. (b) When a relevant document is not retrieved by an IR system for a query, its corresponding precision value in the computation of MAP is assumed as 0.*

$\mathbf{x}$	1	2	3	4	5	6	7	8	9	10
$\log_{\mathbf{2}}(\mathbf{x})$	0.0	1.0	1.5	2.0	2.3	2.5	2.8	3.0	3.1	3.3


max planck institut
informatik
MAX-PLANCK-GESELLSCHAFT

Language Models

Problem 2. Consider the following document collection consisting of five documents and five terms

	a	b	c	d	e
d_1	1	1	0	1	1
d_2	0	4	2	2	0
d_3	2	1	0	1	2
d_4	1	0	2	1	0
d_5	2	4	2	0	0

(a) Compute the three query likelihoods $P(q|d_1)$, $P(q|d_3)$, and $P(q|d_4)$ for the query $q = \{\,\mathsf{b}, \mathsf{d}\,\}$ assuming a multinomial language model (i.e., $P(q|d) = \prod_{t \in q} P(t|d)$) with MLE probabilities $P(t|d)$.

[2 points]

(b) Compute the two query likelihoods $P(q|d_1)$ and $P(q|d_5)$ for the query $q = \{\,\mathsf{a}, \mathsf{e}\,\}$ assuming a multinomial language model (i.e., $P(q|d) = \prod_{t \in q} P(t|d)$) with probabilities $P(t|d)$ estimated using Jelinek-Mercer smoothing ($\lambda = 0.5$).

[2 points]

(c) Assuming a uniform document prior $P(d)$, show that it is equivalent to rank documents by query likelihood $P(q|d)$ and document likelihood $P(d|q)$. That is, both return documents in the same order.

[2 points]


max planck institut
informatik
MAX-PLANCK-GESELLSCHAFT

Link Analysis

Problem 3. Graph (5 nodes)

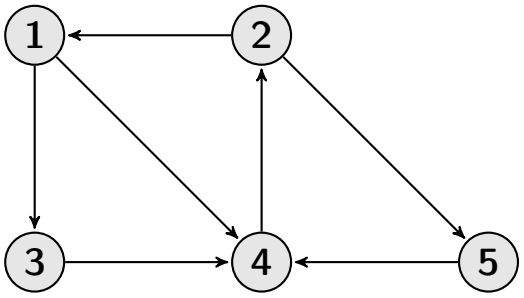
(a) Consider the non-ergodic Markov chain described by the following transition probability matrix

$$P = \begin{bmatrix} 0.1 & 0.9 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.8 & 0.0 & 0.2 \\ 0.0 & 0.0 & 0.0 & 0.3 & 0.7 \\ 0.0 & 0.6 & 0.0 & 0.0 & 0.4 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix}$$

Which two conditions necessary for ergodicity does the Markov chain violate? Explain your answers!

[2 points]

(b) Consider the following directed graph $G(V, E)$



Determine the transition probability matrix P of the Markov chain induced by PageRank for $\epsilon = 0.2$. As intermediate steps, please also provide the matrices T and J capturing transitions and random jumps of the random surfer.

[2 points]

(c) Compute the vector $\pi^{(1)}$ obtained after the first iteration of the power method, when using $\pi^{(0)} = \frac{1}{5} \cdot \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix}$ as an initial state probability distribution.

[1 point]


max planck institut
informatik
MAX-PLANCK-GESELLSCHAFT

Compression

Problem 4.

(a) Compress the following sequence of term frequencies

$$\langle 13, 41, 5, 261 \rangle$$

using *Gamma encoding*. Please specify for each number the bits that represent the length in unary and the offset in binary (e.g., 13 : **u** `1110` **b** `101`)

[1 point]

(b) Compress the following sequence of document identifiers

$$\langle 133, 197, 453, 460 \rangle$$

using *gap encoding* and *variable-byte encoding*. Please specify each number in the gap-encoded sequence together with the bytes representing it (e.g., 133 : `00000001 10000101`)

[2 points]

(c) Compress the following sequence of characters using LZ77:

`abracadabra$`

Please specify the sequence of $(back, count, new)$ triples that you obtain. **[2 points]**


MAX-PLANCK-GESELLSCHAFT
max planck institut
informatik

ADDITIONAL MATERIAL

Linear algebra

- Identity matrix: n-by-n matrix I such that $I_{ij} = 1$ iff $i = j$ and $I_{ij} = 0$ otherwise
- Product with identity matrix: $AI = IA = A$ for all n-by-n matrices A
- Matrix inverse: $A^{-1}A = AA^{-1} = I$
- Transpose identities: $(A^T)^T = A$ for all A; $(AB)^T = B^T A^T$ when the product is well-defined
- Inverse of a product: $(AB)^{-1} = B^{-1}A^{-1}$ if A and B are invertible
- Inverse of orthogonal matrices: $A^T = A^{-1}$ iff A is orthogonal

Probability & Statistics:

- Bayes' Theorem: $\Pr[A|B] = \dfrac{\Pr[B|A]\ \Pr[A]}{\Pr[B]}$
- Law of Total Probability: $\Pr[B] = \sum_{i=1}^{n} \Pr[B|A_i] \Pr[A_i]$ for disjoint events A_i with $\sum_{i}^{n} \Pr[A_i] = 1$
- Expectation: $\mathbf{E}[X] = \sum_{k=1}^{n} k f_X(k)$ and Variance: $\mathbf{Var}[X] = \mathbf{E}[X^2] - \mathbf{E}[X]^2$ for a discrete RV X with density function f_X
- Markov inequality: $\Pr[X \geq t] \leq \dfrac{\mathbf{E}[X]}{t}$ for $t \geq 0$ and a non-neg. RV X
- Chebyshev inequality: $\Pr[|X - \mathbf{E}[X]| \geq t] \leq \dfrac{\mathbf{Var}[X]}{t^2}$ for $t > 0$ and a non-neg. RV X
- Sample Mean: $\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ and Sample Variance: $S_X^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- For an estimator $\hat{\theta}$ of parameter θ over i.i.d. samples $\{X_1, X_2, .., X_i, .., X_n\}$,
  - If $\mathbf{E}[X_i] = \mu$, then $\mathbf{E}[\hat{\theta}_n] = \mu$
  - If $\mathbf{Var}[X_i] = \sigma^2$, then $\mathbf{Var}[\hat{\theta}_n] = \dfrac{\sigma^2}{n}$
  - Standard Error: $se(\hat{\theta}) = \sqrt{\mathbf{Var}[\hat{\theta}_n]}$
  - Mean Squared Error: $MSE[\hat{\theta}_n] = (\mathbf{E}[\hat{\theta}_n] - \theta)^2 + \mathbf{Var}[\hat{\theta}_n])$


max planck institut
informatik
MAX-PLANCK-GESELLSCHAFT