Assignment 1 – Information Retrieval

Task 1:

a) The invention of printing press in 1450 offered the possibility to easy and fast replication of texts. This lead for example to faster publication of knowledge (specifically science) which had massive impact on better (jnternational) research and knowledge. People who work as translator during that time might lose impact due to the printing press invention.

b) Vanevar Bush anticipated in his article „As we may think" the idea the World Wide Web. He talked about a device which could save all books, news and information and offer it to humanity in lighting pace. This is possible throughout the WWW.

Task 2:

a) *Information retrieval* is the task of finding material and satisfying the information need on large collections of unstructured data. Challenge is to find relevant documents, which are documents that fit into the given information need. The setting is not defined, so information retrieval can be done in a couple of scenarios.

Web search is more specific: It's about finding information in the web. Here, documents are websites.

b) Relevance comes along with a given information need. The users information need is expressed by his query. Documents are relevant, if they fit into the context of the information need. Thereby, relevance can be defined binary (relevant, not relevant) or by a spectrum (0, 0.1, … 1).

c) 1. Performance: Optimize runtime of the search engine.

2. Dynamic Data: New websites have to be listed in the engines indexes.

3. Scalability: It must work with a bunch of users. Ist possible, that even more users are using the engine in the future.

4. Adaptability: Index structures and rankings should be modifyable.

5. Spam: Not wanted pages should be filtered out.

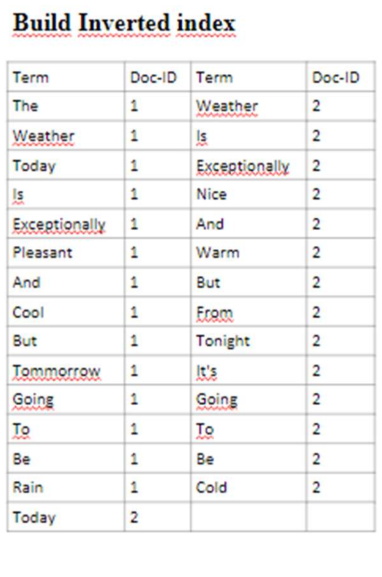
Task 3:

a) 1) aquarium AND ocean
2) tropical fish ocean || tropical fish AND NOT saltwater
3) NOT tropical fish ocean

b) query = (tropical AND ocean AND fish) OR (tropical AND ocean AND aquarium AND NOT tank)

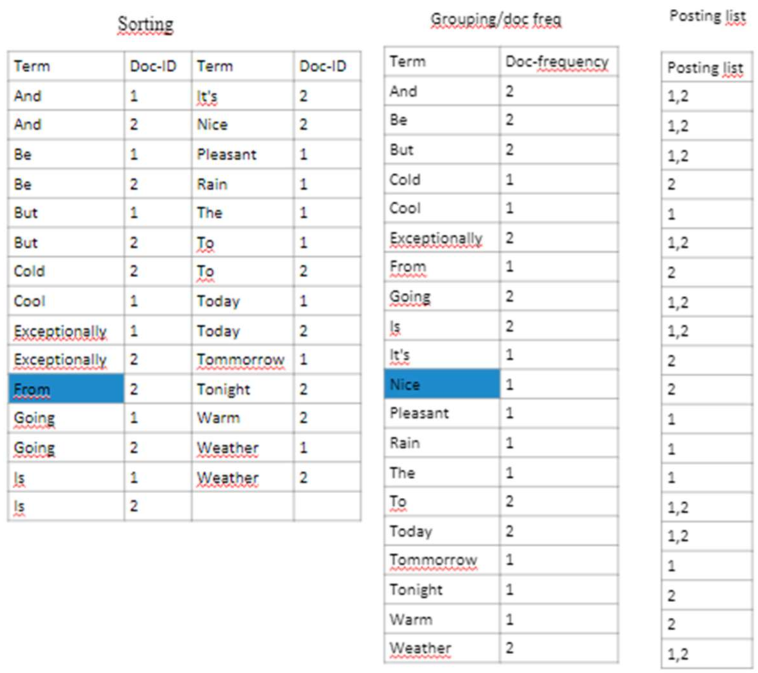
Jacob Abb 1162694
Bahne-Thiel Peters 1159242

Task 4:



Build Inverted index

Term	Doc-ID	Term	Doc-ID
The	1	Weather	2
Weather	1	Is	2
Today	1	Exceptionally	2
Is	1	Nice	2
Exceptionally	1	And	2
Pleasant	1	Warm	2
And	1	But	2
Cool	1	From	2
But	1	Tonight	2
Tommorrow	1	It's	2
Going	1	Going	2
To	1	To	2
Be	1	Be	2
Rain	1	Cold	2
Today	2		



Sorting

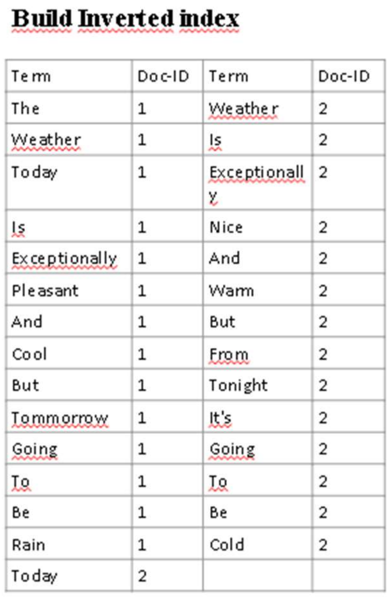
Term	Doc-ID	Term	Doc-ID
And	1	It's	2
And	2	Nice	2
Be	1	Pleasant	1
Be	2	Rain	1
But	1	The	1
But	2	To	1
Cold	2	To	2
Cool	1	Today	1
Exceptionally	1	Today	2
Exceptionally	2	Tommorrow	1
From	2	Tonight	2
Going	1	Warm	2
Going	2	Weather	1
Is	1	Weather	2
Is	2		

Grouping/doc freq

Term	Doc-frequency
And	2
Be	2
But	2
Cold	1
Cool	1
Exceptionally	2
From	1
Going	2
Is	2
It's	1
Nice	1
Pleasant	1
Rain	1
The	1
To	2
Today	2
Tommorrow	1
Tonight	1
Warm	1
Weather	2

Posting list

Posting list
1,2
1,2
1,2
2
1
1,2
2
1,2
1,2
2
2
1
1
1
1,2
1,2
1
2
2
1,2



Build Inverted index

Term	Doc-ID	Term	Doc-ID
The	1	Weather	2
Weather	1	Is	2
Today	1	Exceptionally	2
Is	1	Nice	2
Exceptionally	1	And	2
Pleasant	1	Warm	2
And	1	But	2
Cool	1	From	2
But	1	Tonight	2
Tommorrow	1	It's	2
Going	1	Going	2
To	1	To	2
Be	1	Be	2
Rain	1	Cold	2
Today	2		

a)