Universitetet
i Stavanger

DAT630-2016 Fall - Trial Part II

Questions		Type	Grading
1	DAT630-2016-trial info	Writing assignment	Manual score
2	Similarity	Simple choice	Automatic score
3	Indexing	Writing assignment	Manual score
4	Retrieval	Composite	Automatic score
5	Retrieval Evaluation	Composite	Automatic score
6	Retrieval	Writing assignment	Manual score
7	PageRank	Composite	Automatic score
8	PageRank	Simple choice	Automatic score
9	Entity retrieval	Match / pairing	Automatic score
10	Entity linking	Writing assignment	Manual score

DAT630-2016 Fall - Trial Part II

Exam start time:	16.11.2016 04:30	PDF created	14.11.2016 23:03
Exam end time:	16.11.2016 06:30	Created by	Krisztian Balog
		No. of pages	8

Section 1

1 QUESTION

DAT630-2016-trial info

You can use

- Calculator
- All written (printed) material
- All electronic material brought on a pendrive (PDFs, slides, MS Excel files, python code, etc.)
- Any program available on the PC (MS Excel, Adobe Acrobat, etc.)
- **No online resources**

Scoring of multiple choice questions:

- 2 or 3 points if correct
- 0 if unanswered
- -1 if incorrectly answered

For all computations, provide numbers up to 3 digits after the dot (e.g., 0.7, 0.25, 0.333).

If you have any comments about the exam, write them here

2 QUESTION

Similarity

We are given two documents, A and B, with term vectors, and we compute their cosine similarity. Then, we multiply all values by 2 in the term vector of A, and divide all values by 2 in the term vector of B. How will cosine similarity change? (3p)

Select an alternative:

It will be 0.5 times the original

It will not change

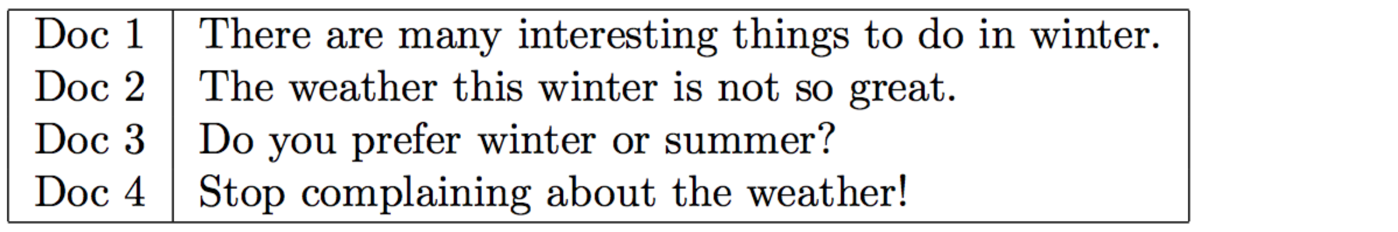
It will be 0.25 times the original

It will be 2 times the original

It will be 4 times the original

3 QUESTION

Indexing



Doc 1	There are many interesting things to do in winter.
Doc 2	The weather this winter is not so great.
Doc 3	Do you prefer winter or summer?
Doc 4	Stop complaining about the weather!

Given the above set of documents, create an inverted index with *position information*. (10p)

- Apply standard tokenization, lowercasing, and stopword removal (but no stemming).
- Use a standard English stopword list; submit the list of words you identified as stopwords.
- Stopwords do not get indexed, but their positions count. For example, if you have "word1 stopword word2", then the position of word1 is 1 and the position of word2 is 3.
- Show one posting list per line. Use : to separate the payload.
  For example: "x => y1:z1, y2:z2, ..." (You should now what x, y, and z stand for.)

Fill in your answer here

4 QUESTION

Retrieval

	term 1	term 2	term 3	term 4	term 5	term 6	length
Document 1	3	0	5	2	10	5	25
Document 2	4	3	5	2	1	5	20
Collection	100	50	80	93	100	25	1000

$$BM25(q,d) = \sum_{t \in q} \frac{f_{t,d} \cdot (1 + k_1)}{f_{t,d} + k_1(1 - b + b\frac{|d|}{avgdl})} \cdot idf_t \qquad idf_t = \log \frac{N}{n_t}$$

Compute retrieval scores using the BM25 algorithm.

- The collection row shows the number of documents that contain the given term; the collection contains 1000 documents in total.
- The average document length in the collection is 50.
- The BM25 parameters are k1 = 1.2 and b = 0.75.
- Use base-10 logarithm for the computations!

Answers: (4x4p)

The query is a single term, "term 2".

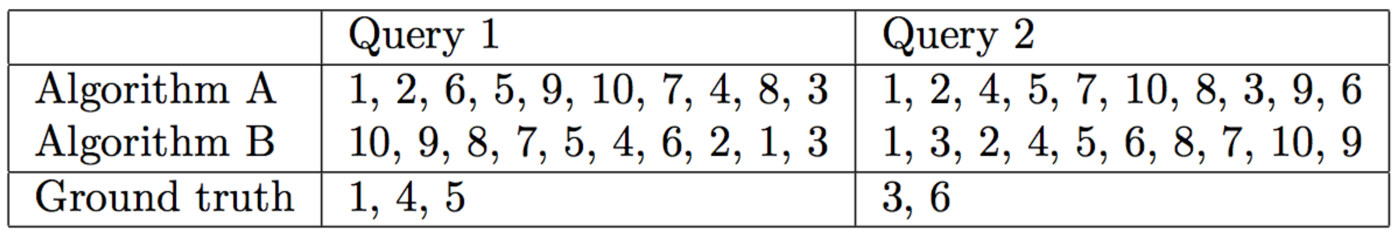
- BM25 score of Document 1:
- BM25 score of Document 2:

The query is "term2 term2 term5".

- BM25 score of Document 1:
- BM25 score of Document 2:

5 QUESTION

Retrieval Evaluation



	Query 1	Query 2
Algorithm A	1, 2, 6, 5, 9, 10, 7, 4, 8, 3	1, 2, 4, 5, 7, 10, 8, 3, 9, 6
Algorithm B	10, 9, 8, 7, 5, 4, 6, 2, 1, 3	1, 3, 2, 4, 5, 6, 8, 7, 10, 9
Ground truth	1, 4, 5	3, 6

The table shows, for two queries, the document rankings produced by ranking two different algorithms along with the list of relevant documents according to the ground truth. We assume that relevance is binary.

Answer the questions below. (5x2p)

- What is P@5 (precision at rank 5) of Algorithm A on Query 1? □
- What is the Average Precision of Algorithm A on Query 1? □
- What is the Reciprocal Rank of Algorithm B on Query 2? □
- What is the Mean Reciprocal Rank of Algorithm B? □
- Which algorithm has higher Mean Average Precision? □ (Algorithm A, Algorithm B, they have the same)

6 QUESTION

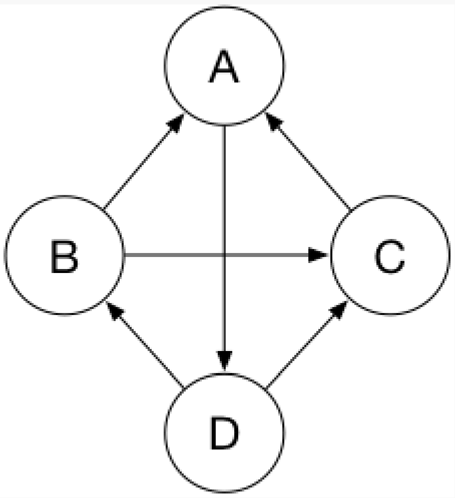
Retrieval

Explain the role of smoothing in Language Modeling. Also explain what would the effect be of setting the smoothing parameter in Jelinek-Mercer smoothing to 0 or to 1. (4p)

Fill in your answer here

7 QUESTION

PageRank



Compute the PageRank values for the following graph for two iterations. (10p)

The probability of a random jump (i.e., the parameter q) is 0.2.

