Information Retrieval Models

Brian Mutisyo

Harvard University

Fall 2022

Contents

1 INFORMATION RETRIEVAL

Information retrieval (IR) is defined as the process of finding information or materials of an unstructured nature that satisfies a particular information need. The materials which we seek information from are always most often organized as documents in large collections and these documents are mostly stored as unstructured text [1]. The most common form of information retrieval is web search where a particular query is made to find information of a particular need. The same system is applied to tasks like email search functionalities, digital library systems, legal systems and even medical databases where keyword searches are performed frequently.

The main goal of information retrieval is to help users find relevant information from an organized collection of documents. The information being retrieved tends to vary and could be in the form of text retrieval, image retrieval, video retrieval and even audio. recently, modern retrieval systems have employed the use of more statistical algorithms that have facilitated retrieval of non-textual information.

The study in of itself involves a wide range of topics like NLP, machine learning and data mining. These techniques develop fundamental algorithms and models that help to retrieve information from large collection effectively and thus information retrieval has become a fundamental component of modern computing. The vast information in the internet era call for effective information retrieval techniques which make it easier to access information in a meaningful way.

1.1 Information Retrieval Models

Information retrieval models are simply mathematical frameworks that are used to represent and analyze the process of accessing and retrieving information. They take into account a number of factors like characteristics of a user's query, relationships between documents found in the same or different collection, and the importance of certain words in a particular document. These models may represent text in different ways including vector form and notation and where text numbers are large forming high dimensional spaces. Information retrieval models are thus critical tools for organizing and accessing large collections of information and continue to shape how we interact with textual data.

In the next section we introduce some of the well-know information models and retrieval strategies for documents.

1.2 The Boolean Model

The boolean model in information retrieval generally uses boolean logic to retrieve relevant information from a database or a collection of given documents, and is based on set theory and boolean algebra. The boolean model suggests that documents are sets of terms and queries are boolean expressions on terms i.e. both the documents to be searched and the user's query are conceived as a set of terms[1].

The most common boolean operators used in information retrieval are AND, OR, and NOT. Generally, words are logically combined with the Boolean operators AND, OR, and NOT. For example, we might have two statements like x or y which mean that either x OR y must be satisfied i.e. at least one of these statements must be met. The Boolean expression X and Y goes further to say that both x AND y must be satisfied.

The Boolean model predicts that each document is either relevant or non-relevant - where documents are labeled as relevant or irrelevant depending on whether they contain the words or not. Since a document is judged as relevant or irrelevant based on the search query, there is no concept of a "partial match" between documents and queries[1].

Boolean models are simple and efficient but they do not provide ranking of the results based on relevance of the documents to the query. They are quite common and widely used where the search space is small and well-defined. Additionally, their inability to identify partial matches can lead to poor performance[2].

1.3 The Probabilistic Model

In information retrieval, the probabilistic model employs the use of probability theory to estimate the relevance of a document to a given query. The probabilistic model calculates the similarity between the query and the document in terms of probabilities. This framework relies on the assumption that the presence of certain words in a document is a great indication of the relevance of the document to the query, and the absence of it signifying irrelevance to the particular query. [3].

Typically, a probabilistic model uses Bayes Theorem and Maximum likelihood estimation to calculate the final score/measure of relevance. The probabilities and weights for each term are calculated based on the Probability Ranking Principle [3] which just assigns probabilities to terms in the query and using these as 'evidence' in calculating the final probability that a document is relevant to the query. At the end, the documents are ranked based on the combined probability scores and the most relevant documents are those that score the highest on this scale.

1.4 The Vector Space Model

The vector space model is one of the most widely used language model in information retrieval. The basic functionality of this model is that words contained in documents are represented as vectors in some n dimensional space. This space has special properties; for instance the *set of words* is that which contains only the unique words in the given document collection.

