D5: Databases and Information Systems
Advanced Topics in IR, SS 2016
Dr. Vinay Setty and Dr. Jannik Strötgen
Assignment 3, due: **9 Jun 2016**

max planck institut informatik

Please submit your solution as a PDF and source code together in a single tar/zip file, name your pdf file in the following format: "Lastname-Immatrikulationnumber-AssigmentX.pdf", also include Immatrikulationnumber in the title of your document and email it to `atir16@mpi-inf.mpg.de` *before the due date mentioned above!*

Effectiveness Measures for Novelty & Diversity (20 Points)

Problem 1.
Two retrieval systems return the results given in the table on the left for an ambiguous query.

	$\mathbf{R}_1$	$\mathbf{R}_2$
1.	d_1	d_4
2.	d_6	d_3
3.	d_2	d_7
4.	d_8	d_8
5.	d_5	d_1

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9
a	0	2	0	1	0	2	1	0	2
b	1	0	1	0	1	0	2	1	2

We further know that there are two query aspects **a** and **b** (each equally popular among users) and have collected the graded relevance assessments given in the table on the right.

(a) Compute standard nDCG for the two retrieval results. Use the maximum of the two graded labels assigned to a document for the two query aspects as its unified graded label.

(b) Compute intent-aware nDCG (nDCG-IA) for the two query results.

(c) Compute α-nDCG ($\alpha = 0.5$) for the two query results. Use query aspects as information nuggets and treat documents with a graded label in $\{1, 2\}$ as relevant.

Submodularity (20 Points)

Problem 2.
Carbonell and Goldstein [3] describe Maximum Marginal Relevance (MMR) as a greedy selection rule. Analogous to IA-Select by Agrawal et al. [1], we can alternatively cast MMR into the following optimization problem

$$\arg\max_{S \subseteq R} \sum_{d \in S} \left(\lambda \cdot \text{sim}(q, d) - (1 - \lambda) \cdot \max_{d' \in S, d' \neq d} \text{sim}(d, d') \right) \quad \text{s.t.} \quad |S| = k$$

where R is the set of all documents, q is the query, and S denotes the selected set of k documents.

(a) Is the objective function of the above optimization problem *submodular*? Prove your answer.

(b) Does the greedy selection rule given in [1] thus provide an approximation guarantee?

Maximum Marginal Relevance (Programming Assignment) (20 Points)

Problem 3.
In the previous two assignments you already indexed TREC sample corups using elasticsearch. We now want to compare the results obtained by MMR for different choices of λ.

(i) Implement Maximum Marginal Relevance (MMR). As a first step, determine the similarities $\text{sim}(q, d)$ for the given query q (using the Okapi BM-25 scoring model). These documents constitute the set R from which you now select the subset S. The first document to be included in S is the one having highest $\text{sim}(q, d)$. Now, include more documents in S using the greedy selection rule and computing $\text{sim}(d, d')$ using the precomputed normalized vectors.

(ii) Determine the top-10 results for the queries:

(a) Hubble Telescope Achievements

(b) African Civilian Deaths

(c) Implant Dentistry

(d) Radio Waves and Brain Cancer

(e) Alzheimer's Drug Treatment

using $\lambda \in \{0.1, 0.5, 0.9, 1.0\}$. Please compare the precision and nDCG values for the different λ values using the relevance judgments from "judgments.tsv" from the first assignment.