NLP FALL-2017 FINAL PROJECT REPORT

Team Name: WEBSCRAPERS

Team Members:

1. Teja Mukka: txm162230
2. Ravikiran Kolanpaka: rxk171530
3. Rohit Sindhu: rks160030

Problem Description:

The problem that we are trying to access in this project is divided into 4 parts:

1. Choose a good corpus with at least:
    a. 1000 articles
    b. 100000 words
2. Article-Sentence based indexing:
    a. Create a NLP Pipeline to create keyword search index
        i. Segment the News articles into sentence.
        ii. Tokenize the sentences into words.
        iii. Index the word vector per sentence into search index.
    b. Natural language query parsing and search:
        i. Run a search/match with the search query word vector against the sentence word vector (present in the Lucene/SOLR search index.
3. Semantic search index creation.
    a. Segment the News articles into sentences.
        i. Tokenize the sentences into words.
        ii. Lemmatize the words to extract lemmas as features.
        iii. Stem the words to extract stemmed words as features.
        iv. Part-of-speech (POS) tag the words to extract POS tag features.
        v. Syntactically parse the sentence and extract phrases, head words, OR dependency parse relations as features.
        vi. Using WordNet, extract hypernymns, hyponyms, meronyms, AND holonyms as features.
    b. Index the various NLP features as separate search fields in a search index:
        i. Run a search/match against the separate or combination of search index fields created from the corpus.
4. Improve the shallow NLP pipeline results using a combination of deeper NLP pipeline features:
    a. Use different combination of weights of the features to see if the results improve.

Proposed Solution:

1. We decided to solve the problem with a combination of Wordnet based tools such as NLTK, Stanford Dependency Parser for natural language processing and Apache Solr to do customized search in the corpus.
2. We used the corpus file **rural.txt** which was downloaded from www.nltk.org/nltk_data/ [Australian Broadcasting Commission 2006]
3. Used NLTK to tokenise each of the sentences.
4. Used Apache Solr tool to maintain a server instance and create indexes.
5. We tokenized the sentences and created JSON structure.
6. Sent the JSON data to Solr to create and save indexes.
7. We invoke Solr Server through pysolr based python program.
8. We pass the desired query into the python program.
9. The query is sent to Solr with 3 levels of search:
    a. Plain sentence (tokens) based search.
    b. Tokens along with the following features to search.
        i. POS tags
        ii. Stem
        iii. Lemma
        iv. Hypernyms
        v. Hyponyms
        vi. Holonyms
        vii. Meronyms
        viii. Synonyms
        ix. Head Words
    c. Tokens and features with selected weights on them. Some features could be set off completely.
10. Solr installation details are present in the README.txt file included in the package.

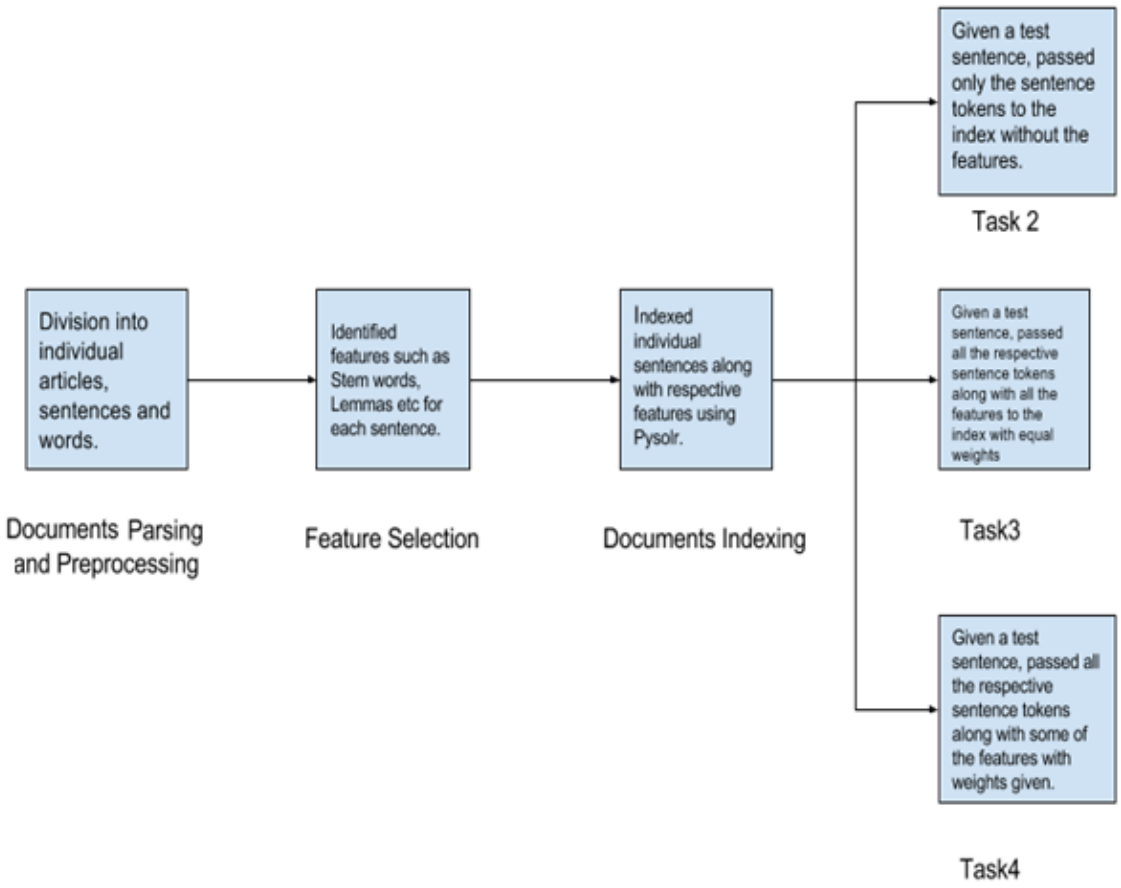
Implementation Details:

We used the following software tools for our implementation:

1. Python 3.5
2. Apache Solr version 7.1.0
3. Nltk-3.2.5
    a. Wordnet
    b. Stopwords
    c. WordNetLemmatizer
    d. PorterStemmer
    e. StanfordDependencyParser
4. Pysolr
5. PrettyTable

We have used a corpus with total 2424 Articles and 1690625 words in it.

Architectural Diagram:



Code / Files and their Role:

1. **Project_Part3.py**: This is a python library which returns feature vector for each of the word in a sentence.
2. **Indexer.py**: This python program creates index in the Solr server for each of the line in the corpus.
3. **project_main.py**: Project main file that takes a query input and returns results for part 2,3,4 separately.
4. **get_corpus_size.py**: This is a small python program that indicates the size of corpus used.
5. **Rural.txt**: The main corpus file with 2424 Articles and over 1.5 million words.
6. **prettytable.py**: A third party python library to print the feature vector in a tabular manner.

Experiments and Results:

We ran our tests on multiple different queries and found interesting results.

Overall in most of the cases results from part 2 became worse in part 3 where we are using all the features without any specific weights for search.

However, results improved in part 4 where we provided different weights to the features.

Few features in part have null values for some of the words, we have a default text for those NULL Statements as '*fetchedNULLvalue*'.

Some of our queries and their results are as following:

SAMPLE QUERY 1: "CAN CANCER BE TREATED BY SNAIL"

1. Top 10 Output of part2:

length of the second results: 3033

- Text ID is '619_5 '. The giant African snail is also a major risk , and clothing , and footwear is being searched for its eggs , which can be carried in soil material
- Text ID is '1197_3'. `` The vessels were treated in open water by a contractor with AFMA , '' he said
- Text ID is '1037_0'. Laboratory tests of an extract from olive leaves have shown it can kill prostate and breast cancer cells
- Text ID is '1101_0'. New research has shown eating the sprouts that are often found in mixed salads can help curb cancer
- Text ID is '1512_2'. The Federal Government had treated the payments as income , and demanded tax be paid
- Text ID is '1700_1'. More than 2,500 megalitres of treated sewage water will be diverted from the Derwent River
- Text ID is '750_1 '. The study by the Queensland Cancer Fund and the University of Sydney found twice as many people as previously thought faced cancer-causing substances at work , including pesticides , UV rays and dust
- Text ID is '1618_5'. `` Also we can coat the CCA treated pine logs and make posts for vineyards and that sort of thing with the recycled poly
- Text ID is '748_1 '. The study by the Queensland Cancer Fund and the University of Sydney has found twice as many people as previously thought are exposed to cancer-causing substances at work , including pesticides , UV rays and dust
- Text ID is '1465_1'. It is the second year in a row that a snail from a packing shed in the Sunraysia region has been found by US quarantine officials

2. Feature List from Part 3:

Feature list for the third task is as following:

Feature Names	can	cancer	be	treated	by	snail
POS Tags	MD	NN	VB	VBN	IN	NN
Lemma	can	cancer	be	treated	by	snail
Stem	can	cancer	be	treat	by	snail
Synonym	buttocks	Cancer	beryllium	treat	aside	escargot
Hypernym	container	malignant_tumor	metallic_element	interact	fetchedNULLvalue	gastropod
Hyponym	beer_can	carcinoma	fetchedNULLvalue	brutalize	fetchedNULLvalue	edible_snail
Meronym	fetchedNULLvalue	fetchedNULLvalue	fetchedNULLvalue	fetchedNULLvalue	fetchedNULLvalue	fetchedNULLvalue
Holonym	fetchedNULLvalue	fetchedNULLvalue	fetchedNULLvalue	fetchedNULLvalue	fetchedNULLvalue	fetchedNULLvalue
HeadWord	treated	--	--	--	--	--

3. Top 10 Output of part 3 [*please notice that results are bad compared to part 2 output*]:

length of the third task results: 3

- Text ID is '1197_3'. `` The vessels were treated in open water by a contractor with AFMA , '' he said
- Text ID is '1512_2'. The Federal Government had treated the payments as income , and demanded tax be paid
- Text ID is '1246_3'. The season in this part of the world has been treated well , with some saying it is one of the best they have seen

4. Weights Selected for part 4:
    a. **Stem^4 AND Lemma^4 AND Synonyms^4**

5. Top 10 Output of part 4 with weighted query [*please notice that results are better compared to part 2 output*]:

length of the fourth task results: 3017

- Text ID is '10_0 '. New research is under way to investigate whether South Australian sea snails could eventually be used to treat cancer
- Text ID is '619_5 '. The giant African snail is also a major risk , and clothing , and footwear is being searched for its eggs , which can be carried in soil material
- Text ID is '1197_3'. `` The vessels were treated in open water by a contractor with AFMA , '' he said
- Text ID is '1037_0'. Laboratory tests of an extract from olive leaves have shown it can kill prostate and breast cancer cells
- Text ID is '1101_0'. New research has shown eating the sprouts that are often found in mixed salads can help curb cancer
- Text ID is '1512_2'. The Federal Government had treated the payments as income , and demanded tax be paid
- Text ID is '750_1 '. The study by the Queensland Cancer Fund and the University of Sydney found twice as many people as previously thought faced cancer-causing substances at work , including pesticides , UV rays and dust
- Text ID is '1700_1'. More than 2,500 megalitres of treated sewage water will be diverted from the Derwent River
- Text ID is '1618_5'. `` Also we can coat the CCA treated pine logs and make posts for vineyards and that sort of thing with the recycled poly
- Text ID is '748_1 '. The study by the Queensland Cancer Fund and the University of Sydney has found twice as many people as previously thought are exposed to cancer-causing substances at work , including pesticides , UV rays and dust

SAMPLE QUERY 2: "DOES DOCTOR FACE RACISM IN RURAL AUSTRALIA"

1. Top 10 Output of part 2:

length of the second results: 5646

- Text ID is '132_0 '. Mining and resource companies are starting to face staff shortages as the jobs boom in rural Australia continues
- Text ID is '67_0 '. Overseas trained doctors are being warned against practising in rural Australia , because they could face racist abuse from both patients and colleagues
- Text ID is '1735_1'. The former head of atmospheric research at the CSIRO , Dr Graeme Pearman , says Australia will face growing international pressure to meet targets and it does not have any in place
- Text ID is '67_7 '. But he says while getting to know the locals took some time , stories of racism in country areas are exaggerated
- Text ID is '2140_0'. The Rural Doctors Association (RDA) claims the Federal Government is more concerned with ensuring people have a phone rather than a doctor or hospital
- Text ID is '1014_0'. If you are living in remote Australia in the future , instant messages with information from bushfire alerts to doctor visits could pop up on your television
- Text ID is '2185_1'. The state of mind in rural Australia been revealed in the latest Rabobank Rural Confidence Survey
- Text ID is '514_3 '. Mr Crombie , who lives in Brisbane , says he is not a typical farmer , but does not think that stops him being an effective leader of Australia 's most powerful rural lobby group
- Text ID is '2383_1'. The group of six men and four women have been sent to Port Augusta in South Australia to face deportation
- Text ID is '1742_0'. Giant hardware chain Bunnings could face prosecution for allegedly bringing grapevines into South Australia from a phylloxera risk area in Victoria

2. Feature list from part 3:

Feature list for the third task is as following:

Feature Names	does	doctor	face	racism	in	rural	Australia
POS Tags	VBZ	VB	NN	NN	IN	JJ	NNP
Lemma	doe	doctor	face	racism	in	rural	Australia
Stem	doe	doctor	face	racism	in	rural	australia
Synonym	Department_of_Energy	Doctor_of_the_Church	expression	racism	inch	rural	Australia
Hypernym	executive_department	medical_practitioner	external_body_part	bias	linear_unit	fetchedNULLvalue	fetchedNULLvalue
Hyponym	fetchedNULLvalue	abortionist	countenance	anti-Semitism	fetchedNULLvalue	fetchedNULLvalue	fetchedNULLvalue
Meronym	Department_of_Energy_Intelligence	fetchedNULLvalue	beard	fetchedNULLvalue	em	fetchedNULLvalue	Australian_Alps
Holonym	fetchedNULLvalue	fetchedNULLvalue	head	fetchedNULLvalue	foot	fetchedNULLvalue	Australia
HeadWord	does	--	--	--	--	--	--

3. Top 10 Output of Part 3 [*please notice that results are bad compared to part 2 output*]:

length of the third task results: 8

- Text ID is '975_5 '. `` In addition , with interest rates around the world rising , does suggest it will be a softer year
- Text ID is '467_4 '. `` All this does is add to our manufacturing base a very important strategic asset , not only for our business but for the farmers of Australia as well , '' he said
- Text ID is '47_52 '. If I look at New Zealand , part of the horticulture industry in New Zealand does use what I call cheaper labour
- Text ID is '914_2 '. With grape growers struggling with poor prices and an oversupply , Mr Brownhill does not , however , expect much improvement in the short term
- Text ID is '1250_72'. He does concede , though , that there is a bit of a way to go to convince consumers about eating the food that 's made from g-m crops
- Text ID is '1306_4'. `` And if there is , does it have enough funding to get it going and if there is n't why is there one ? `` Because it 's one of the best returns on money that they can get
- Text ID is '1486_4'. `` What the plant does is because it lives in very arid regions it needs to get its water from the ground water so it puts the most of its energy into creating this huge single root that penetrates very , very deeply into the regolith , into the cover of the landscape to the ground water , '' he said
- Text ID is '1116_7'. `` What it does is it undermines the strength of existing brands and it means that the temptations for those brands to try and meet similar price points is there because they obviously want to be able to sell their wine

4. Weights Selected for part 4:
    a. **Stem^4 AND Lemma^4 AND Synonyms^4**
5. Top 10 Output of part 4 with weighted query [*please notice that results are better compared to part 2 output*]:

length of the fourth task results: 5632

- Text ID is '67_0 '. Overseas trained doctors are being warned against practising in rural Australia , because they could face racist abuse from both patients and colleagues
- Text ID is '132_0 '. Mining and resource companies are starting to face staff shortages as the jobs boom in rural Australia continues
- Text ID is '143_5 '. `` The impact is felt more significantly in rural areas because traditionally rural doctors have worked longer hours on average than metropolitan doctors
- Text ID is '200_0 '. Rural doctors say new funding for mental health services must take into account the lack of psychologists in regional Australia
- Text ID is '1735_1'. The former head of atmospheric research at the CSIRO , Dr Graeme Pearman , says Australia will face growing international pressure to meet targets and it does not have any in place
- Text ID is '2140_0'. The Rural Doctors Association (RDA) claims the Federal Government is more concerned with ensuring people have a phone rather than a doctor or hospital
- Text ID is '2402_0'. While rural Australia continues to cry out for doctors , there 's now alarm over the falling number of doctors willing to work full time
- Text ID is '67_7 '. But he says while getting to know the locals took some time , stories of racism in country areas are exaggerated
- Text ID is '1014_0'. If you are living in remote Australia in the future , instant messages with information from bushfire alerts to doctor visits could pop up on your television
- Text ID is '923_0 '. Rural doctors in Tasmania claim they are being devalued by the introduction of nurse practitioners in remote areas

SAMPLE QUERY 3: "DOES GRAIN PRICE AFFECT DUE TO INQUIRY ON AWB"