Algebraically, the terms or the words are the axes of the space. And documents are vectors from the origin pointing to that point in that space. As the number of words grow, the complexity of the vector space increases. For instance if we have, a thousand words in total, we get a vector space with a thousand dimensions. As the number of words grows, the size of the vector space grows by the same proportion. Another key property of vectors in the vector space model is that they are mostly *sparse vectors*. A representation of such a vector becomes one whose most entries are zero entries because of the documents in the vector space only contain a few hundred or thousand words contained in the vector space. [1]

1.5 The Language Model

The main idea behind language models as used in information retrieval is that documents can be ranked based on their likelihood of generating the query [3]. In their simplest form, language models can be defined as a tool or a mechanism for generating a piece of text using a distribution of all the possible word sequences [3]. To implement this system, we use a set of probabilistic tools to model and estimate the probability that a given document is relevant to the query. These probability distributions for the documents are used to predict the likelihood of observing the particular query terms. So in other terms instead of using the query to predict he likelihood of observing a document, we use the text of document to predict the probability of observing the query [2]

Language models are used in many application such as speech recognition, machine translation and part-of-speech tagging [2] and they use natural language processing techniques to understand the meaning of a search query and retrieve the relevant documents.

2 MODELLING INFORMATION MODELS

2.1 Baseline Study: Boolean Retrieval Model

The implementation of a Boolean Model is based on whether or not the documents contain the query terms so considerations of whether the keywords are present or absent in a document or title are made. More specifically, the AND operator retrieves documents that contain all of the specified search terms,

the OR operator retrieves documents that contain at least one of the search terms, and the NOT operator excludes documents that contain a specific term. For instance, all elements of the term document matrix are either 1, to indicate presence of the term in the document, or 0, to indicate the absence of the term in the document.

2.1.1 Processing Boolean Queries

For processing the Boolean queries and documents in Boolean Model, we will consider the use of an inverted index.

2.1.2 Inverted Index

The inverted index is the key data structure that underlies all modern information retrieval systems. An inverted index is a database index storing a mapping from content, such as words or numbers, to its locations in a table, or in a document or a set of documents. The purpose of an inverted index is to allow fast full-text searches, at a cost of increased processing when a document is added to the database. [2]. Instead of scanning the entire collection of documents, the text is preprocessed and all unique terms are identified . This list of unique terms is referred to as the index. For each term , a list of documents that contain the term is also stored. This list is referred to as posting list.

2.1.3 Construction of the inverted index

An inverted index consists of two components - a list of each distinct term referred to as the index and a set of lists referred to as posting lists. The index represents the words that are in the collection of documents and the postings lists the document id's of the documents in our collection.

We follow the following set of procedures to construct an inverted index for the dataset:

1. First we tokenize the text into a list of tokens. This was carried out by first obtaining the list of sentences in any given document. The list of sentences were tokenized further into a list of individual words contained in the document.

2. The list of words we obtain in 1 above are processed further using linguistic processing techniques. We remove stopwords and identify a list of all the unique words in each document. These form our indexing terms.

3. Within a document collection, we assume that each document has a unique serial number, known as the document identifier (docID). These numbers can be assigned manually if not provided, as long as they are all unique identifiers for their respective documents.

4. Next we map each of our indexing terms to the documents in which they appear in. We create a linked list with pointers that point to all the documents containing a given word.

3 METHODOLOGY

The document retrieval I build in this project works in a two fold manner. One side is where the user inputs query while the other one is involved with presenting the user with the information that best matches that query. Typically, the input of the user is pre-processed into a series of machine understandable format and the same is done to the series of documents that we wish to match the query to. There are a couple of ways to do this, but here the implementation of this system can be broken down into 3 key steps:

1. Query Processing
2. Document Processing
3. Matching of queries to the relevant documents

3.1 Document Processing

Normal text is hard to be understood by computers hence the need for a machine friendly representation. Additionally, raw text data may contain unimportant information that might make it difficult to understand the text. Document processing involves presenting textual data into this understandable format. For this part I performed the document pre-processing in the following main steps. The data that we have is a collection of newspaper articles that contain the content and title of newspaper articles from various news media outlets. It is not labelled, so we don't know which documents are relevant and which ones are not.

