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

SS 2019
28. June 2019
Sheet 9

Analyzing Massive Data Sets

Exercise 1: Fuzzy IR-model (homework)

Given is the following corpus *K* of documents:

- $D_1 = \{$Augsburg, Europe, soccer$\}$
- $D_2 = \{$soccer, Bundesliga$\}$
- $D_3 = \{$Bundesliga, Augsburg$\}$
- $D_4 = \{$Augsburg, Europe$\}$

a) Determine the **Jaccard indices** for the terms in the documents $D_j, j = 1, ..., 4$ to get a notion of **term similarity**.

b) Compute the **fuzzy degree of membership** $W(D_j, t_i)$ for each term t_i in each document D_j.

c) The following query Q is given:

Q = (Augsburg **OR** Europe) **AND NOT** (soccer)

Evaluate this query using the fuzzy model. Determine the **Top-2** results.

Exercise 2: Vector Space Model (live)

Consider the following **vocabulary** V=$\{$cat, bird, pet, dog$\}$ and the following two **documents**:

- $D_1 = \{$cat, pet, pet, dog$\}$
- $D_2 = \{$bird, pet, pet, pet, dog, dog$\}$

Furthermore the two following **queries** are given:

- $Q_1 =$ 'cat', 'pet', 'dog'
- $Q_2 =$ 'cat', 'pet', 'dog', 'bird'

a) Calculate the **term weights** for the terms actually present and the documents using **TF/IDF**.

b) Determine the **similarity** between the **documents** D_1, D_2 and the following **queries** Q_1, Q_2 using the **euclidean distance** and the **cosine correlation** and determine **the best matching documents for each query.**

Please note: For calculating the **term weights of the queries** use the **IDF values** of the documents calculated in **subtask a)**.

Exercise 3: Effectiveness Metrics (homework)

Consider a **query Q** with following **properties**:

- Q does **not return** 160 documents, which were **not relevant** for the query.
- Q **returns** 15 documents, which were **not relevant** for for the query.
- Q does **not return** 180 documents in total.
- Q is executed on a **corpus** of 300 documents.

a) Calculate **Precision**, **Recall** and **Fallout** of **query Q**.

b) Some queries can only be compared to other queries considering **Precision and Recall in combination**. Calculate the $\mathbf{F_1}$**-measure** for Q.

c) Which **changes** of the **document-conditions** are necessary to **improve the F-measure**, considering the **type-errors**?

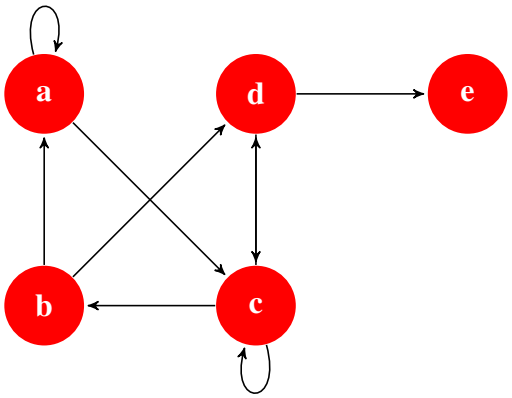
Exercise 4: Mean Average Precision (MAP) (homework)

Given are the following sets of documents delivered in response to three queries (**r** stands for **relevant**, **n** - for **not relevant**):

Ranking w.r.t. the query 1	**r**	**n**	**n**	**n**	**r**	**r**	**n**	**r**	**r**	**r**
Ranking w.r.t. the query 2	**n**	**r**	**r**	**n**	**n**	**n**	**n**	**r**	**n**	**n**
Ranking w.r.t. the query 3	**r**	**r**	**r**	**r**	**n**	**r**	**n**	**n**	**n**	**n**

Calculate **Mean Average Precision (MAP)** and draw a recall precision graph for all documents.

Exercise 5: PageRank (live)



a) Start the **power iteration** for the graph from above and perform the **first 3 iterations**. Which **scores** do you expect for the nodes?

b) Start the **power iteration** for the graph from above again and perform the first **3 iterations**. Before you start the power iteration adjust the **web matrix** using the **PageRank equation** with $\beta = 0.8$. Which **scores** do you now expect for the nodes **compared** to the **scores from subtask a)**?

c) Is there a **better solution** to **avoid problems** than using the PageRank equation from subtask b)? **Explain** your solution, start the **power iteration** for the graph again and perform the **first 3 iterations**.