	Iteration 0	Iteration 1	Iteration 2
A	0.25	□	□
B	0.25	□	□
C	0.25	□	□
D	0.25	□	□

8 QUESTION

PageRank

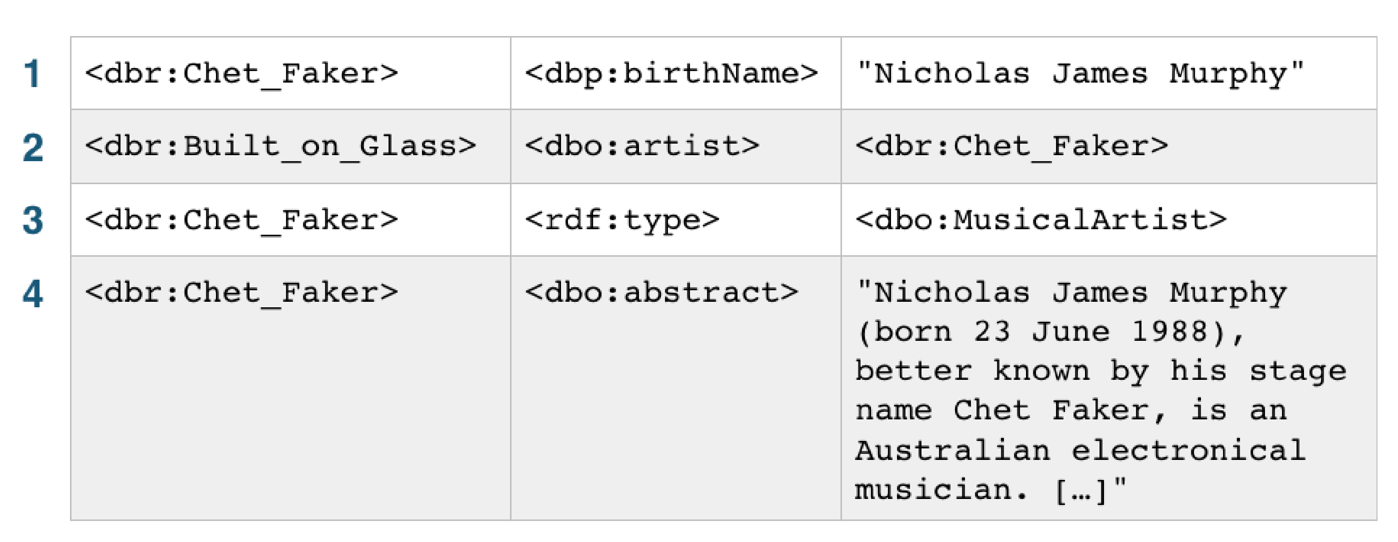
Assume that Page A has 10 in-links and Page B has 2 in-links. Which one has higher PageRank? (2p)

Select an alternative:

Page A
It's not possible to tell
Page B
They have the same

9 QUESTION

Entity retrieval



1	<dbr:Chet_Faker>	<dbp:birthName>	"Nicholas James Murphy"
2	<dbr:Built_on_Glass>	<dbo:artist>	<dbr:Chet_Faker>
3	<dbr:Chet_Faker>	<rdf:type>	<dbo:MusicalArtist>
4	<dbr:Chet_Faker>	<dbo:abstract>	"Nicholas James Murphy (born 23 June 1988), better known by his stage name Chet Faker, is an Australian electronical musician. […]"

We want to create a fielded document representation for the entity "Chet Faker" given the information associated with him in a knowledge base. We use three fields:

- names: literal objects that contain the name of the entity
- attributes: all literal objects that are not already in names
- inlinks: all incoming relations (subjects where the given entity stands as object)

Select for each RDF triple from the above image the field that it should be mapped to (or NONE if that triple is not mapped to any of the three fields). (4x2p)

Each correct answer is 2p, each incorrect answer is -1p.

Please match the values:

	names	attributes	inlinks	NONE
1				
2				
3				
4				

10 QUESTION

Entity linking

We have an entity linking system that only returns entity annotations above a given confidence threshold. First, we run this system on some input text using 0.1 as the threshold. Then, we change the threshold to 0.9 and run the system on the same input. How will precision and recall change? (4p)

Fill in your answer here