3.1.1 Tokenization

Tokenization is the process of breaking down larger chunks of text like paragraphs into smaller pieces like sentences. Sentences can further be broken down into words. These smaller tokens are referred to as *tokens* and result in a set of words representing the whole document.

The general approach to splitting a paragraph into its individual sentences employs use of segmentation strategies like splitting around punctuation marks like full stops or by splitting on whitespaces to turn sentences into words. The former becomes problematic due to varying sentence ending delimiters like fullstops, question marks and exclamation marks. This problem could be solved by using stemming

algorithms which typically would do better in such tasks.

3.1.2 Normalization

Normalization involves the conversion of the text into a standardized consistent format. This step is essential since it helps the later processes performed on the same piece of text when there is a need to *recognize* and process the same text. Normalization involves converting all text to the same case (upper or lower), removing punctuation, converting numbers to their word equivalents or even removing unnecessary numbers.

For an even more homogeneous text output, we use normalization techniques like stemming and lematization in our analysis.

3.1.3 Stemming

Stemming can be described as the process of trimming a word down to its stem. It involves removing parts of the word that inflate it like the prefix, the suffix, the infixes and the circumfixes. The output of this gives the stem of the word. For instance, in our analysis we get the following stemmed words from a random sample of words:

Word	**Prefix/Suffix**	**Stem**
importance	**-ance**	import
laughing	**-ing**	laugh
uncomfortable	**un-** + **-able**	comfort
studies	**-es**	studi

The stemming process we use here is not robust for normalization. Not all the words can be stemmed into their right root word. Stemming produces words which sometimes carry no meaning, or carry a different meaning from the original word that was stemmed. Stemming cannot connect words with different forms, but of the same grammatical constructs like is, am and be. All these are stemmed into different words and has the effect of changing the context of the stemmed words resulting in false positives.

3.1.4 Lematization

As opposed to stemming, lemmatization employs the use of word structure, vocabulary, part of speech tags, and grammar relations to trim down word to their root form. The output of lemmatization is called the word's *lemma.* Lemmatization is more robust in the sense that words with the same meaning but with different canonical forms are mapped to the same word and this reduces the false positive rate problem that we faced earlier with stemming. Take for instance, a random sample of lemmatized words from our dataset shown below:

Original Word	Stemmed word	Lemmatized word
changing	chang	change
wrote	wrot	write
better	bet	good
studies	studi	studies

It is important to lemmatize and stem words for document retrieval. Stemming reduces the number of terms used in the *inverted index* thereby saving on storage space. Additionally, it improves the recall of the relevant documents. For example if a query has the search keyword "analyze", the user might also be interested in documents which contain *analysis, analyzing, analyzer* or *analyzed*. For our system to match all these documents, the query and the documents terms are both stemmed to analy- for a better match. However, a downside of this is reduced precision as we shall see in later chapters.

3.1.5 Removing stopwords

One of the reasons why we remove stopwords in our system is to remove words which are otherwise regarded as non-informative for the user's request. These words tend to add little to no value in the retrieval process and removing them reduces the number of unique words/tokens that need to be indexed in the document. The most common stopwords removed in our case are conjunctions, interjections, preposition, pronouns and 'to be' verb forms.

3.2 Implementation of a vector space model

The key characteristic in the implementation of this model is the representation of the words within the documents as vectors, and doing the same for the queries from the users. Like we saw earlier, each term represents a dimension in the vector space we create. The goal of doing this is to find a way of measuring the similarity of a given query vector to a document vector. Mathematically, this can be done using cosine similarity where the angle between two vectors is calculated.

