Assignment 1

Information Retrieval and Web Search

Ahmed Rekik - 790063 & Tim Henning - 789242

November 7, 2017

2. Newspaper Comment Retrieval

2.a. What are the main differences between Web search and Newspaper Comment search?

From a general point of view, Web search represents the seek of information among all existing data in the internet. In the other hand Newspaper comment search represents only search in comments of newspaper article written by newspaper's readers.

Since the newspaper comment are existing in the internet, could the newspaper comment search be represented as a unit of Web search. When the web search covers his sources on all the web, Newspaper comment search focuses only on the comment's contents, thereby in order to seek for specific comments it is more suitable to employ newspaper comment searching instead of web searching to avoid a swamp of other unwanted type of results.

2.b. Why would someone search for newspaper comments? Considering different kinds of comment searchers, i.e. newspaper readers, article authors, social scientists, etc. which are the main reasons for searching comments?

Newspaper comments represent mainly the reaction of reader on the post's subject and contents. These comments could help certain specialist in analyzing the flow of people opinions in terms of the information that they read for scientific, socialist or political goals.

Many readers seek also for comment to interact with them or to look for other opinions, which are suitable with his. The reader author could through comment analyze, what kind of topic the reader looks for and with which ideology should he convey his information in order to please the majority of them.

Comments can also contain critics about the author, and they may reveal other information that were not taken in consideration, which could also please the informational need of others.

3. Text Processing

3.a. What are the pros and cons of using a stopword filter?

The stopping task in indexing preprocessing aims mainly to reduce the number of occured words in the text following a choosen stopwords filter, which are not pertinent for the text context. By eliminating the most occurred words in texts, which denote stopwords in most of cases, we reduce the index space strongly (potentially 40% of existing words in the english language), what advantage the storage saving. Reducing indexes for a text minimize the number of relevant word occured in a text, thereby the response time will be improved in the seeking task. Although the stop words represent usually words that have no major meaning in the text such as pronoun or preposition, their removement may cause search limitation in some texts, which focus on one of those stopwords. Example: Searching for the band "The Who". Other big disadvantage that some of stopword may represent verb preposition and their elimination may change the whole context of phrase.

3.b. Are there any specific stopwords for comments or similar short posts used in the literature?

Generally article's comments are about the context of the article. Some specific words in the topic can occur necessarily more frequently in their comments than other topic's comments. For example in an article about a certain football player, his name should be more frequent in its comments than in comments of an article dedicated to the global warming. Thus, those specific very frequent words in a topic could represent stopwords.

3.c. What are the pros and cons of stemming and lemmatization?

Stemming and lemmatization are used in the preprocessing of texts by reducing each token to its stem, which denotes the domain of the word. By reducing all words to their domain the index space will necessarily be reduced, thus the response time will be improved.

Although this improvement is minimal for some languages as English or French, it may be very crucial for other languages such as the Arabic language. One other big benefit that both of Stemming and lemmatization yield is text or query rectification. While grammatical or conjugation faults may occur in texts or (generally) queries, reducing those fault to their domain term or infinitive will increase strongly in the effectiveness of the seek and avoid involuntary mismatches.

In an other point of view, Stemming and lemmatization may be a hard task to achieve even in preprocessing. Generally processed using algorithms, these may produce errors in refactoring some words by changing their true meaning or even by transforming the word into a stem, that has no meaning, which can yield confusions and decrease the relevance of queries.

3.d. Can stemming lower precision or recall in a simple keyword retrieval system? Explain your answer.

As the Stemming and Lemmatization task are processed with an automatic way through algorithms, may indexed or query words lose their real meaning and hence confusions may occur in matching query words with the indexed word. This issue can low strongly the precision of the result by returning matched data that are irrelevant to the user need with the potentially relevant data. In example, we query " the Biggest organization" that may be transformed to "big organ" and thus returns a bunch of irrelevant information. Following this reasoning the recall metric do not fear any decreasement since the relevant information are also returned unless if preprocessed words lose completely their meaning so it will not be returned if it is indexed.

4. (Programming) Indexing

4.a. Print a list of the comments (comment text only; no other attributes) in your dataset that match the queries: