Semantic Search Engine

Natural Language Processing
CS- 6320
(Dr. Mithun Balakrishna)

by

Shruti Harihar – sxh164530
Annesha Chowdhury – axc152830
Susmitha Manda - sxm161830

Contents:

Introduction

Semantic search is all about generating more relevant results and improving search accuracy by predicting the searcher's intent and the contextual sense of terms as they appear in the searchable dataspace. Semantic search systems use '*semantics' – the science of meaning in language* to retrieve relevant search results by considering various features like context of search, synonyms, generalized and specialized queries and natural language. Many major search engines like Google incorporate some of these elements of semantic search.

Problem Description

The goal of natural language processing is to allow a kind of interaction that non-programmers can obtain useful and relevant information from computing systems. Our challenge is to show that a computer can not only recognize a list of words but can also understand the content and retrieve relevant results that a user expects. To solve this challenge, we took the help of semantics which helps in answering the queries in a more relevant and insightful manner. In this project we implemented a semantic search application that produces improved results using NLP features and techniques. Our project implements a keyword-based strategy and an improved strategy using NLP feature and techniques.

Proposed Solution

In this project, we implemented a semantic search application that produces improved results using NLP functionalities and techniques. This project includes various techniques like tokenizing, index creation, query parsing and searching, evaluating the obtained search results and finally an attempt to improve shallow NLP pipeline results by using a combination of deeper NLP pipeline features. At the end of the project, we have tried a method to improve on the results of the deep NLP Pipeline by switching off and on some NLP features. Through this project, standard user-specific queries are processed token by token, with several included natural language features; the query text is matched with the indexed corpus which returns a list of the top relevant results related to the query. We have scored the results according to their relevance using the improved BM25F Algorithm. For indexing the corpus, we used Whoosh indexing.

Dataset Description

The data which have been considered for our semantic search engine has been taken from the NLTK Data website. Both datasets have been published by The Australian Broadcasting Commission in 2006 and consists of rural news and science news articles. The data is in pure text natural language format separated by spaces, commas and full stops.

The entire data has been divided into two categories:

a) **Rural News data:** It consists of 301271 words.
b) **Science News data:** It consists of 362693 words.

Implementation Details:

Programming Tools

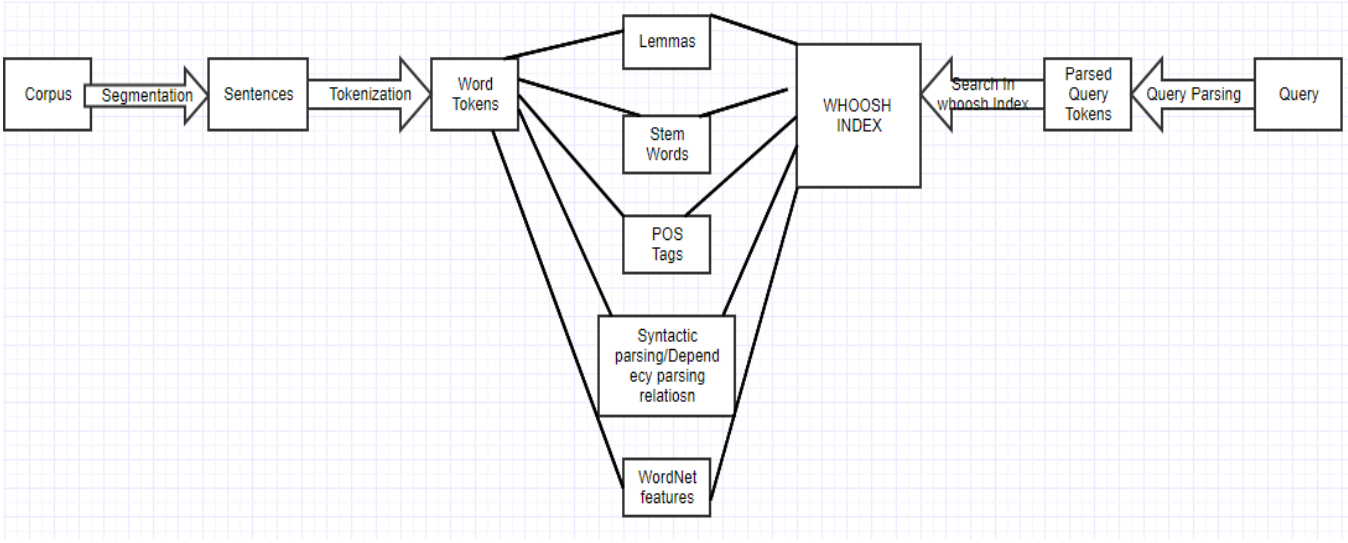
- **Version:** Python 3.6:2.
- **IDE:** PyCharm Community Edition3.
- **Indexing Tool:** *Whoosh 2.7.4*
- **Dependency Parser Tool:** Spacy