3.2.1 Handling Queries & Documents

In a vector space model, queries are also treated the same way as how documents are treated and are also represented as vectors in the space. Ranking of documents is calculated by *proximity* of a document to the query vector in the entire vector space. The proximity can be calculated using the euclidean distance between the query vector and the document vector. However, classical distances like euclidean distance do not do well with sparse data like this one. Using this distance metric tends to make large documents appear irrelevant to most queries, which are typically short [1]. As an alternative, angles are used instead of distance. Documents are ranked according to the angle between the query and the document. For

instance, the angle between two documents of maximal similarity is 0 and so on. Generally, ranking can be done in two ways: [1]

- Rank documents in decreasing order of the *angle* between the query and the document.
- Rank documents in increasing order of the *cosine of the angle* between the query and the document.

To compute a *similarity score* between query and the set of documents that we have and rank order of the documents based on their relevance to the query, it is typical to use weighted vectors. [4]. There are various to compute these weights, and in this implementation we will use cosine similarity which uses angles and term frequency - inverse document frequency(tf-idf) which calculates a similarity score based on the dot product.

3.2.2 Cosine Similarity

Cosine similarity is a measure of the similarity between two vectors in a vector space. In an information retrieval model it is used to calculate the relevance of a document to a query. It is also often used together with other techniques such as term weighting and term frequency-inverse document frequency (TF-IDF) for more robust results. To calculate it, we vectorize both the query and the document, and the typical convention used is to represent each document as a vector of tf-idf weights. Then we compute the dot product between the query and the document vector to get a similarity score for each document. This dot product value represents After this, we just sort the documents by similarity scores and the most relevant are the highest-scoring documents appearing first in our list.

3.2.3 term frequency - inverse document frequency(tf-idf)

Tf-idf is a numerical measure of the importance of a word in a document or a collection of documents. In information retrieval, it is used to help identify which words are the most relevant in a given document. The main idea behind tf-idf is that common words in a document are less important compared to the less common words in the same document. This is because the common words, such as 'the', 'and' and 'because' do not provide much information to understand the text. On the other hand, less common words are highly likely to be specific to the topic that the document describes, hence end up providing more information.

More practically in our system, we use tf-idf to rank the importance of words in a document relative to the entire corpus. For example, if a query is submitted to the system, we perform two main actions. The tf-idf values for each word in the query are calculated and compared to the tf-idf values for each word in the documents in the corpus. The documents with the highest tf-idf values for the query words

are typically ranked higher in the search results.

