University of Augsburg, Institute of Computer Science
Prof. Dr. P. Fischer
J. Kastner, L. Rudenko

SS 2019
05. July 2019
Sheet 10

Analyzing Massive Data Sets

Exercise 1: Effectiveness Metrics (homework)

a) A database D and a query Q with following properties is given:

- D contains 1000 documents.
- Query Q returns 150 documents.
- 120 of the returned documents are relevant for the user.
- 810 documents that were not returned are not relevant for the user.

Calculate the **fallout** considering the given properties.

b) Calculate the **recall** R of a query Q if F-Measure $F_1(p, r) = 1/2$ and Precision $P = 1/2$?

Exercise 2: nDCG (homework)

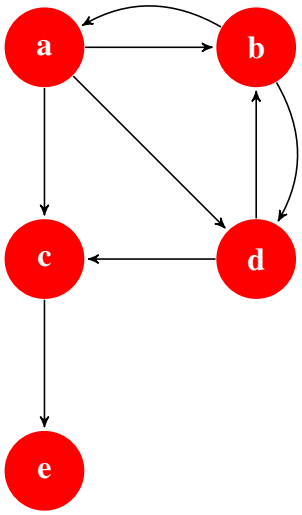
Consider the following ranking of documents with **non-binary relevance 0-3**:

2, 0, 1, 3, 2, 2, 0, 1, 1, 3, 2, 0

Calculate **normalized DCG (nDCG)** for all documents in the list.

Exercise 3: Transition Matrices (homework)

The following graph G is given:

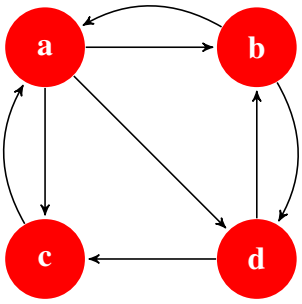


a) Specify the compact representation of the **transition matrix** for the graph G.

b) Assume only the blocks of size $s = 2$ fit into main memory. Specify the compact representation of the **transition matrix** for the graph G with this restriction.

Exercise 4: TrustRank (live)

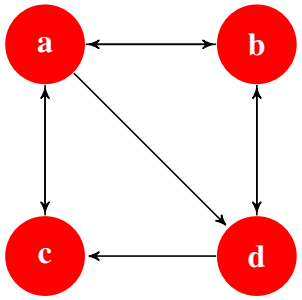
Compute the **TrustRank** of each page of the following graph, assuming only B is a **trusted** page. Use $\beta = 0.8$.



Exercise 5: PageRank with MapReduce (homework)

In this exercise you should implement **PageRank** algorithm using MapReduce approach. You can again rely on our simulator, which will provide means for a fixed number of iterations. Note that your MapReduce implementation should also handle the **dead ends**.

Exercise 6: HITS: Hubs and Authorities (live)



a) Given the Graph G from above. First of all determine the adjacency matrix A and A^T for the graph G.

b) Compute now the ranks for Hubs $h_j^{(t)}$ and Authorities $a_j^{(t)}$ using the HITS algorithm. You can rely on matrices from the previous step. Perform the iterations $t_i \mid i \in \{1, ..., 3\}$ and show the values for $h_j^{(t)}$ and $a_j^{(t)}$.

Note: Use following formulas for the normalization of $h_j^{(t)}$ and $a_j^{(t)}$: $\lambda = \frac{1}{\sqrt{\sum h_i^2}}$, $\mu = \frac{1}{\sqrt{\sum a_i^2}}$ - as stated in the original HITS paper.