A Simple Document Retrieval System for Boolean Queries

Selim Öztürk 2013401027

March 9th 2018

I. Preprocessing steps:

1. Lowercase all
2. Remove punctuations, keep distance info
3. Split the text in a list
4. Remove stopwords
5. Stem words

# of tokens before: 2665309
# of tokens after: 2072477
# of unique terms before: 84512
# of unique terms after: 71897

20 most frequent terms before:

the 139040
of 72162
to 71395
in 53812
and 53451
said 52080
a 49670
for 26385
mln 25697
it 22050
dlrs 20471
3 18802
on 18704
reuter 18468
pct 17438
is 16554
that 15178
its 15149
from 15015
by 14811
will 14593

20 most frequent terms after:

to 71395
said 52080
mln 25710
on 23758
dlr 23665
reuter 19514
3 18802
pct 17438
that 15257
from 15015
by 14811
at 14233
vs 13867
year 13012
bank 11886
wa 11721
compani 11146
billion 10422
ha 10007
share 9624
would 9048

II.
both dictionary and positional index are python dicts namely a hashmap

dictionary is of the form {'token': line_number_in_index }
This enables super fast search because we only read the necessary line from the index

index is of the form { doc_id:[positons], ..}