  Whoosh is a pure Python search engine library which runs without using a compiler. The indexed text which is created by Whoosh is in Unicode. In Whoosh, the Indexing of text, parsing of queries, implementation of natural language features as well as the scoring algorithms are all customizable and replaceable. In our project, we wrote our own customized analyzers for lemmatization and Part-Of-Speech tagging. To install Whoosh 2.7.4 on your machine, you need to have Python 3 working along with 'pip'.

Command to install Whoosh: pip install whoosh

- NLTK toolkit

Architectural Diagram



Results and error analysis

We implemented Whoosh Indexing tool in our project for indexing various features that are created from the corpus. Whoosh is a fast, Python search engine library. Indexing of text, parsing of search queries, level of information stored in each field, scoring algorithms, etc., are all customizable, replaceable and extensible. Whoosh helps in indexing free-form or structured text and then quickly find matching documents based on simple or complex criteria.

We performed various tasks which lead us to obtain the following results:

Task 1: Task 1 includes the creation of corpus which was taken from the NLTK website. The corpus was chosen in such a way that it consists of 1000 articles and 100,000 words to meet our requirements.

Task 2: This task implements shallow NLP pipeline by using keyword search index creation which includes segmenting the articles into sentences followed by words. A word vector is created and is indexed into search index with the help of Whoosh Indexing tool. The next step implemented is the query parsing and search technique that runs a search/match with the search query word vector against the sentence word vector.
When a query sentence is given, it is tokenized into words which are matched with the sentence word vector that is created the corpus and the following top 10 sentence matches are returned.

Task 3: In task 3 we implemented Deeper NLP pipeline to perform Semantic Search Index Creation that involves many features like Stemming, Lemmatization, POS Tagging, Syntactically Parsing a sentence etc., that are structured into separate search fields to form an index. This index is used for running queries to obtain top 10 query results.

Task 4: Improvement over Task 3's NLP Pipeline:
In this task, we have performed a trial and error method to improve the search results obtained in Task 3.
Features considered: WordNet Hypernyms, Hyponyms, Meronyms, Holonyms.
The WordNet features have been removed for the Query Parser in Task 4 and it was noted that we were getting better results for specific queries. Our logic behind this implementation is that when we generate the WordNet

features for the Query, the word (SYNSET) comes in all the query index fields which matches with all the corpus index fields belonging to WordNet features. Once we remove this feature from the Query Parser, we are getting better results.

Below are the different scenarios that are applied to the Tasks:

Scenario 1: Specific Queries

- When a specific query is parsed into the 3 tasks, an improvement was observed starting from task 2 till task 4.
- In task1, the results obtained were not much relevant and the exact match to the query was not found in the top 10 searches.
- When the same query is passed into task 3, the percentage of the relevant results increased and the exact match to the query was found in a lower rank.
- When passed into task 4, most relevant and an exact match result was found in the first place.

Example:

Query: Where is the most earth like planet?

Results from Task 2 – 10 Results**,** 5 relevant answers, no exact match result.

