Semantic Search

Vaishnavi Bajirao Bhosale
Department of Computer Science, Fall 2018
The University of Texas at Dallas
Email: vbb160030@utdallas.edu

Abstract

Semantic search is a data searching technique that anticipates the intention behind the user's query and then find the content from the given corpus. The goal of semantic search is to provide the end user with the most relevant search engine results possible. This project aims at implementing a semantic search application with the help of an improved strategy using Natural Language Processing (NLP) features. Index is built on the corpus using SOLR for extracting the related features in such a way that retrieved result is relevant to the user's input query. In order to extract the important features NLTK, rake-nltk python libraries and WordNet interface are used. The improvement in results for semantic search is achieved by using a deeper NLP pipeline.

keywords: *semantic search, NLP features*

1. INTRODUCTION

Most existing information retrieval (IR) systems do not take much advantage of natural language processing (NLP) techniques due to the complexity of applying NLP to IR [8]. In this paper, I demonstrate a semantic search application that will produce an improved result using NLP features. Deeper NLP pipeline and query parsing with search are two important tasks performed in the application. The process of segmenting and tokenizing is done to get the sentences from articles and words from sentences. I have extracted lemmas, stemmed words, POS tags, syntactic phrases, hypernyms and hyponyms as features. All these features are indexed as separate fields in SOLR. To perform semantic Search, I added the data to Apache SOLR index (systematic arrangement of documents) using client API like python. For query processing, I segmented, tokenized and extracted the sentences, words and features of user input query. Pysolr is a lightweight python wrapper used as an interface that queries the server and returns result based on the query [5]. When a user runs a search in SOLR, the search query is processed by a request handler which calls a query parser which interprets the terms and parameters of a query and finally the query response is returned based on the index.

2. PREVIOUS WORK

Many semantic search engines are designed and implemented previously for different working environments. Bhagwat and Polyzotis propose a semantic-based file system search engine – Eureka, which uses an interference model to build the links between files and a FileRank metric to rank the files according their semantic importance [7]. R. Guha delivered a system that uses W3C's Resource Description Framework with the schema vocabulary provided by RDFS for describing resources and their inter-relations [6]. Yumao Lu investigated how semantic features can improve search relevance and that is the first study of this approach on a web scale [8]. He presents a set of features that incorporate semantic knowledge into the retrieval process and demonstrate that these carefully designed features significantly improve relevance; particularly for difficult queries [8].

3. CORPUS

For the purpose of this project, I am using the Reuters-21578 Corpus, which has news related articles of multiple categories such as food, economics and so on. The corpus is a collection of documents that appeared on Reuters newswire in 1987. The documents were assembled and indexed with categories. The corpus contains nearly 750 articles with 100,000 words.

4. APPROACH

The proposed system has the following steps:

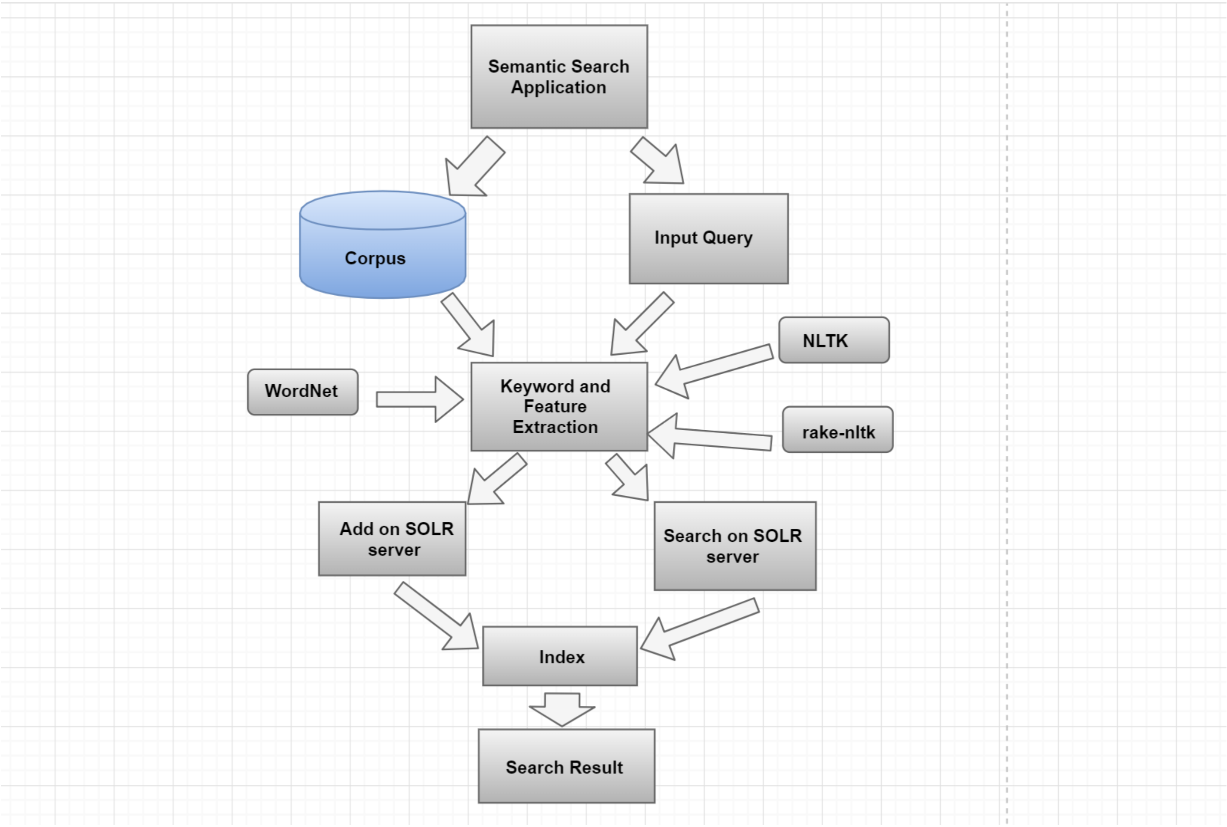
I. Read the file data from the corpus
II. Tokenize the words and extract the keywords and features using nltk, rake-nltk libraries and WordNet interface
III. Add all these features in SOLR server. SOLR creates the index by using these features as separate fields
IV. Read the input query provided by the user
V. Tokenize and extract keywords plus features from the query. Build the new query using these extracted features
VI. Search the query on SOLR server
VII. Using previously built-in index, SOLR will return the top 10 search results