4 RESULTS AND DISCUSSION

4.1 Using tf-idf vector scores as a similarity measure

When using tf-idf as a similarity measure, I implemented a simple retrieval system for a vector space model and use these measure to rank documents. Here are the results when the user types in the query ***financial markets at wall street***.

```
Enter a search query:
  financial markets at wallstreet
Processing Querry...please wait..

--------------------------------
Displaying search results

DocumentID: 137487
Title of Retrieved Document 1: China's red tide arrives on Wall Street's shore
Similarity to query: 0.3669

Sumary of Retrieved Document:
----------------------
 The stock market's   despite the fact that the party was dreadful. From the close of trading in 2015, the S&P 500 has plummeted 6. 9 percent, the Dow free fell 7. When we include December's dismal slide, stocks are already into correction territor
y, down 10 percent. Many may be wondering just which other culprits helped trigger this New Year's 2016 meltdown. For starters, let's consider the fact that the world economy is growing more slowly than we all would like. But the world's markets ar
e also reacting to the US markets after eight years of   percent economic growth, flaccidly bouncing between zero percent and 2 percent. China doesn't even follow our accounting standards, nor does it have any legitimate market structure. One minut
e it has circuit breakers, the next it cancels them, and the next it closes its mysterious markets for a day. There are no clear, bonafide rules, and only a complete fool would invest either long or short in China on his or her own. It's not pleasa
nt to think that the US markets get pushed around by a   economy. And, speaking of Communists, North Korean ruler Kim Jong Un's nuclear bomb "science experiment" last week only added more tension to the situation. No president who has served two fu
ll terms has experienced a more erratic eight years of gross domestic product expansion, with not a single year's growth exceeding 3 percent. Take a look at just the past two years' growth: :   . In fact, if the Atlanta Fed's   2015 estimate is cor
rect, of the last eight quarters, three were below 1 percent. That is not the sign of a strong economy, and therein lies the stock markets' major problem. Unless you need the money tomorrow, stick to a   strategy  —   this isn't our first correcti
on and won't be our last, and selling into correction territory has typically not proven to be a fruitful endeavor.


DocumentID: 78640
Title of Retrieved Document 2: Why Brexit Is So Bad for the Economy
Similarity to query: 0.3493

Sumary of Retrieved Document:
----------------------
 Great Britain's currency, the pound, had fallen to its lowest levels since 1985, and the FTSE and DAX plummeted. In the U. S. markets opened in the red, gold (a commodity that many investors flee to at times of uncertainty) was up, and traders aro
und the globe prepared for a volatile day amid the question of what the future will look like with the U. K. untethered from the European Union. It's understandable, then, that Great Britain's historic move to shed its formal integration with Europ
e after almost six decades and the resignation of its prime minister in one fell swoop would send markets into a bit of a frenzy. While the country won't have to face the daunting task of creating a new currency (a point of serious concern during G
rexit and the Scottish referendum) many of its political and economic policies remain intertwined with the larger European body. Great Britain will be the first country to actually leave the European Union, and the first to have to make sense of th
e relatively brief guidance on how to do so provided by Article 50. The severing of this tie means that important tenets of governing, like major laws, regulations and trade arrangements will need to be updated or renegotiated independently, both w
ithin the EU and with trade partners outside. Then there's the fact that Great Britain will need a new leader to oversee these new changes, since Prime Minister David Cameron —  who called for the vote in the first place —  stepped down after its
conclusion. Part of the explanation for the markets' reaction Friday morning is the fact that many didn't anticipate that the vote would actually swing in the direction of those advocating leaving the EU. All of that said, the markets have so far
not been nearly as volatile as they were during periods of actual economic crisis, such as the financial meltdown of 2008. And as Justin Wolfers noted, American markets moved more on the news of Brexit than they have on news of U. S. presidential
elections over the past 60 years. (Of course there are those who may outright benefit from the melee, betting against safety and stability by doing things like shorting British currency or U. K. There's reason to fear that the spiral and volatilit
y, at least for the U. K. will not end anytime soon. Additionally some major economic players and investors, such as global investment banks, may shift their global strategies by moving jobs out of the U. K. after publicly supporting the campaign
to stay attached to the E. U.


DocumentID: 81495
Title of Retrieved Document 3: Wall Street Is Also Relieved the FBI Cleared Clinton
Similarity to query: 0.3466

Sumary of Retrieved Document:
----------------------
 Until now, the clearest indication of Wall Street's preference for a Hillary Clinton presidency revealed itself after the first debate, which Clinton was widely viewed to have handily won. "Soon after the debate ended, stock markets celebrated the
news of Trump's loss," my colleague Derek Thompson wrote last month. "Markets in the U. S. U. K. and Asia soared, the price of crude oil rose, and the currencies of America's closest trading partners, such as Mexico and Canada, ticked up as well. F
ollowing the debate, as well as Trump's "  talk" scandal, the polls and the markets tightened again after FBI Director James Comey announced, 11 days before the election, that the agency would look into more of Clinton's emails. Cause and effect de
finitely *not* totally clear, but Clinton went from 81% in our forecast   letter to 65% now. According to AP, the Dow Jones index futures "jumped about 200 points," ending a   slide that started with the renewed negative attention Comey's letter di
rected toward the Clinton campaign late last month. "The SP 500 tumbled about 20 points in the 40 minutes after Comey's first letter was made public," noted Bloomberg. On Monday morning, after Comey's most recent announcement, the rally continued,
with the Dow surging more than 300 points, the SP 500 up more than 40 points (more than 2 percent) and the Nasdaq 2. And the Mexican peso, which has tended to falter as Trump's chances improve, surged about 2. It is more than a little ironic that
between the two major party candidates, Clinton is the one who has most forcefully called for reforming Wall Street and raising taxes on the wealthiest Americans. Nevertheless, Clinton seems to offer markets what Trump, seemingly by design, does n
ot: predictability. "Markets want continuity and essentially they want what they have priced in and both point towards Clinton," one market analyst told Reuters. " Regardless of the outcome of tomorrow's election, the markets, like millions of Ame
ricans, will need some time to recover.
```