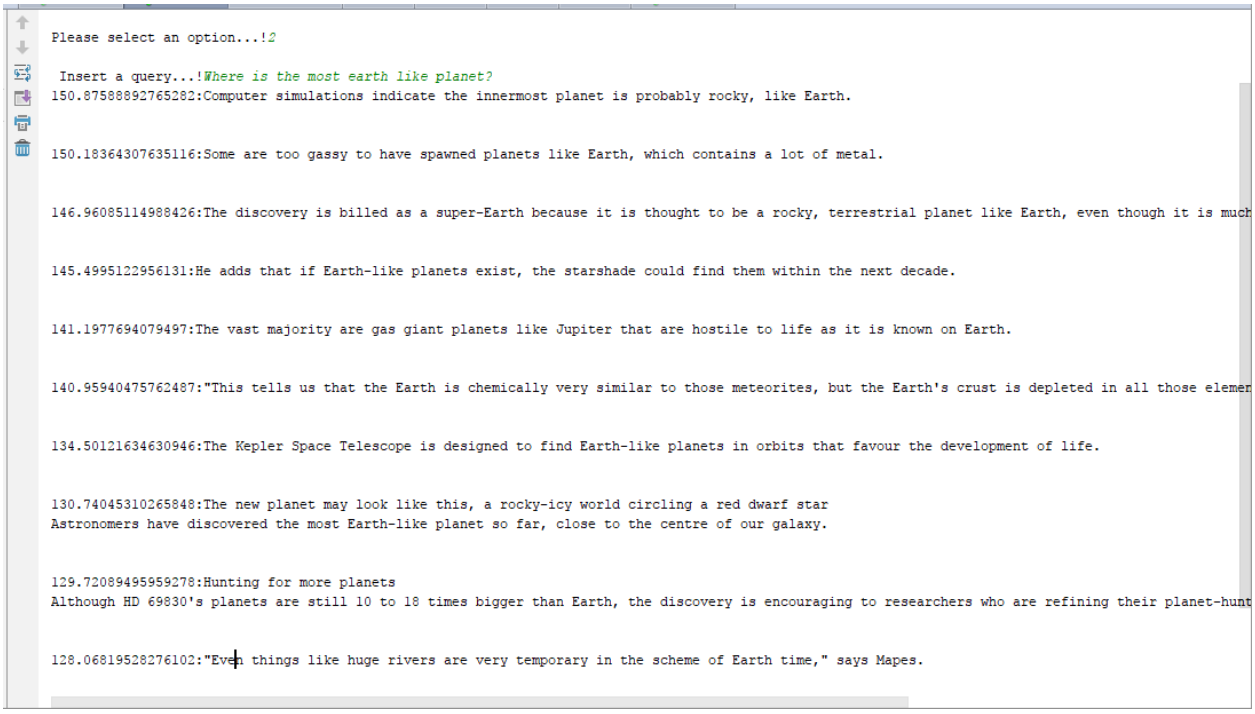
Results from Task 3 – 10 Results**,** 7 relevant answers, exact match result was at position 8 from the top.

Results from Task 4 – 10 Results**,** 7 relevant answers, exact match result was at position 1 from the top.

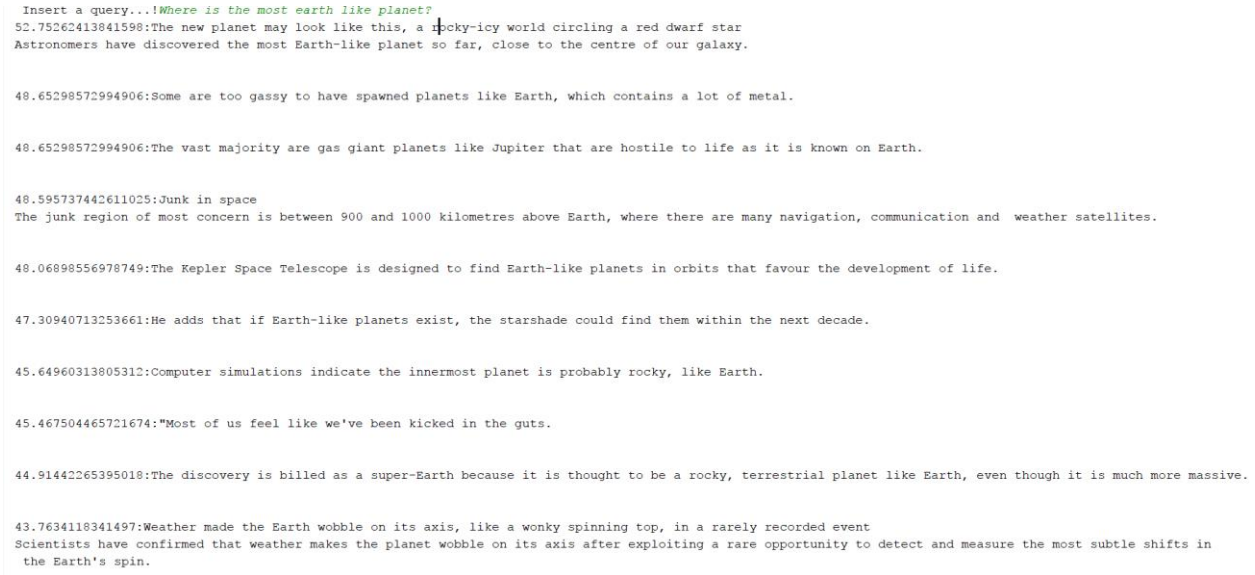
Task 2

```
 Insert a query...!Where is the most earth like planet?
18.849353535747916:The Kepler Space Telescope is designed to find Earth-like planets in orbits that favour the development of life.


18.44215947390885:The vast majority are gas giant planets like Jupiter that are hostile to life as it is known on Earth.


17.69026648373635:Hunting for more planets
Although HD 69830's planets are still 10 to 18 times bigger than Earth, the discovery is encouraging to researchers who are refining their planet-hunting
 techniques to find smaller, more Earth-like worlds.


17.584758060820224:He adds that if Earth-like planets exist, the starshade could find them within the next decade.


17.21107812840457:This is the cutest, most Disney-esque of the planets.


16.519872246304843:But if the shield blocks most of the starlight, planets may be easier to find
A huge daisy-shaped shield that would block out light from parent stars could be used to find Earth-like planets in other solar systems, a US astronomer says.


16.24685453591287:Some are too gassy to have spawned planets like Earth, which contains a lot of metal.


16.229335250491648:They include nearby stars of the right size, age and composition to have Earth-like planets circling them, scientists say.


15.615287474768145:Junk in space
The junk region of most concern is between 900 and 1000 kilometres above Earth, where there are many navigation, communication and  weather satellites.
```