Explanation

I. Initially, get all the data from the corpus and tokenize the data to get sentences using sent_tokenize from nltk library. For each sentence, tokenize it using word_tokenize to get the individual words. Index is a file name concatenated with a sentence count number.
II. Rapid Automatic Keyword Extraction (RAKE) algorithm, is a domain independent keyword extraction algorithm which tries to determine key phrases in a body of text by analyzing the frequency of word appearance and its co-occurrence with other words in the text [2]. Using RAKE, let's take into consideration only the top six keyword phrases.
III. For each token obtained from word_tokenize, get all the synsets using WordNet interface. Synset is a set of synonyms that share a common meaning [3]. For every synset, determine the hypernyms and hyponyms.
IV. Lemmatization and stemming both word normalization techniques are performed using nltk package to get the root forms of inflected words. WordNetLemmatizer and LancasterStemmer built-in methods are used to get the lemma and stem respectively.
V. A part-of-speech tagger is used to process a sequence of words and attaches a part of speech tag to each word.
VI. The python dictionary containing all the extracted keywords and features like lemma, stem, hypernym, hyponym, POS tags, top 6 phrases, word tokens, whole sentence and id with the index value is created for each sentence. And finally, the entire data containing all the dictionaries is added to SOLR.
VII. The data is being fed into a Solr index: a *document* containing multiple *fields*, each with a *name* and containing *content*, which may be empty [4]. One of the fields is designated as a unique ID field which in this case contains the file name with the sentence number.
VIII. Get the query from the user. Extract the keywords and features mentioned earlier.

IX. Create the query which contains all the extracted features in the required format.

X. The query is passed to built-in search method which will implicitly call the query parser that runs a query against the entire index and returns the top 10 search results.

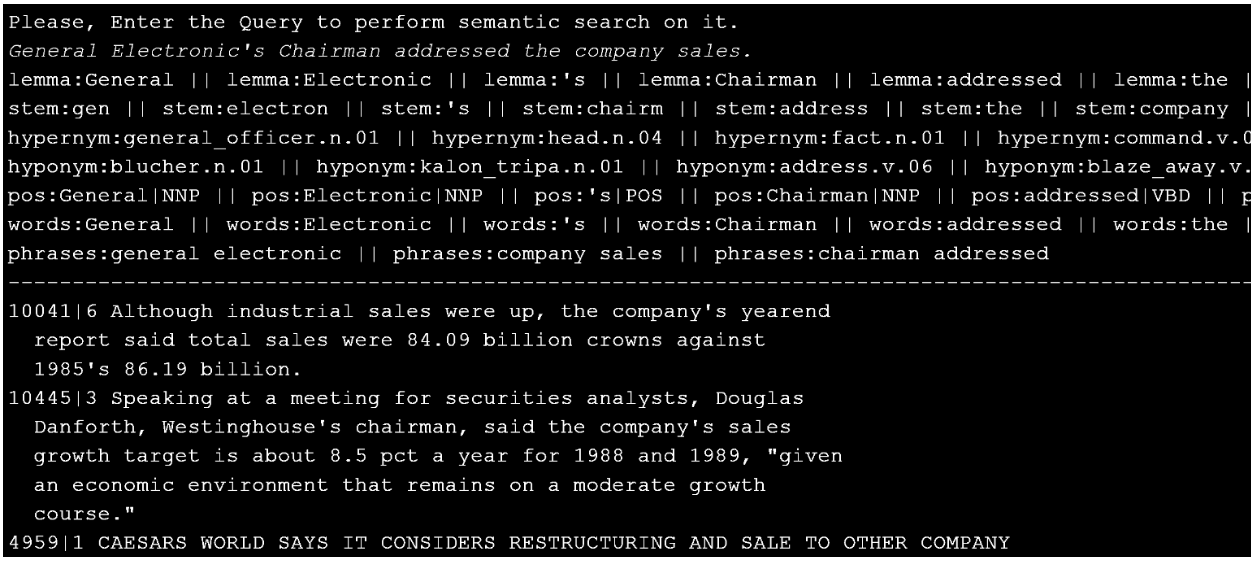


Architecture Diagram

5. EXPERIMENTS

There were numerous experiments performed when programming this project. The goal was to get the relevant data for user's input query. Various distinct NLP features were used to obtain the search result. Hypernyms and hyponyms are part of that list. Hypernymy can be defined as a relation from concepts to superordinate while hyponymy is a relation from concepts to subtypes. In addition to that WordNet interface returns all possible hypernyms and hyponyms but for a better result I considered only the first one as it has the highest possibility and gives the best result. Following screenshots show the difference between the search result if hypernyms and hyponyms are used.

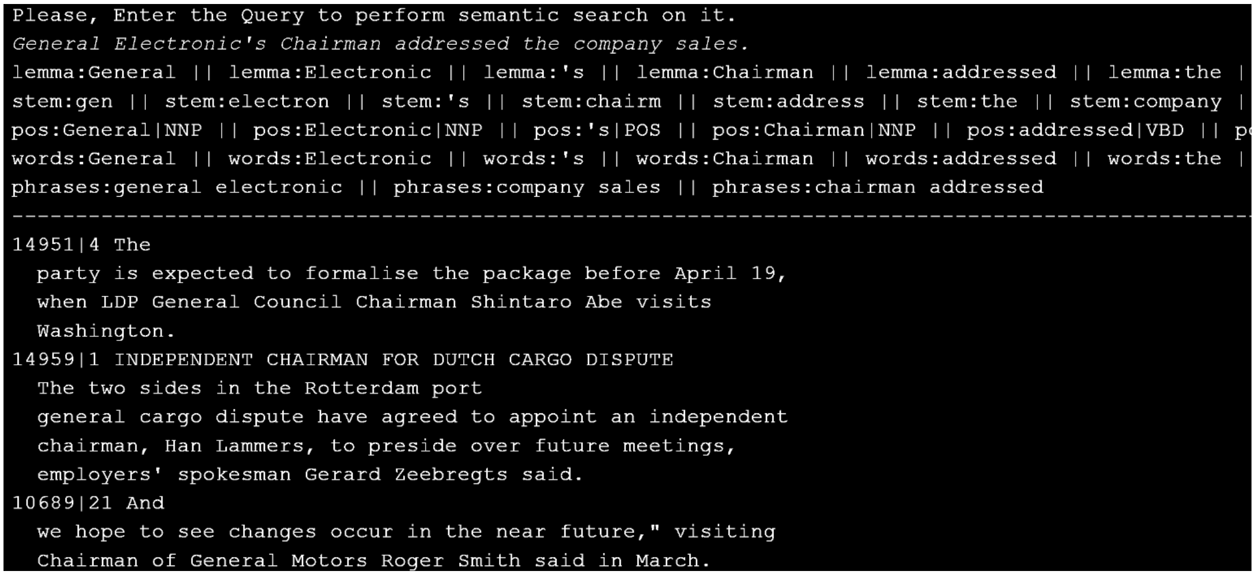
Query: General electronics chairman addressed the company sales.



```
Please, Enter the Query to perform semantic search on it.
General Electronic's Chairman addressed the company sales.
lemma:General || lemma:Electronic || lemma:'s || lemma:Chairman || lemma:addressed || lemma:the |
stem:gen || stem:electron || stem:'s || stem:chairm || stem:address || stem:the || stem:company |
hypernym:general_officer.n.01 || hypernym:head.n.04 || hypernym:fact.n.01 || hypernym:command.v.0
hyponym:blucher.n.01 || hyponym:kalon_tripa.n.01 || hyponym:address.v.06 || hyponym:blaze_away.v.
pos:General|NNP || pos:Electronic|NNP || pos:'s|POS || pos:Chairman|NNP || pos:addressed|VBD || p
words:General || words:Electronic || words:'s || words:Chairman || words:addressed || words:the |
phrases:general electronic || phrases:company sales || phrases:chairman addressed
-------------------------------------------------------------------------------------------------
10041|6 Although industrial sales were up, the company's yearend
  report said total sales were 84.09 billion crowns against
  1985's 86.19 billion.
10445|3 Speaking at a meeting for securities analysts, Douglas
  Danforth, Westinghouse's chairman, said the company's sales
  growth target is about 8.5 pct a year for 1988 and 1989, "given
  an economic environment that remains on a moderate growth
  course."
4959|1 CAESARS WORLD SAYS IT CONSIDERS RESTRUCTURING AND SALE TO OTHER COMPANY
```

Search result with Hypernyms and Hyponyms features used

As per the user's query, the one with using hypernyms and hyponyms contains information about the company sales and what is said by the chairman in the search result while this relevant information is missing in the second case.



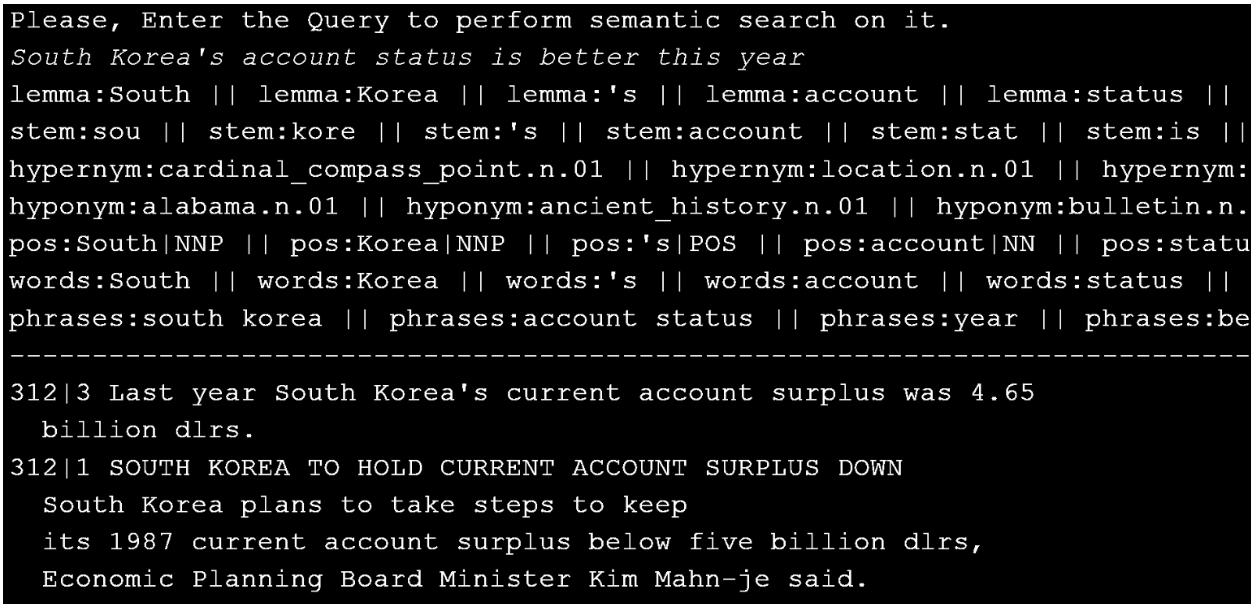
```
Please, Enter the Query to perform semantic search on it.
General Electronic's Chairman addressed the company sales.
lemma:General || lemma:Electronic || lemma:'s || lemma:Chairman || lemma:addressed || lemma:the |
stem:gen || stem:electron || stem:'s || stem:chairm || stem:address || stem:the || stem:company |
pos:General|NNP || pos:Electronic|NNP || pos:'s|POS || pos:Chairman|NNP || pos:addressed|VBD || po
words:General || words:Electronic || words:'s || words:Chairman || words:addressed || words:the |
phrases:general electronic || phrases:company sales || phrases:chairman addressed
-------------------------------------------------------------------------------------------------
14951|4 The
  party is expected to formalise the package before April 19,
  when LDP General Council Chairman Shintaro Abe visits
  Washington.
14959|1 INDEPENDENT CHAIRMAN FOR DUTCH CARGO DISPUTE
  The two sides in the Rotterdam port
  general cargo dispute have agreed to appoint an independent
  chairman, Han Lammers, to preside over future meetings,
  employers' spokesman Gerard Zeebregts said.
10689|21 And
  we hope to see changes occur in the near future," visiting
  Chairman of General Motors Roger Smith said in March.
```

Search result without Hypernyms and Hyponyms features used

Stemming and Lemmatization are also a crucial part of the NLP features which are necessary of the better search result. Accurate POS tagging to queries is very important because POS tagging accuracy directly affects the search effectiveness. Following screenshots show how the information we get can be irrelevant if we miss these features.

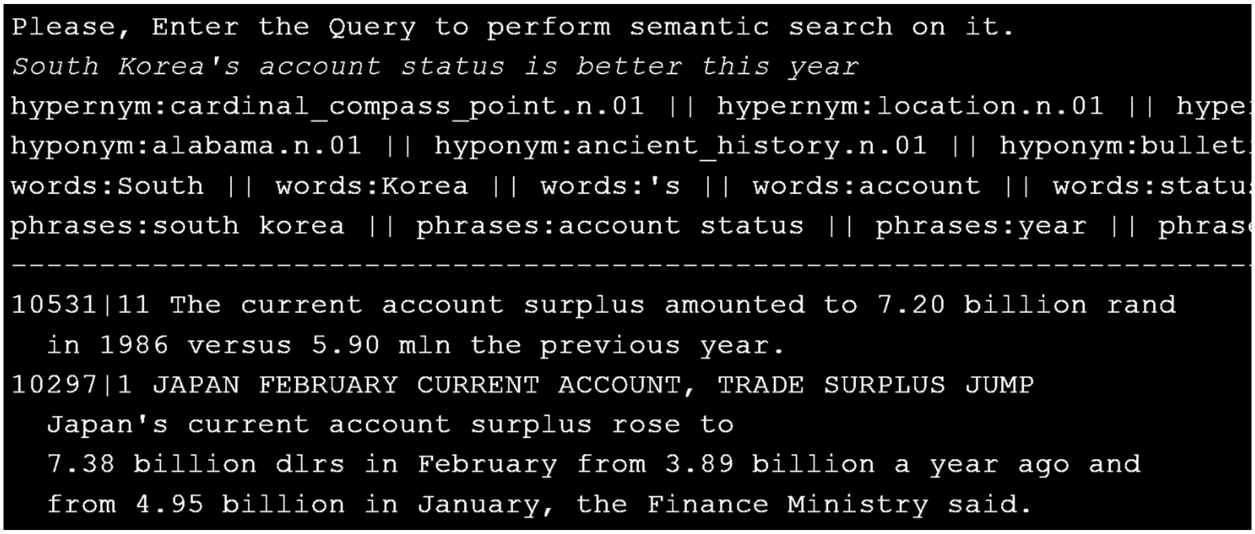
Query : South Korea’s account is better this year.

`



```
Please, Enter the Query to perform semantic search on it.
South Korea's account status is better this year
lemma:South || lemma:Korea || lemma:'s || lemma:account || lemma:status ||
stem:sou || stem:kore || stem:'s || stem:account || stem:stat || stem:is ||
hypernym:cardinal_compass_point.n.01 || hypernym:location.n.01 || hypernym:
hyponym:alabama.n.01 || hyponym:ancient_history.n.01 || hyponym:bulletin.n.
pos:South|NNP || pos:Korea|NNP || pos:'s|POS || pos:account|NN || pos:statu
words:South || words:Korea || words:'s || words:account || words:status ||
phrases:south korea || phrases:account status || phrases:year || phrases:be
---------------------------------------------------------------------------
312|3 Last year South Korea's current account surplus was 4.65
  billion dlrs.
312|1 SOUTH KOREA TO HOLD CURRENT ACCOUNT SURPLUS DOWN
  South Korea plans to take steps to keep
  its 1987 current account surplus below five billion dlrs,
  Economic Planning Board Minister Kim Mahn-je said.
```

Search result by using stem, lemma and POS tags



```
Please, Enter the Query to perform semantic search on it.
South Korea's account status is better this year
hypernym:cardinal_compass_point.n.01 || hypernym:location.n.01 || hype
hyponym:alabama.n.01 || hyponym:ancient_history.n.01 || hyponym:bullet
words:South || words:Korea || words:'s || words:account || words:statu
phrases:south korea || phrases:account status || phrases:year || phras
----------------------------------------------------------------------
10531|11 The current account surplus amounted to 7.20 billion rand
  in 1986 versus 5.90 mln the previous year.
10297|1 JAPAN FEBRUARY CURRENT ACCOUNT, TRADE SURPLUS JUMP
  Japan's current account surplus rose to
  7.38 billion dlrs in February from 3.89 billion a year ago and
  from 4.95 billion in January, the Finance Ministry said.
```

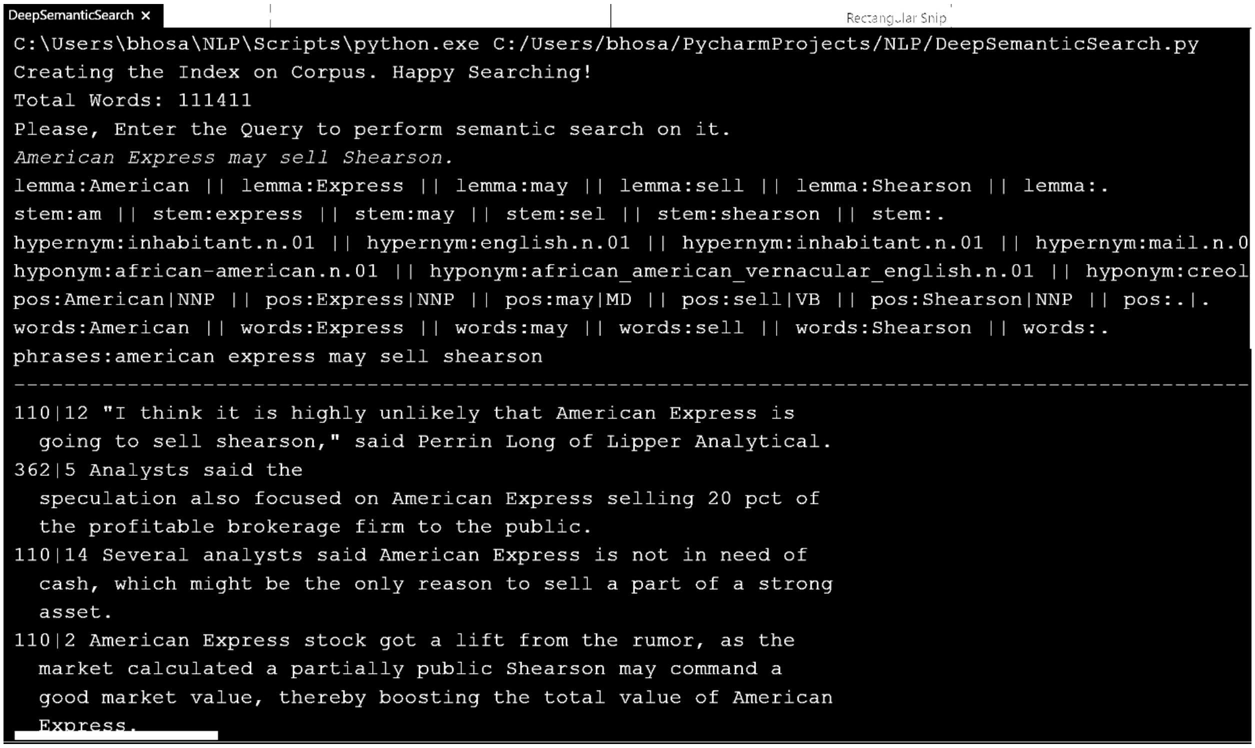
Search result without using stem, lemma and POS tags

6. RESULT

According to the user’s query, the most relevant information from the corpus with the help of indexing is found and the system prints the top 10 search result along with all the extracted features of the input query.

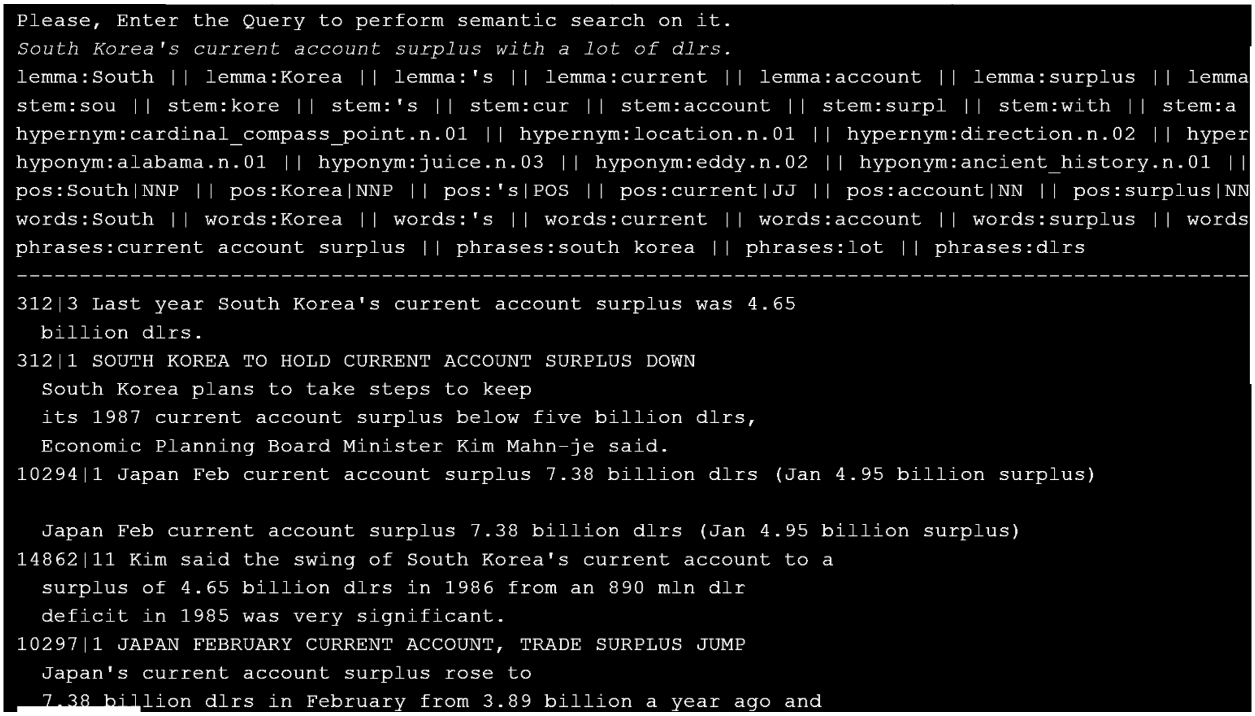
Some screenshots for example run of the system are as follows:

I. Query: American Express may sell Shearson.



```
C:\Users\bhosa\NLP\Scripts\python.exe C:/Users/bhosa/PycharmProjects/NLP/DeepSemanticSearch.py
Creating the Index on Corpus. Happy Searching!
Total Words: 111411
Please, Enter the Query to perform semantic search on it.
American Express may sell Shearson.
lemma:American || lemma:Express || lemma:may || lemma:sell || lemma:Shearson || lemma:.
stem:am || stem:express || stem:may || stem:sel || stem:shearson || stem:.
hypernym:inhabitant.n.01 || hypernym:english.n.01 || hypernym:inhabitant.n.01 || hypernym:mail.n.0
hyponym:african-american.n.01 || hyponym:african_american_vernacular_english.n.01 || hyponym:creol
pos:American|NNP || pos:Express|NNP || pos:may|MD || pos:sell|VB || pos:Shearson|NNP || pos:.|.
words:American || words:Express || words:may || words:sell || words:Shearson || words:.
phrases:american express may sell shearson
----------------------------------------------------------------------------------------------------
110|12 "I think it is highly unlikely that American Express is
  going to sell shearson," said Perrin Long of Lipper Analytical.
362|5 Analysts said the
  speculation also focused on American Express selling 20 pct of
  the profitable brokerage firm to the public.
110|14 Several analysts said American Express is not in need of
  cash, which might be the only reason to sell a part of a strong
  asset.
110|2 American Express stock got a lift from the rumor, as the
  market calculated a partially public Shearson may command a
  good market value, thereby boosting the total value of American
  Express.
```

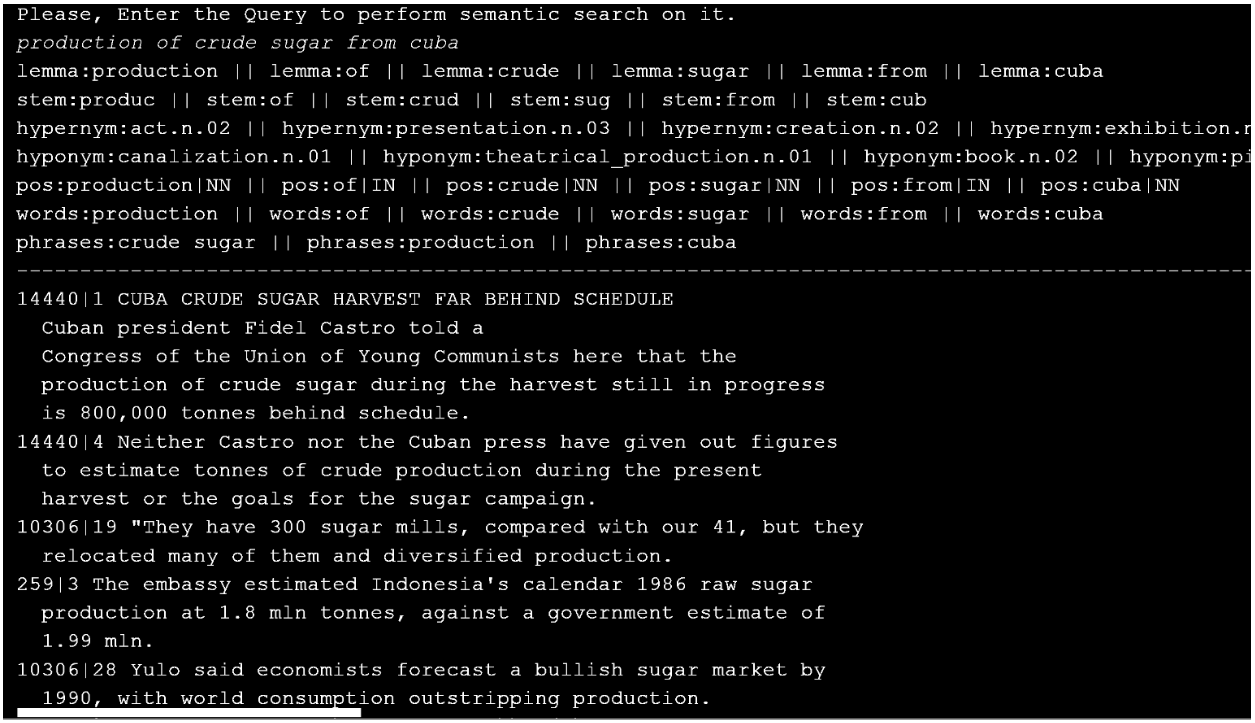
II. Query: South Korea’s current account surplus with a lot of dlrs.



```
Please, Enter the Query to perform semantic search on it.
South Korea's current account surplus with a lot of dlrs.
lemma:South || lemma:Korea || lemma:'s || lemma:current || lemma:account || lemma:surplus || lemma
stem:sou || stem:kore || stem:'s || stem:cur || stem:account || stem:surpl || stem:with || stem:a
hypernym:cardinal_compass_point.n.01 || hypernym:location.n.01 || hypernym:direction.n.02 || hyper
hyponym:alabama.n.01 || hyponym:juice.n.03 || hyponym:eddy.n.02 || hyponym:ancient_history.n.01 ||
pos:South|NNP || pos:Korea|NNP || pos:'s|POS || pos:current|JJ || pos:account|NN || pos:surplus|NN
words:South || words:Korea || words:'s || words:current || words:account || words:surplus || words
phrases:current account surplus || phrases:south korea || phrases:lot || phrases:dlrs
----------------------------------------------------------------------------------------------------
312|3 Last year South Korea's current account surplus was 4.65
  billion dlrs.
312|1 SOUTH KOREA TO HOLD CURRENT ACCOUNT SURPLUS DOWN
  South Korea plans to take steps to keep
  its 1987 current account surplus below five billion dlrs,
  Economic Planning Board Minister Kim Mahn-je said.
10294|1 Japan Feb current account surplus 7.38 billion dlrs (Jan 4.95 billion surplus)

  Japan Feb current account surplus 7.38 billion dlrs (Jan 4.95 billion surplus)
14862|11 Kim said the swing of South Korea's current account to a
  surplus of 4.65 billion dlrs in 1986 from an 890 mln dlr
  deficit in 1985 was very significant.
10297|1 JAPAN FEBRUARY CURRENT ACCOUNT, TRADE SURPLUS JUMP
  Japan's current account surplus rose to
  7.38 billion dlrs in February from 3.89 billion a year ago and
```