For comparison, I ran the system with another query: ***The ethics of machine learning and state of artificial intelligence*** and the following output is generated showing the top documents and their summaries.

```
Enter a search query:
  financial markets at wallstreet
Processing Querry...please wait..

--------------------------------
Displaying search results

DocumentID: 137487
Title of Retrieved Document 1: China's red tide arrives on Wall Street's shore
Similarity to query: 0.3669

Sumary of Retrieved Document:
-----------------------
 The stock market's   despite the fact that the party was dreadful. From the close of trading in 2015, the S&P 500 has plummeted 6. 9 percent, the Dow free fell 7. When we include December's dismal slide, stocks are already into correction territor
y, down 10 percent. Many may be wondering just which other culprits helped trigger this New Year's 2016 meltdown. For starters, let's consider the fact that the world economy is growing more slowly than we all would like. But the world's markets ar
e also reacting to the US markets after eight years of   percent economic growth, flaccidly bouncing between zero percent and 2 percent. China doesn't even follow our accounting standards, nor does it have any legitimate market structure. One minut
e it has circuit breakers, the next it cancels them, and the next it closes its mysterious markets for a day. There are no clear, bonafide rules, and only a complete fool would invest either long or short in China on his or her own. It's not pleasa
nt to think that the US markets get pushed around by a   economy. And, speaking of Communists, North Korean ruler Kim Jong Un's nuclear bomb "science experiment" last week only added more tension to the situation. No president who has served two fu
ll terms has experienced a more erratic eight years of gross domestic product expansion, with not a single year's growth exceeding 3 percent. Take a look at just the past two years' growth: :  . In fact, if the Atlanta Fed's   2015 estimate is cor
rect, of the last eight quarters, three were below 1 percent. That is not the sign of a strong economy, and therein lies the stock markets' major problem. Unless you need the money tomorrow, stick to a   strategy  —   this isn't our first correcti
on and won't be our last, and selling into correction territory has typically not proven to be a fruitful endeavor.


DocumentID: 78640
Title of Retrieved Document 2: Why Brexit Is So Bad for the Economy
Similarity to query: 0.3493

Sumary of Retrieved Document:
-----------------------
 Great Britain's currency, the pound, had fallen to its lowest levels since 1985, and the FTSE and DAX plummeted. In the U. S. markets opened in the red, gold (a commodity that many investors flee to at times of uncertainty) was up, and traders aro
und the globe prepared for a volatile day amid the question of what the future will look like with the U. K. untethered from the European Union. It's understandable, then, that Great Britain's historic move to shed its formal integration with Europ
e after almost six decades and the resignation of its prime minister in one fell swoop would send markets into a bit of a frenzy. While the country won't have to face the daunting task of creating a new currency (a point of serious concern during G
rexit and the Scottish referendum) many of its political and economic policies remain intertwined with the larger European body. Great Britain will be the first country to actually leave the European Union, and the first to have to make sense of th
e relatively brief guidance on how to do so provided by Article 50. The severing of this tie means that important tenets of governing, like major laws, regulations and trade arrangements will need to be updated or renegotiated independently, both w
ithin the EU and with trade partners outside. Then there's the fact that Great Britain will need a new leader to oversee these new changes, since Prime Minister David Cameron —  who called for the vote in the first place —  stepped down after its
conclusion. Part of the explanation for the markets' reaction Friday morning is the fact that many didn't anticipate that the vote would actually swing in the direction of those advocating leaving the EU. All of that said, the markets have so far
not been nearly as volatile as they were during periods of actual economic crisis, such as the financial meltdown of 2008. And as Justin Wolfers noted, American markets moved more on the news of Brexit than they have on news of U. S. presidential
elections over the past 60 years. (Of course there are those who may outright benefit from the melee, betting against safety and stability by doing things like shorting British currency or U. K. There's reason to fear that the spiral and volatilit
y, at least for the U. K. will not end anytime soon. Additionally some major economic players and investors, such as global investment banks, may shift their global strategies by moving jobs out of the U. K. after publicly supporting the campaign
to stay attached to the E. U.


DocumentID: 81495
Title of Retrieved Document 3: Wall Street Is Also Relieved the FBI Cleared Clinton
Similarity to query: 0.3466

Sumary of Retrieved Document:
-----------------------
 Until now, the clearest indication of Wall Street's preference for a Hillary Clinton presidency revealed itself after the first debate, which Clinton was widely viewed to have handily won. "Soon after the debate ended, stock markets celebrated the
news of Trump's loss," my colleague Derek Thompson wrote last month. "Markets in the U. S. U. K. and Asia soared, the price of crude oil rose, and the currencies of America's closest trading partners, such as Mexico and Canada, ticked up as well. F
ollowing the debate, as well as Trump's "  talk" scandal, the polls and the markets tightened again after FBI Director James Comey announced, 11 days before the election, that the agency would look into more of Clinton's emails. Cause and effect de
finitely =not= totally clear, but Clinton went from 81% in our forecast   letter to 65% now. According to AP, the Dow Jones index futures "jumped about 200 points," ending a   slide that started with the renewed negative attention Comey's letter di
rected toward the Clinton campaign late last month. "The SP 500 tumbled about 20 points in the 40 minutes after Comey's first letter was made public," noted Bloomberg. On Monday morning, after Comey's most recent announcement, the rally continued,
with the Dow surging more than 300 points, the SP 500 up more than 40 points (more than 2 percent) and the Nasdaq 2. And the Mexican peso, which has tended to falter as Trump's chances improve, surged about 2. It is more than a little ironic that
between the two major party candidates, Clinton is the one who has most forcefully called for reforming Wall Street and raising taxes on the wealthiest Americans. Nevertheless, Clinton seems to offer markets what Trump, seemingly by design, does n
ot: predictability. "Markets want continuity and essentially they want what they have priced in and both point towards Clinton," one market analyst told Reuters. " Regardless of the outcome of tomorrow's election, the markets, like millions of Ame
ricans, will need some time to recover.
```