Task 3



Task 4



<u>Scenario 2: Passing exact match words & modified match words</u>

- When an exact match query from the corpus is passed into task 2, the percentage of the relevant results were high, and an exact match was found in the first place. But when the same query is modified and passed, irrelevant results were obtained with no exact match in the top 10 results.
- However, when this case was applied in task 4, the percentage of relevant results were high in both the exact and modified cases and an exact match was also found in the top position.

Example:

Exact Query: What do you do when you pick up a pen and write?

- **Worked in Task 2 & Task 4** – High percentage of relevant results.

Modified Query: What am I doing when I am picking up my pen and writing?

- **Worked in Task 4** - High percentage of relevant results

- **Did not work properly in Task 2** – Irrelevant results.

Task 2 – not modified

```
 Insert a query...!What was blocked by floodwaters today?
21.591887787828863:Floodwaters close highway
A major highway between the Northern Territory and Western Australia remains blocked by floodwaters today.


15.7933061088243:What has blocked the pre-Big Bang view from theoreticians was the mathematical expression of what was happening - based on certain assumptions
 about space-time.


10.792486772849802:The Department of Natural Resources says it had warned farmer representative that fines would be imposed if access was blocked.


9.920655629084337:"Then you're looking at a situation where very important communications are being monitored, filtered and potentially blocked inappropriately
 by your email provider."


9.906585667209395:Climate change
Around 11,000 years ago what was the Arafura plain was flooded by rising seas as the ice age ended.


9.794930654254577:Floodwaters affect search for banana prawns
Floodwaters and muddy river sediments are making it hard for trawlers in the Gulf of Carpentaria this season.


9.725495748550113:"People decided to intervene in nature and supply their own food rather than relying on what was provided by the gods.


9.166027169783494:By extension, the researchers envisioned what insects would have looked like millions of years ago when the air was 35% oxygen.


8.979594674081858:Gulf graziers appeal for assistance as floodwaters rise
Graziers in Queensland's north-western Gulf country want more state help for families affected by the worst flooding in 100 years.
```