III. Query: production of crude sugar from cuba



```
Please, Enter the Query to perform semantic search on it.
production of crude sugar from cuba
lemma:production || lemma:of || lemma:crude || lemma:sugar || lemma:from || lemma:cuba
stem:produc || stem:of || stem:crud || stem:sug || stem:from || stem:cub
hypernym:act.n.02 || hypernym:presentation.n.03 || hypernym:creation.n.02 || hypernym:exhibition.r
hyponym:canalization.n.01 || hyponym:theatrical_production.n.01 || hyponym:book.n.02 || hyponym:pi
pos:production|NN || pos:of|IN || pos:crude|NN || pos:sugar|NN || pos:from|IN || pos:cuba|NN
words:production || words:of || words:crude || words:sugar || words:from || words:cuba
phrases:crude sugar || phrases:production || phrases:cuba
--------------------------------------------------------------------------------------------------
14440|1 CUBA CRUDE SUGAR HARVEST FAR BEHIND SCHEDULE
  Cuban president Fidel Castro told a
  Congress of the Union of Young Communists here that the
  production of crude sugar during the harvest still in progress
  is 800,000 tonnes behind schedule.
14440|4 Neither Castro nor the Cuban press have given out figures
  to estimate tonnes of crude production during the present
  harvest or the goals for the sugar campaign.
10306|19 "They have 300 sugar mills, compared with our 41, but they
  relocated many of them and diversified production.
259|3 The embassy estimated Indonesia's calendar 1986 raw sugar
  production at 1.8 mln tonnes, against a government estimate of
  1.99 mln.
10306|28 Yulo said economists forecast a bullish sugar market by
  1990, with world consumption outstripping production.
```

7. CONCLUSION

Thus, we can conclude that the corpus to the user's input query is where the system gets the most relevant information. Semantic search in Natural Language Processing has some amazing real-world applications. Moreover, the use of a deeper NLP pipeline can help the system in a big way as made evident by our experimental results. All in all, searcher's intent was found to be the most important aspect.

The search result from SOLR can be improved by boosting different features like POS tags, lemma or any other features. The weights can be applied to the most relevant features, the ones which helped getting better result. By analyzing the result in trial and error basis or using some algorithm, we can decide the appropriate weights for the relevant features. Some advanced features like headwords, meronyms, holonyms or dependency parse relations can be extracted to get the more relevant result. In this case, I have considered only top six phrases obtained by rake-nltk. However, including few more or all the phrases might have improved the results. WSD (Word Sense Disambiguation) can be implemented to improve the search result.

8. REFERENCES

[1] http://www.nltk.org/nltk_data/
[2] https://pypi.org/project/rake-nltk/
[3] http://www.nltk.org/howto/wordnet.html
[4] http://lucene.apache.org/solr/
[5] https://github.com/django-haystack/pysolr
[6] R. Guha, R. McCool, and E. Miller, "Semantic search," in *Proceedings of WWW '03*, Budapest, 2003

[7] H. Dong, F. K. Hussain, and E. Chang, “A survey in semantic search technologies,” in 2nd IEEE International Conference on Digital Ecosystems and Technologies. IEEE, 26-29 February 2008, pp. 403–408.

[8] Yumao Lu, Fuchun Peng, Gilad Mishne, Xing Wei, and Benoit Dumoulin. Improving web search relevance with semantic features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP ’09, pages 648–657, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.