Foundations of Information Retrieval

Exercise 1

Exercise session: 03.11.2016
Questions to: Markus Rokicki, rokicki@L3S.de

1. Query optimization

1. A document collection with 125,000 documents contains film reviews. Given is the following query:

(NOT horror) AND (film OR animation) AND (action OR comedy)

Specify the most efficient order of execution for this query that can be determined from the following table:

Term	DF
horror	75,000
film	62,000
animation	3,000
action	41,000
comedy	40,000

Is the order you proposed always optimal?

2. Inverted index

Given is the following document collection:

D1:
Ice Age 4 was released in 34 territories.
D2:
Ice Age 4 (original Ice Age: Continental Drift) is a 2012 American computer-animated comedy film.

Create an inverted index for this document collection. **Tokenization rules**: word wise, case-folding, ignore punctuation. **Stop list**: *was, in, is, a.* Include TF and DF values at a suitable position in the index.

Which search results can be obtained from this index for the following queries?

Q_1= *Ice Age*
Q_2= *Tokyo Drift*