Concordia University

COMP 479

Information Retrieval

Fall 2015

Project 2 Report

Presented to :

Dr. Sabine Bergler

by :

Samer Ayoub
26750265

1) Added Features:

- Relevance ranked retrieval system , using OKAPI BM-25 RSV formula where b = 0.75, k = 1.6
- Ranked matches that contains all terms in the query (AND based query)
- OTHERWISE
  Offers the closest matches of the query if at least 1 term of the query exists in the collection (OR based query)
- It offers top N ranked matches, N is a user input bounded by total matches
- Displays document Body as well as new ID, Title for a query result.

2) Approach and Design:

- Adding term frequency:
  Inverted index creation was modified so that each posting includes the term frequency
- Still Using SPIMI to create inverted index and to add terms frequencies.
- Indexer.py :
  - Creates inverted index using SPIMI and save it on disk
  - Read and parses the corpus files
  - Saves block indices on disk
  - Apply all sorts of normalization and stemming passed by user

- Searcher.py:
  - Offers choices to user to recreate the index or retrieve it from disk.
  - Offers all types of normalization in case of recreating the index.
  - Displays the properties of currently existing inverted index once retrieved.
  - Tokenizes user query and finds the postings lists of each term of the query in the index and pass it to the ranker as a list of lists of postings to be processed.
  - Display results passed by Ranker as: document ID, title and body.
  - Offers top N matches or alternative closest match.

- Ranker.py
  - List of postings lists is passed to the `rankDocsForQuery` by the Searcher.py
  - Passes one posting list of one query term at a time to `docScoreOfTerm` to assign a score to every document of that term using OKAPI-BM25 RSV formula and return it back to `rankDocsForQuery`.
  - `rankDocsForQuery` passes 2 copies of each return List of (docID, score) to the two methods `combineScores` and `freeCombineScores` to merge and combine scores according to AND, OR boolean operations and return it back as dictionaries to `rankDocsForQuery`.
  - `rankDocsForQuery` sorts Both dictionaries in non-increasing order according to assigned docs scores to be returned to the Searcher.py

3) Aborted ideas:

- Implementing Zones for Titles, authors
- Saving queries results on disk for faster retrieval.

4) Learned lessons:

- It gave me the chance to design, practice and test probabilistic search techniques.
- Implementing fast algorithms in merging scores of several postings list.
- more practicing for python.
- Implementing 2 types of search queries (AND & OR)

5) Checking the effect of changing OKABI formula parameters

K [1.2, 2]	B [0,1]	Top 5 Rankings for "George Bush" Doc ID
1.2	0.5	08593 20719 20891 16780 04008
1.2	0.9	08593 20719 20891 16780 04008
1.9	0.5	08593 20719 20891 16780 04008
1.6	0.75	08593 20719 20891 16780 04008
1.9	0.9	08593 20719 20891 16780 04008

It makes no difference for the top ranking matches

6) Basic Queries:

Query 1: Democrats' welfare and healthcare reform policies

```
Enter r to Retrieve the inverted index or c to create it     -----> r
======================================================================
                Retrieving Inverted Index
======================================================================

*************** Existing Inverted Index Properties: ****************

Remove Punctuation :                                  False
Case Fold :                                           False
Filter Numbers :                                      False
Filter StopWords :                                    False
Stemming :                                            False
Memory Block Size :                                   128  Kb
Total Number of Tokens :                              2939024  tokens
Total Number of dictionary terms:                     85791  terms
Total number of postings:                             627764  postings

**********************************************************************
search     ?    (y/n)    ----->                    y

Please, enter your query term or phrase    -----> Democrats' welfare and healthcare reform policies
Sorry, no matching Document IDs were found corresponding to your query.
Do you want a match that may not include all query terms ??    (y/n)    -----> y
search returned  14333  results
Top Number of Documents desired     ----->        1

Matching documents IDs, Titles and Bodies:
=======================================<<< Search Results  >>>=======================================
Document ID:          01895
Title:                REAGAN ADMITS IRAN ARMS OPERATION A MISTAKE
                    <<< Body >>>
    President Reagan, fighting to regain
public confidence in the wake of the Iran arms scandal,
admitted tonight that the clandestine operation wound up as an
arms-for-hostages deal and, "It was a mistake."
    "When it came to managing the NSC (National Security
Council) staff, let's face it, my style didn't match its
previous track record," Reagan said in a television address to
the American people.
    "I have already begun correcting this," he added in his
prepared remarks.
    Reagan's speech, widely regarded as critical to his hopes
of repairing his presidency, was his first detailed response to
last week's scorching Tower commission report on the secret
sale of arms to Iran and diversion of profits to U.S.-backed
contra rebels in Nicaragua.
    Reagan said he had been silent on the scandal while he
waited for the truth to come out and admitted, "I've paid a
price for my silence in terms of your trust and confidence."
    He said that a few months ago, he told the American people
he did not trade arms for hostages in the 18-month covert
operation.
    "My heart and my best intentions still tell me that is true,
but the facts and the evidence tell me it is not," Reagan said.
    "There are reasons why it happened, but no excuses. It was a
mistake," he said.
    Reagan again said that the original Iran initiative was to
develop relations with those who might assume leadership in a
post-Khomeini government.
```

