T-61.5020 Statistical Natural Language Processing

Exercises 3 — Information retrieval
Version 1.0

1. There are 10 000 documents in a database. A user makes a query that should give six relevant documents. Two competing search engines return ordered lists of ten documents. The relevant ones are marked with + in the table below. Consider the quality of the results using the measures mentioned below.

   – Precision
   – Recall
   – Fallout
   – Accuracy
   – Error
   – F-measure
   – Uninterpolated average precision

Document	relevance	
	engine 1	engine 2
d1	+	+
d2	+	+
d3	-	+
d4	+	-
d5	-	+
d6	-	-
d7	-	+
d8	-	-
d9	-	+
d10	+	-

2. Word w_1 exists in 21 documents and word w_2 in 500 documents in a collection of 10 000 documents. Word w_1 occurs total 101 times in all the documents, w_2 700 times. How much these words should be weighted in a search engine? Try Inverse Document Frequency (IDS) and Residual Inverse Document Frequency (RIDF) as weighting methods.

3. Let's use the following made-up news headlines as data:

   d_1: Future of Moscow track at stake — Formula 1 racing continues in Austria.

   d_2: Self-confident Schumacher: Track of Sepang suits Ferrari.

   d_3: Collision of stars — Wreck worth over 35 euros left in the track of Hungaroring.

   d_4: Five formula cars in wreck, Schumacher blames Coulthard for the collision.

   d_5: On the trail of meteors to find the origin of the galaxy.

   d_6: Dozens of planets fall from their tracks yearly in our galaxy.

   d_7: A star turned out to be a planet.

   Make a document–word matrix for words Schumacher, track, formula, collision, galaxy, star, planet, and meteor. Use stemming. How similarities of documents change if you reduce the dimension to two using Singular Value Decomposition (SVD)? Consider specially the correlation between the documents d_5 and d_7.