

Exercise 1: part 1

- Consider the simple web graph shown in slide 58
- Compute the matrix iteration M^k (use excel, formula MMULT, look in the help to know to use it)
- Find k such that $M^k = M^{k+1}$
- How can you obtain the page rank scores for the nodes using this matrix (M^k)?

Exercise 2: part 1

- ❑ Consider the simple web graph shown in slide 58
- ❑ Imagining to walk on this graph starting from an arbitrary initial node and following one of the outgoing links
- ❑ At each time you choose one of the links with equal probability
- ❑ Simulate this process (for 40 steps) and count how many times you visit a node
- ❑ Estimate the probability of visiting a node as the number of visits of that node divided by the total number of visits (40)
- ❑ Compare these probabilities with the pagerank scores.

Exercise 2: part 1

- Make small groups of 3 people
- Consider the slide 25 (Exploratory Search)
- Select a domain, e.g., travels, digital photography, music, politics, technology, etc.
- For each type of search (lookup, learn, investigate) define two tasks of the user for that type of search
- Think about information tools, that are available, and how these can be used to solve these tasks
- Then think about new tools that can better support some of these tasks.

Exercise 3: part 1

- Think about the **similarities** and **differences** of collaborative filtering and google page-rank.
- Evaluate them with respect to the following dimensions:
 - How they exploit User-Generated-Content, i.e., any kind of content that is produced by the users (e.g., ratings, reviews, pictures, etc.)
 - Are these tools context-dependent, i.e., do they behave differently if the context of the user changes (e.g., what the user knows about a topic, or the user activity on the tool in the 30 mins before the current time, or if its is night or morning)
 - When the heaviest part of the computation is performed (before or during the query reply)?



Part 2

Exercise 1.2

- Consider these documents
 - Doc1: breakthrough drug for schizophrenia
 - Doc2: new schizophrenia drug
 - Doc3: new approach for treatment of schizophrenia
 - Doc4: new hopes for schizophrenia patients
- A: Draw the term-document incidence matrix for this document collection
- B: Draw the inverted index representation for this collections, as in Figure 1.3 (page 7)

Exercise 1.3

- For the document collection shown in Exercise 1.2, what are the returned results for these queries
 - A: schizophrenia AND drug
 - B: for AND NOT (drug OR approach)

Exercise 1.4

- Adapt the merge for the queries:
Brutus AND NOT Caesar
Brutus OR NOT Caesar
- Can we still run through the merge in time $O(x + y)$?
- If not, what can we achieve?

Exercise 1.5

- What about an arbitrary Boolean formula?
*(Brutus OR Caesar) AND NOT
(Antony OR Cleopatra)*

- Can we always merge in “linear” time?
 - Linear in what?
- Can we do better?

Exercise 1.7

- Recommend a query processing order for

*(tangerine OR trees) AND
(marmalade OR skies) AND
(kaleidoscope OR eyes)*

Term	Freq
eyes	213312
kaleidoscope	87009
marmalade	107913
skies	271658
tangerine	46653
trees	316812

Exercise 1.8

- If the query is:
 - **friends AND romans AND (NOT countrymen)**
- how could we use the freq of countrymen in evaluating the best query evaluation order?
- Propose a way of handling negation in determining the order of query processing.

Exercise

- Try the search feature at <http://www.rhymezone.com/shakespeare/>
- Write down five search features you think it could do better

Exercise 2.5, 2.6

- 2.5: Why are skip pointers not useful for queries of the form x OR y ?
- 2.6: We have a two-word query. For one term the postings list consists of the following 16 entries:
 - [4,6,10,12,14,16,18,20,22,32,47,81,120,122,157,180]
- and for the other it is the one entry postings list:
 - [47]
- Work out how many comparisons would be done to intersect the two postings lists with the following two strategies. Briefly justify your answers:
 - a. Using standard postings lists
 - b. Using postings lists stored with skip pointers, with a skip length of \sqrt{P} , as suggested in Section 2.3.