Task 4 – not modified

```
 Insert a query...!What was blocked by floodwaters today?
72.91195083660764:Floodwaters close highway
A major highway between the Northern Territory and Western Australia remains blocked by floodwaters today.


66.89466132809136:What has blocked the pre-Big Bang view from theoreticians was the mathematical expression of what was happening - based on certain assumptions
 about space-time.


47.38408705101491:The Department of Natural Resources says it had warned farmer representative that fines would be imposed if access was blocked.


46.75767974834878:Floodwaters affect search for banana prawns
Floodwaters and muddy river sediments are making it hard for trawlers in the Gulf of Carpentaria this season.


45.158560515966144:Flooding causes detour
Floodwaters in the Northern Territory are still blocking a major highway today.


39.15083210455302:"What we didn't realise at the time was the stock was blown over," he said.


38.40316268729046:But what surprised the researchers was evidence of salmonella bacteria.


37.81119066381495:"The Minister telephoned us and advised us that this is what the Government was going to do.


37.05094490472662:What she found was a merry-go-round of gathering and stealing.
```

Task 2 – modified

```
 Insert a query...!what is blocking the highway?
11.550843070556228:There is also speculation the Government is considering a $1 billion roads package aimed at Queensland's Bruce and Hume Highways.


10.958386310162096:Australia's highways below standard: survey
The first independent audit of the nation's highways shows half are below standard and may be unsafe to drive on.


9.864741070654295:"The question is what?


9.835921603469483:EFA is also worried that this Kazaa ruling could one day lead to email providers inappropriately monitoring, filtering or blocking messages.


9.78740009578329:Flooding causes detour
Floodwaters in the Northern Territory are still blocking a major highway today.


9.710746760910443:"It is very easy to see what is in the rocks, what is coming out of the rocks," Clark says.


9.47493539317122:The drug works by blocking the effects of the hormone progesterone, which a woman needs to start and maintain a pregnancy.


9.170188006237726:There's also a chance that immune cells could grow around the outside of the capsules, blocking the flow of insulin, she says.


9.144411198411035:What the songs mean is, again, unknown."


9.058729927130909:"The market in Sydney is just, 'What have you got, what have you got?
```

Task 4 – modified

```
 Insert a query...!what is blocking the highway?
34.72731352049564:Flooding causes detour
Floodwaters in the Northern Territory are still blocking a major highway today.


34.218461566143404:The drug works by blocking the effects of the hormone progesterone, which a woman needs to start and maintain a pregnancy.


31.192066433953165:EFA is also worried that this Kazaa ruling could one day lead to email providers inappropriately monitoring, filtering or blocking messages.


31.192066433953165:There's also a chance that immune cells could grow around the outside of the capsules, blocking the flow of insulin, she says.


31.192066433953165:Hanke says TGN1412 was designed to activate its target protein - rather than blocking it as many antibody drugs do.


26.429086032983527:"We know what's in the sea and what's in the land.


26.073481426712227:Ages later, about 65 million years ago, the Andes began to rise on the western edge of South America, blocking the river's passage to the
 Pacific and shifting its flow to the east.


25.431674402633984:"What we need to do is find what kind of animal toads really are, what bits of the landscape they use, how they use it and what they depend on."


25.34232075297152:"The market in Sydney is just, 'What have you got, what have you got?


25.34232075297152:We find that a rate of 5% is what is needed to explain what we see."
```

Scenario 3: After using NLP Pipelines on General Queries

- This scenario displays the difference between the results obtained before and after usage of NLP Pipelines.
- When no pipelines were used, the accuracy and the relevance of the results were quite low. When the pipelines were induced there was a drastic change in the accuracy and relevance of the results.

Example:

Query: What are the effects of Atkins diet?

Results from Task 2: 10 Results**,** 3 relevant results

Results from Task 4: 10 Results**,** 6 relevant results

Task 2

```
Insert a query...!what are the effects of atkins diet?
19.571774788705525:The Atkins diet builds on a long history of low-carbohydrate diets that reaches into the 19th century.


14.404661846535292:The Atkins diet stresses lashings of meat, butter and other dairy products - high-fat foods typically limited in classic diets - but cuts
 potatoes, rice and pasta to negligible levels and greatly limits intake of fruit and vegetables.


13.112747318088426:So, they are thought to mask some of the effects of global warming.


12.59074161217282:But the research shows low GI diets are better than high protein diets in reducing 'bad' or LDL (low density lipoprotein) cholesterol associated
  with cardiovascular disease.


11.87504775232052:The scientists are also monitoring seagrass beds in Thailand, home to shellfish, to see the effects of the 2004 tsunami.


11.859444616012464:The side-effects are the same as those of a spontaneous natural abortion, and include bleeding as part of the normal response."


11.644401383933612:Berk describes these effects as "the biology of hope", and says they are linked to the anticipation of a positive mood state.


11.574243768839214:"We are going to try to have an experimental demonstration of these effects.


11.418948844930739:"In spite of being part of the staple diets of these populations, their consumption is limited by the flatulence they produce."


11.34336595115331:The effects of the internet on society are still being debated, the researchers note in an article in the Journal of Computer-Mediated
 Communication.
```

