Assignment 2

Information Retrieval and Web Search

Ahmed Rekik - 790063 & Tim Henning - 789242

November 12, 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 while newspaper comment search represents only search in comments of newspaper article written by newspaper's readers.

The ranking mechanisms can be very different as incoming links are only well suited to rank webpages but not comments and comments can be for example ranked better according to their upvotes (if present).

Web search is usually done on a much larger corpus than comment search, too. Comment search could even be regarded as a subset of a general web search.

Information noise such as ads and duplicated content is a much smaller problem with comment search than with web search. But on the other hand comment search has different characteristics that need to be considered such as the hierarchical structure of most comment pages.

To avoid a swamp of other unwanted type of results, it is more suitable to use a specialized comment search engine when the information need of the user focuses on newspaper comments.

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?

In general, newspaper comments represent the reader's reaction in term of an article content. Those comments have their benefits and they could be useful for the article author but also for readers.

The readers of the newspaper could want to find articles that support their opinion or articles that were praised by other readers in the comments.

The article authors may need to find heavily discussed topics as an inspiration for new articles. They may also want to know if there are comments that criticize the content or found point out a wrong information.

Social scientists could use comment search to find out what articles result in hate speech or how political opinions are distributed.

3. Text Processing

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

The task of removing stopwords following a previously chosen stopword list while preprocessing the texts for indexing aims mainly at reducing the number of words which are not pertinent for the text context. By eliminating the most occurring words in texts, which denote stopwords in most cases, we reduce the storage space needed for the index heavily (potentially 40% of existing words are stopwords in the English language). This can also improve the response time while searching because less information has to be read from disk.

Although the stop words usually represent words that have no major meaning in the text (such as pronouns or prepositions), their deletion may cause search limitation for some queries that focus on one of those stopwords. Searching for the band "The Who" will be nearly impossible after removing the stopwords before indexing. Another big disadvantage is that some of the stopwords may represent verb preposition and their elimination may change the whole context of phrase. Searching for negated queries could be harder then.

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 count of indexed terms will necessarily be reduced, thus the response time will be improved (random accesses to many posting lists are prevented and sequential read can be performed on one large list).

Although this improvement is not very large 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 faults to their domain term or infinitive will increase strongly in the effectiveness of the seek and avoid involuntary mismatches. Even without faults in the query or in the text, words with the same stem often contain the same information and are relevant for the same query.

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 a query results.

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

As the Stemming and Lemmatization tasks are processed in an automatic way through algorithms, indexed or query words may lose their real meaning and hence confusions may occur in matching query words with the indexed word. This issue can lower the precision of the results strongly by returning data that is irrelevant for the users information need. For example, when searching for "the Biggest organization", the query may be transformed to "big organ" and thus returns a bunch of irrelevant information with the relevant ones.

Following this reasoning the recall metric does not fear any decrease since the relevant information is also still returned if the query is stemmed or lemmatized in the same way as the corpus.

4. (Programming) Indexing

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

```
Searching:  October
Found 48 results in 0.21ms.
But sadly most of people's referendum are declared as "not
    binding" from the beginning - as are the coming one's in
    Italy for Lombardy and Venetia on the 22nd of October.
The Red Army was created after the success of the October
    Revolution  in 1917 led by V.I. Ulyanov (Lenin).\nThirty
    five years later the Red Army smashed the fascist armies of
    Germany, Finland, Hungary, Romania, Italy, Croatia and
    Slovakia.\nThe bodies of tens of thousands rascist/anti-
    semitic fascists fertilized Soviet territory helping the
    recovery of Soviet lands.\nYou stand with all that is
    immoral.
```

```
Yep. Between 6-8 million between April 1945 and October 1949.
Yeah but they're always wet as October.  That's always a plus.
May called Swamp on October 10th on JCPOA. It was decided then.
    Brits are ahead of Trump in the lineup for Tel Aviv.

Searching:  jobs
Found 2847 results in 2.29ms.
You name three countries brougt to their knees by islamic
    invasions and terrorism,sjw terrorist infiltration of govt
    jobs but its soon going to end. Camps for the terrorists and
     their support base.. just wait n watch history repeat.
consequences are that his political career is over. This
    situation clearly demonstrate that EU and PACE is not about
    solutions. That is not their job. Their job is to execute
    orders and this guy see the full consequences of thinking
    independently and toward peace.
"form a film and production crew for the purposes of producing
    a video documentary based on its research associated with Mr
    . Guelen." so the job is similar to other USA 'consulting
    firm', like film about Maidan, White Helmet, ISIS beheading
    series ( wonder when the season finale is.)
Agree-and he is doing the same job, the British elite did 80
    years or more before...
So true.\n\nAcademically questioning holocaust will make you
    lost job in University and land you in a jail.

Searching:  Trump
Found 7455 results in 6.08ms.
Who believes you Trump? You are the biggest self-loving
    opportunistic liar in politics. You arm Saudi Arabia and
    call Iran terrorists. You want peace in the middle east but
    support the settlement building of Israel in Palastine. You
    claim that the actions against Qatar are in support of your
    vision on Iran, yet you have a military base in Qatar from
    where you bomb Syria. Go away Trump!
Just months ago, Donald Trump was thanking Wikileaks and asking
     them to expose more about Hillary Clinton. Now he wants to
    arrest and jail the founder?
Trump lies about crime numbers, terrorist attacks, his taxes,
    ...
I think she has the goods on Trump.
No need.  Trump is just a wee little froggie in a great, big
    swamp.
```

```
Searching:  hate
Found 2247 results in 1.73ms.
I am a bosnian turk. I support Russia Putin and RT and support
    BRICS. Hate NWO and USA and NATO.\nMost of the Turks are
    just sheeps with no awareness no understanding what is going
     on earth.\nBut not all Turks my Russian friend.
Do you blame Pinochet, or the people that made it possible for
    him to commit those atrocities? \nOh and yes Russian names,
    though mostly western educated jews. As much as i hate
    jewbashing.
You don't hate to think about it actually
Symbols and words are hate crimes, but bullets and explosions
    are not
True tru and true again. 90's were disgucting disaster.While
    those oligarch liberals were pocketing on USSR wealth, and
    flushign it into foreign banks, regular working Russian
    people were starving. And of course those liberals miss
    those days when they could easily rob, and wish to get back
    to those times. They hate Russia, they only love money they
    can steal from the people, and those people who support them
     are so gullible and dumb.
```

Implementation Choices

- Stopword List: as there are no special stopwords in our use-case for RT.com, we are using nltk.corpus.stopwords (Porter et al 2,400 stopwords for 11 languages)
- Stemmer: using nltk.stem.porter.PorterStemmer (widely regarded as the most universal stemmer in NLTK)
- Lemmatizer: not used at the moment (search results were sufficient without it)