"It's clear from the Board's report however that I let my personal concern for the hostages spill over into the geo-political strategy of reaching out to Iran.

"I asked so many questions about the hostages' welfare that I didn't ask enough about the specifics of the total Iran plan," he said.

The commission, headed by former Republican Sen. John Tower, said Reagan's "intense compassion" for Americans being held by pro-Iranian groups in Lebanon had resulted in an unprofessional and unsatisfactory policy.

It portrayed 76-year old Reagan as a man who did not know or care much about the wide-ranging, probably illegal activities of his National Security Council (NSC) staff, which hatched the operation.

Reagan said he endorsed all of the Tower commission's recommendations about the running of the NSC, adding, "In fact, I'm going beyond its recommendations, so as to put the house in even better order."

He noted that he had appointed former Senate Republican leader Howard Baker as his new chief of staff and said he hoped Baker would help him forge a new partnership with Congress, "especially on foreign and national security policies."

He said his new national security adviser, Frank Carlucci, was rebuilding the national security staff "with proper management discipline."

Reagan said that almost half the NSC professional staff now consisted of new people.

He said that FBI Director William Webster, his new nominee to head the CIA, "understands the meaning of 'rule of law'".

Reagan also announced that Tower had agreed to serve as a member of his Foreign Intelligence Advisory Board, which acts as a watchdog on the nation's covert activities.

But he said that he had issued a directive barring the NSC staff itself from undertaking covert operations -- "No ifs, ands, or buts."

Tonight's speech was a far cry from Reagan's initial strong defense of his Iran policy.

In a televised speech last November 13, Reagan called charges that he ransomed hostages and undercut the U.S. war on terrorism "utterly false."

As recently as two months ago in his State of the Union speech, Reagan said that "serious mistakes were made" but defended the basic policy as one that had worthy goals.

By contrast, tonight's speech had an apologetic tone that was a marked departure from Reagan's usual upbeat, confident demeanor.

He said he took full responsibility for his own actions "and for those of my administration."

"As angry as I may be about activities undertaken without my knowledge, I am still accountable for those activities. As disappointed as I may be in some who served me, I am still the one who must answer to the American people for this behavior," Reagan said.

Reagan said the message that the nation should move on had come from Republicans and Democrats in Congress, from allies around the world -- "and if we're reading the signals right, even from the Soviets."

His remark seemed to be a reference to a new Soviet willingness to reach an agreement on eliminating medium-range nuclear missiles in Europe.

Reuter

Query 2: Drug company bankruptcies

```
search     ?    (y/n)    ----->                  y

Please, enter your query term or phrase    -----> Drug company bankruptcies
Sorry, no matching Document IDs were found corresponding to your query.
Do you want a match that may not include all query terms ??    (y/n)    -----> y
search returned  5291  results
Top Number of Documents desired     ----->        1

Matching documents IDs, Titles and Bodies:

====================================<<<    Search Results    >>>====================================
Document ID:          04050
Title:                JAPANESE BANKRUPTCIES DECLINE IN FEBRUARY
                    <<< Body >>>
    Japan's corporate bankruptcies in
February fell 10.8 pct from January to 1,071 cases and total
debts dropped 49.4 pct to 149.40 billion yen, the Tokyo
Commerce and Industry Research Co said.
    February bankruptcies fell 14.9 pct from a year earlier,
the 26th straight monthly decline, and debts fell 54.3 pct.
    The lower number of bankruptcies in February reflected a
relaxation of money market conditions and reduced bill
settlements due to fewer operating days, it said.
    Bankruptcies caused by the strength of the yen against the
dollar totalled 69, or 6.4 pct of those in February, with debts
of 25.52 billion yen, the research firm said.
    This compared with 64 with debts of 125.59 billion yen in
January, it said.
    Currency-linked bankruptcies since November 1985, when the
dollar's depreciation against the yen began to affect Japanese
export-linked firms, totalled 772, with cumulative debts of
660.53 billion yen, it said.
    The value of the yen against the dollar rose to an average
153.49 yen per dollar in February from 184.62 a year earlier.
    Bankruptcies usually decline in the first quarter of the
year due to fewer operating days and for seasonal reasons.
    Bankruptcies are expected to increase in the quarter
starting April 1 due to expectations of slow consumer spending,
low wage increases for the 1987/88 fiscal year which starts in
April, and slow capital spending by manufacturers, the company
said.
    Bankrupcties among export-linked subcontractors will rise
due to a recent shift by major manufacturers to overseas
production, it added.
 REUTER
_____________________________________________________________________________________________

==============================================================================================
```