Here is the table of results that we obtain if we look at the similarity scores for the 5 documents that rank the highest for the first query term.

DocID	DocTitle	SimilarityScore
137487	China's red tide arrives on Wall Street's Shore	0.3669
78640	Why Brexit Is So Bad for the Economy	0.3493
81495	Wall Street Is Also Relieved the FBI Cleared Clinton	0.3466
44607	George Soros Cites 'Buyers Remorse' After Brexit	0.3446
67762	Here's your complete preview of this 4-day week's big market-moving events	0.3385

Table 1: Results for query *financial markets at wallstreet*.

Here is the table of results that we obtain if we look at the similarity scores for the 5 documents that rank the highest for the second query term - *the ethics of machine learning and state of artificial intelligence*.

DocID	DocTitle	SimilarityScore
75603	Amazon is Making It Easier for Companies to Track You	0.3805
31087	Tech Billionaires Create Fund to Prevent Robot Apocalypse - Breitbart	0.3200
71384	Apple is facing a crisis of salesmanship	0.3021
71054	No need for a CS degree from Stanford work on one of Google's hottest teams	0.2956
69681	Here are the most exciting things Google announced at its giant conference	0.2853

Table 2: Results for query *the ethics of machine learning and state of artificial intelligence*.

4.1.1 Discussion

Above we look at the top 5 results returned by the system for two different queries. The results above present a few things to note. The similarity scores for the query *financial markets at wallstreet* are all

less than 50%, with the highest similarity score being 0.3669. For the second query *the ethics of machine learning and state of artificial intelligence*, the system returns similarity scores that are also less than 50%, with the top ranked document representing a 38% match with the query.

