Information Retrieval

Project 1

Ken Bauwens & Michaël Adriaensen

20143225 20141452

Introduction

We are using the IR library pylucene. For our test dataset, we decided on using the reddit dataset.

Available functionalities

Preprocessing

Lucene contains a bunch of preprocessing functionalities for documents and queries. It supports the following preprocessing:

- Tokenization
- Stemming
- Stop word removal
- Text normalization
- N-Gram Tokenization

Indexing

Lucene's indexing is based on inverted indexing and postings. You can change the index options per field and choose to add positions and frequencies to the index.

Efficient indexing

We did not find any support for distributed indexing in lucene. There are other solutions like Elasticsearch [4] which are built around lucene and which can facilitate distributed indexing, but we don't think that distributed indexing will be necessary for an indexing system used by a single hospital for example.

Lucene allows for dynamic indexing. An index will be kept in memory until a size limit is reached and will then be flushed to a Directory.

Lucene also allows compressions with the **org.apache.lucene.codecs** module.

Lucene segments the index and uses a merging policy. The standard merging policy allows for tiered indexing.

Lucene uses segments.the complete index consists of segments which are also fully functional indices, which can then be merged.

Querying

Lucene supports different types of queries. By having separate queries for each query-type, lucene can optimize each query type. Some of the supported queries are:

- Boolean queries
  - Lucene has support for boolean queries and several other query types. This allows lucene to optimize different query types.
- Phrase queries
  - Lucene supports phrase queries
- Wild card queries
  - Wild card queries are possible in Lucene. These are computed using a state machine.
- N-gram queries
  - Phrase queries can be optimized using N-grams. Using phrase queries on a k-gram field is equivalent to k-gram query in the course material.

Lucene queries also have synonym expansion and a spelling checker with several distance measures (eg. Levenstein). There does not seem to be any other query expansion.
Relevance feedback is not implemented in lucene, but can be implemented by the user.

Ranked retrieval

By default, lucene filters results through boolean logic in a querier, and then assigns scores to these documents using a scoring model. It supports the following scoring models [1]:

- Vector space model
- Probabilistic models (Okapi BM25, DFR, …)
- Language models

Dataset

The reddit dataset [2] consists of 51 csv-files. Each file covers a specific topic (a “subreddit”) and each row in the files represents a single comment. Each row also contains metadata like the author, time of posting, …
According to the readme.md: *“All text is normalized to lowercase, tokenized using a TreebankTokenizer from natural, then joined with spaces. This results in punctuation being separated from words, a desired effect.”* [3]

We noticed that the files contained a lot of duplicate comments, so we preprocessed the data by removing these duplicates. After preprocessing and indexing we got the following statistics:

- Size of data: +- 25MB
- Time to index: +- 5.12 seconds

Queries

Some example queries and their results can be found in the appendix

REFERENCES

[1]https://lucene.apache.org/core/7_5_0/core/org/apache/lucene/search/package-summary.html#scoring

[2]https://github.com/linanqiu/reddit-dataset

[3]https://github.com/linanqiu/reddit-dataset/blob/master/README.md#Comments

[4]https://www.elastic.co/products/elasticsearch

Appendix

Query: pulp fiction
subreddit: movies
metareddit: entertainment
text: pulp fiction marvin he got shot
author: theguywithnohat
score: 12.042684555053711

subreddit: movies
metareddit: entertainment
text: the first time i saw pulp fiction when vincent accidentally killed marvin also the entire ending of the departed everything from the rooftop scene till the very end
author: wtfmynamegotdeleted
score: 4.544185638427734

subreddit: movies
metareddit: entertainment
text: six times in the theater each for am lie pulp fiction   and the avengers   i think the first wins because i drove 90 miles to see it several times and also saw it in an art house cinema while vacationing on the west coast
author: MatttheBruinsfan
score: 3.4653255939483643

subreddit: movies
metareddit: entertainment
text: saw pulp fiction 11 times it was in theaters for almost a year and i spent time in three different states that year so i d go to the theater with different people at different times there was always someone who had nt seen it plus it s a bad muther fucker
author: SonnyLouse
score: 3.0175652503967285

subreddit: movies
metareddit: entertainment
text: denis villeneuve has a scifi movie coming out later this year story of your life which should give us a better sense of how he handles science fiction even so everything he s directed so far has been excellent so i m not at all worried it will be forgotten
author: AJensen227
score: 1.2773762941360474

Query: pulp -fiction
Returns nothing

Query: p*p -fiction
subreddit: anime
metareddit: entertainment
text: holy shit when i saw sculpture in the title i thought you meant a small figure not a full on bust it s absolutely glorious good job if i was a millionaire i d take 310 one for ever pushup squat situp and km run everyday by saitama
author: RisaKoba
score: 1.0

subreddit: anime
metareddit: entertainment
text: protip never enter in an airplane with joseph in fact avoid vehicles altogether
author: NoPenNameGirl
score: 1.0

subreddit: anime
metareddit: entertainment
text: gt is it wrong to pickup girls in a dungeon i guess you re alright with fanservice so i ll recommend no game no life
author: Narglepuff
score: 1.0

subreddit: anime
metareddit: entertainment
text: hi everyone new to this subr just wanted to say hi and have been a long time fan of anime in general just wanted to know are there any other rpg online gaming fantasy type world anime that are good or especially new ive seen the below already grimgar of fantasy and ash still watching sword art online gun gale online is it wrong to pickup girls in a dungeon log horizon accel world
author: onceEPIC
score: 1.0