Query 3: George Bush

```
search     ?    (y/n)    ----->                  y

Please, enter your query term or phrase    -----> George Bush
search returned  13  results
Top Number of Documents desired     ----->        3

Matching documents IDs, Titles and Bodies:

====================================<<<    Search Results    >>>==================================
Document ID:         08593
Title:               HAIG SAYS HE PLANS TO RUN FOR U.S. PRESIDENT
                   <<< Body >>>
    Former U.S. Secretary of State
Alexander Haig has announced his intention to run for
president.
    Haig, a former general and NATO chief who served seven
presidents but never held elected office, said he will formally
announce his bid to run for the Republican presidential
nomination at a news conference later today.
    Haig's aides say he starts far back in the field of
Republican hopefuls, which is currently dominated by
Vice-President George Bush and Senate Minority Leader Robert
Dole, neither of whom have yet declared their intentions.
 REUTER
_____________________________________________________________________________________________

Document ID:         20719
Title:               SAUDI ROLE IN GULF PRAISED BY U.S. OFFICIALS
                   <<< Body >>>
    Saudi Arabian Crown Prince Abdullah
bin Abdul Aziz was thanked by the Reagan administration for his
country's close, and closed-mouthed, cooperation with
Washington in the Gulf, a senior U.S. official said.
    "The Saudis are being very cooperative. It would be nice if
the Saudis would go more public, but it's their real estate,"
said the official who asked not to be named.
    He declined to describe what sort of help the Saudis were
providing, saying that Saudi officials are reluctant to
acknowledge their role in the Gulf where the United States has
stationed forces to protect shipping lanes.
    The prince met Vice President George Bush on Monday after
U.S. naval forces attacked offshore Iranian oil platforms in
what Washington said was retaliation for an Iranian attack on a
ship moored off Kuwait and flying the U.S. flag.
    Asked at the start of the meeting how he felt about the
attack, the prince, who is here on an official visit, replied,
"I believe what the United States has done is their
responsibility as a superpower."
    The senior U.S. official said his remark was an endorsement
of the U.S. attack.
 Reuter
_____________________________________________________________________________________________

Document ID:         20891
Title:               SAUDI CROWN PRINCE MEETS WITH US VICE PRES BUSH
                   <<< Body >>>
    Saudi Arabia's Crown Prince Abdullah
bin Abdul Aziz met for an hour with Vice President George Bush
on Monday after U.S. naval forces destroyed one Iranian oil
platform in the Gulf and raided another.
```

Asked at the start of the meeting how he felt about the attack, the Crown Prince, who is here on an official visit, replied, "I believe what the United States has done is their responsibility as a superpower."

His remark appeared to be an implicit endorsement of the U.S. action, which the Pentagon said came in retaliation for last Friday's Iranian missile attack on a U.S.-flagged Kuwaiti tanker.

Administration officials said Bush had assured the Crown Prince the United States would "stay the course" in the Gulf.

They said Prince Abdullah, who is deputy prime minister of Saudi Arabia and commander of the kingdom's national guard, was "very supportive" of the U.S. role in the strategic waterway.

Before meeting with Bush, the Crown Prince paid a brief courtesy call on President Reagan.

During his stay in Washington, he was also scheduled to meet with Deputy Secretary of State John Whitehead, Defense Secretary Caspar Weinberger and leaders of the House and Senate foreign policy committees.