The implication of this is that we do not get the most relevant document to the search query. The most relevant in our case would be documents which have a similarity score that is as close as possible to 1. However much, the retrieved documents are not that different from the query and actually do contain some of the information that the user would wish to know as shown by the summaries. The summaries show that the user can still find the some information related to the query as far as this document collection is concerned.

Another key finding we see above is the closeness in similarity scores for the various retrieved documents. From table 1, documents in ranks 2, 3 and 4 differ in similarity score by really small margins. It is hard to know why this is the case, but the easiest reason is that words in both documents are most likely similar, even though they are used to describe different topics.

5 Conclusions

To summarize the conclusions above, we looked at the different information retrieval models that can be used to retrieve information, like the Boolean Model, the Proabilistic Model, the Language Model and the the Vector Space Model. We looked at various ways of pre-processing text (a key step in building an IR system) like tokenization, lematization and stemming. We looked at the statistical and modelling principles governing how these information retrieval systems work and we implemented a vector space model for a database of 10,000 newspaper articles.

The system we built calculates a similarity score to match documents to the relevant queries by use of vector algebra and returns a set of results ranked using this score to the user.

We also found out that the retrieved documents for some random queries that we used provide similarity scores that are quite low. This can be attributed to a number of reasons:

- *Scope*: The newspaper articles dataset that we use in this project is not representative of all the information needs from the user, and this limits the information that is relevant to the queries provided.
- *Bias*: Newspaper articles tend to be biased because they focus on certain topics or particular perspectives, and this may not include all the information that is relevant to a particular query.
- *A small collection of documents to work with*: Typically large information retrieval models contain

millions of documents and compared to the small dataset that we had, our model typically underperforms due to less data.

6 Strengths, limitations & future work

One of the main limitations of using the tf-idf model for information retrieval is that the model does not take into context the words in the document. The model takes the words as they are, based on the bag of words that we build, without considering the words around it. As a result the model is unable to capture with great accuracy the meaning of words that have different meanings when used in different contexts. Additionally, failure to consider synonyms and related words also reduces the effectiveness of the system in retrieving documents that talk about the same concept but with a different set of words.

This model's core assumption also fails. The tf-idf model is based on the assumption that rare words (less frequently occurring words) bear more meaning and importance than the less common words. However, this is not always the case because some words base their importance on their positions in a document or depending on their use - even though these words may have higher frequencies. As as result, such words which are commonly occurring because they talk of common topics are ignored and therefore the system may not be able to effectively retrieve such documents.

Lastly, the information retrieval system presented above does not incorporate user feedback. This makes it hard for the results to be tailored to a user based on their previous queries and their information needs.

However, this model is relatively simple to implement compared to the language and probabilistic models we saw in earlier sections. With this simplicity it is still able to fully capture the importance of words based on their frequency and how commonly they have been used across the documents. This is particularly helpful especially when retrieving documents which do not have the exact word as they appear in the query.

To make it this project more robust, more additional signals and features should be incorporated into the model. For instance, including the position of words in a document into the model may capture more information about the meaning of words and improve overall effectiveness of retrieving relevant documents. For instance, if a user sends more than one query, similarities between the current and previous queries should be calculated and this feedback used in the model to improve performance. Although this was outside the scope of this project, user feedback should also be incorporated into the model to make the system more flexible to user needs.

Thanks to Professor Hanspeter Pfister and Dr. Zhutian Chen for their generous support in supervising

this research study.

References

[1] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press, 2008.

[2] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology behind Search*. USA: Addison-Wesley Publishing Company, 2nd ed., 2011.

[3] D. A. Grossman and O. Frieder, *Information Retrieval: Algorithms and Heuristics*. The Kluwer International Series of Information Retrieval, Springer, second ed., 2004.

[4] E. D. Liddy, "Document retrieval, automatic," *Encyclopedia of Language & Linguistics, 2nd Edition*, 2005.