Task 4

```
Insert a query...!what are the effects of atkins diet?
44.912501416497484:The Atkins diet builds on a long history of low-carbohydrate diets that reaches into the 19th century.

37.673380684753894:"What people do not understand is the potential side effects that diet mixed alcoholic drinks may have on their body's response to alcohol."

37.39441689280889:She reported losing 9 kilograms after eating only meat, cheese and salads, supplemented by minerals and vitamins sold by Atkins Nutritionals,
 the company founded by diet pioneer Robert Atkins in 1989.

36.98760532161257:"Whether or not the effects will be what's desired for the outcome, whether control would be better than eradication, I'd say eradication would
 sound the better idea.

34.94497033899204:Chen and team report seeing a 40-year-old obese woman a month after starting the Atkins diet.

34.94497033899204:Atkins Nutritionals emerged from bankruptcy protection earlier this year, specialising as a company that sells low-carb bars and shakes.

34.87564669312296:But geologists are debating what effect it has.

34.1922452115752:A Ministry of Defence report, which will be made public later this month, says what UFO watchers may be seeing unusual atmospheric effects like
 glowing plasma clouds.

30.747972737710224:The Atkins diet stresses lashings of meat, butter and other dairy products - high-fat foods typically limited in classic diets - but cuts
 potatoes, rice and pasta to negligible levels and greatly limits intake of fruit and vegetables.

30.097692167490298:"Our patient had an underlying ketosis caused by the Atkins diet and developed severe ketoacidosis," say the researchers, adding that mild
 pancreatitis or stomach infection may have contributed to the problem.
```

Scoring Metric(Evaluation):

Scoring Metric for the relevant queries:

- In Whoosh, there are four main algorithms to implement scoring for the relevant search results. They are TF-IDF, DFree, BM25, PL2 Algorithms.
- For this project, we have used the **BM25F algorithm** which is an extension of the baseline BM25 model used to rank documents in a search engine according to their relevance. It mainly uses bag-of-words retrieval method to retrieve a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document. In our project we have used an extension of the algorithm since BM25 leaves out the structure of documents in the weighting process.

How does BM25F work?

The BM25F involves weighting term frequencies according to their field importance, combining them, and then using the result pseudo-frequencies. It considers the breakdown of text in structured documents and then gives varying degrees of importance to different fields. It performs per field BM25 calculation but uses the shared document frequency across multiple fields. In our project, we notice that we get better BM25F scores in Task 3 rather than in Task 2. The reason for this is the deep NLP pipeline implemented in Task 3 which includes features like POS Tagging, Lemmatization and Stemming. We get better matches when we pass the Query to index which further leads to better scores and specific relevant results in Task 3.

Problems Encountered:

For the semantic search engine, Whoosh was chosen since it was a pure Python indexing tool unlike other indexing tools like SOLR. Since Whoosh is new, it was difficult to find good documentation for integration of NLP Features with the indexes. We referred to the open-source main Whoosh source code on Bitbucket and wrote our own Custom Analyzers and filters except the Stemming feature. That took a lot of research and trial-error on our part to implement it from the scratch.

Pending Issues

We had 2-3 specific queries which were returning generalized results for all three tasks. For example, if we give a Query specific to a single line in a article, it may return some general non-semantic results which are not related to the article directly but matches some words present in the parsed query.

Potential Improvements:

1. We can assign varying weights to the features implemented in the deep NLP pipeline. In that way, we can get an idea of which are the top-weighted NLP features to be used for the search engine.

2. For dealing with generalization: We are sending queries parsed with OR method now, in future we can try to improve the specific search results by using a combination of OR as well as AND methods for parsing the Queries.

References

[1**]** http://www.minerazzi.com/tutorials/bm25f-model-tutorial.pdf
[2] A Search Engine for Natural Language Applications: Michael J Carafella,Oren Etzioni
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.79.139&rep=rep1&type=pdf
[3] Whoosh Source Code: https://bitbucket.org/mchaput/whoosh/src