Reuter

==========

search ? (y/n) -----> y

Please, enter your query term or phrase -----> george bush
search returned 4 results

Please, enter your query term or phrase -----> George bush
Sorry, no matching Document IDs were found corresponding to your query.
Do you want a match that may not include all query terms ?? (y/n) -----> y
search returned 143 results

7) More Queries:

Query 1: Shortage of around 250 mln stg in the money market

```
Enter r to Retrieve the inverted index or c to create it     -----> c
Enter block size in kilo bytes    ----->          64
Remove Punctuation ?     (y/n)    ----->          y
Case fold ?     (y/n)    ----->                  y
Remove numbers ?     (y/n)    ----->             y
Remove stop words ?     (y/n)    ----->          y
Stem using Porter Stemmer ?     (y/n)    ----->  n
=====================================================================
                      Creating Inverted Index
=====================================================================
Inverted Index is created in :     121.61395621299744  seconds
search     ?    (y/n)    ----->                  y

Please, enter your query term or phrase    -----> Shortage of around 250 mln stg in the money market
today
search returned  34  results
Top Number of Documents desired     ----->        3

Matching documents IDs, Titles and Bodies:

====================================<<<    Search Results    >>>==================================
Document ID:          00943
Title:                U.K. MONEY MARKET SHORTAGE FORECAST REVISED UP
                    <<< Body >>>
    The Bank of England said it revised up
its forecast of the shortage in the money market today to
around 500 mln stg from its initial estimate of 350 mln.
 REUTER
_____________________________________________________________________________________________

Document ID:          12523
Title:                U.K. MONEY MARKET GIVEN 40 MLN STG LATE HELP
                    <<< Body >>>
    The Bank of England said it had provided
the money market with around 40 mln stg late assistance. This
takes the Bank's total help today to some 537 mln stg and
compares with its estimate of a 700 mln stg shortage.
 REUTER
_____________________________________________________________________________________________

Document ID:          17523
Title:                U.K. MONEY MARKET GIVEN 25 MLN STG LATE ASSISTANCE
                    <<< Body >>>
    The Bank of England said it provided the
money market with late assistance of around 25 mln stg.
    This takes the Bank's total help today to some 137 mln stg
and compares with its latest forecast of a 150 mln stg
shortage.
 REUTER

_____________________________________________________________________________________________


=============================================================================================
```

Query 2: earnings nine segments sharp improvement

```
search     ?    (y/n)    ----->                  y
Please, enter your query term or phrase    -----> earnings nine segments sharp improvement
search returned  1  results
Top Number of Documents desired     ----->        5

Matching documents IDs, Titles and Bodies:
====================================<<<    Search Results    >>>====================================
Document ID:          12949
Title:                POPE AND TALBOT <POP> SEES HIGHER 1ST QTR NET
                    <<< Body >>>
    Pope and Talbot Inc said it
expects first quarter earnings to total about one dlr per
share, compared with a year-earlier loss of nine cts per share.
    Each of the company's business segments contributed to the
sharp improvement, Pope and Talbot said.
    The wood products company also said it expects to release
first quarter results later this month.
 Reuter
```

Query 3: love drug abuse

```
search     ?    (y/n)    ----->                  y
Please, enter your query term or phrase    -----> love drug abuse
Sorry, no matching Document IDs were found corresponding to your query.
Do you want a match that may not include all query terms ??    (y/n)    -----> y
search returned  190  results
Top Number of Documents desired     ----->        1

Matching documents IDs, Titles and Bodies:
====================================<<<    Search Results    >>>====================================
Document ID:          16829
Title:                WALL STREET STOCKBROKERS CHARGED IN DRUG ARRESTS
                    <<< Body >>>
    Sixteen Wall Street stockbrokers were
charged with dealing cocaine, some at a brokerage house federal
officials said was deeply involved in drug trafficking as well
as trading violations.
    A total of 19 people were arrested in the federal
investigation, eight of them affiliated with Brooks, Weinger,
Robbins and Leeds, Inc.
    Manhattan U.S. Attorney Rudolph Giuliani said the
investigation marked the start of a wider probe into widespread
drug abuse in the New York financial district.
    "This case and the implications of it are quite serious.
This is the beginning of this whole area of investigation," he
said at a news conference announcing the indictments.
    The arrests marked the latest phase in a series of scandals
that have rocked Wall Street, most of them involving illegal
insider trading.
    The federal probe coincides with a police investigation  in
which 114 people, including messengers, a security guard and a
New York Telephone company executive, have been arrested for
alleged drug dealing and drug abuse, police said today.
    A request by Giuliani's office for a warrant to search two
Brooks, Weinger offices alleges numerous instances of stock
manipulation as well as other securities law violations.
 Reuter
```

Query 4: abcdefgh

```
search     ?    (y/n)    ----->                  y

Please, enter your query term or phrase    -----> abcdefgh
Sorry, no matching Document IDs were found corresponding to your query.

search     ?    (y/n)    ----->                  n
Goodbye !!!
```