1. Top 10 Output of part 2:

length of the second results: 7840

- Text ID is '2401_17'. The suspension does not prevent Australian wheat from being sold in the US or affect AWB 's ability to trade on commodity exchanges
- Text ID is '839_2 '. ABARE 'S study says the world oil price , feed grain prices and tax arrangements all affect the industry
- Text ID is '1433_0'. Wheat exporter AWB has told Victorian grain growers it does not think the Cole inquiry into allegations of sanction-breaking kickbacks will present any problems for the company
- Text ID is '3_0 '. But an analyst predicts grain prices will drop another $ 20 a tonne on the back of the inquiry into AWB
- Text ID is '1433_6'. `` AWB spokesman Peter McBride says the company does not want to make any public statement pre-empting the outcome of the Cole inquiry
- Text ID is '580_0 '. Despite the oil-for-food inquiry and a sliding share price , wheat handler AWB has delivered a strong half-year result
- Text ID is '529_0 '. President George W Bush does not appear to think there will be need for a US inquiry into AWB , despite pressure from domestic farm groups and some senators
- Text ID is '2_0 '. AWB still has plenty of support among grain growers in central western New South Wales despite the revelations of the Cole inquiry
- Text ID is '1134_1'. This time a challenge has been initiated by Commissioner Terence Cole , who does not want to participate in the case about AWB 's claim to withhold documents from his inquiry
- Text ID is '250_0 '. Prime Minister John Howard has told the Cole inquiry he does not believe he received or read any cables that warned AWB was involved in sanctions-busting

2. Feature List from Part 3:

Feature list for the third task is as following:

Feature Names	does	grain	price	affect	due	to	inquiry	on	AWB
POS Tags	VBZ	VB	NN	NN	JJ	TO	VB	IN	NNP
Lemma	doe	grain	price	affect	due	to	inquiry	on	AWB
Stem	doe	grain	price	affect	due	to	inquiri	on	awb
Synonym	Department_of_Energy	texture	monetary_value	involve	ascribable	to	question	along	AWB
Hypernym	executive_department	atom	value	feeling	right	fetchedNULLvalue	problem_solving	fetchedNULLvalue	fetchedNULLvalue
Hyponym	fetchedNULLvalue	granule	assessment	fetchedNULLvalue	fetchedNULLvalue	fetchedNULLvalue	empirical_research	fetchedNULLvalue	fetchedNULLvalue
Meronym	Department_of_Energy_Intelligence	fetchedNULLvalue	fetchedNULLvalue	fetchedNULLvalue	fetchedNULLvalue	fetchedNULLvalue	fetchedNULLvalue	fetchedNULLvalue	fetchedNULLvalue
Holonym	fetchedNULLvalue	fetchedNULLvalue	fetchedNULLvalue	fetchedNULLvalue	fetchedNULLvalue	fetchedNULLvalue	fetchedNULLvalue	fetchedNULLvalue	fetchedNULLvalue
HeadWord	affect	--	--	--	--	--	--	--	--

1. Top 10 Output of Part 3 [*please notice that results are bad compared to part 2 output*]:

length of the third task results: 3

- Text ID is '168_0 '. High fuel and fertiliser costs , a low Australian dollar and global grain prices will affect plantings for this season 's grains crops
- Text ID is '47_67 '. This will affect everything from urbanisation , migration into major cities , to unemployment
- Text ID is '218_4 '. `` It will certainly affect saleyard numbers and obviously the movement of cattle wherever

1. Weights Selected for part 4:
    a. **Stem^4 AND Lemma^4 AND Synonyms^4**

3. Top 10 Output of part4 with weighted query [*please notice that results are better compared to part 2 output*]:

length of the fourth task results: 7804

- Text ID is '3_0 '. But an analyst predicts grain prices will drop another $ 20 a tonne on the back of the inquiry into AWB
- Text ID is '2401_17'. The suspension does not prevent Australian wheat from being sold in the US or affect AWB 's ability to trade on commodity exchanges
- Text ID is '1433_0'. Wheat exporter AWB has told Victorian grain growers it does not think the Cole inquiry into allegations of sanction-breaking kickbacks will present any problems for the company
- Text ID is '839_2 '. ABARE 'S study says the world oil price , feed grain prices and tax arrangements all affect the industry

- Text ID is '1433_6'. `` AWB spokesman Peter McBride says the company does not want to make any public statement pre-empting the outcome of the Cole inquiry
- Text ID is '250_0 '. Prime Minister John Howard has told the Cole inquiry he does not believe he received or read any cables that warned AWB was involved in sanctions-busting
- Text ID is '2406_0'. The dairy industry is worried that a high level of wheat exports could affect local feed grain prices
- Text ID is '168_0 '. High fuel and fertiliser costs , a low Australian dollar and global grain prices will affect plantings for this season 's grains crops
- Text ID is '1402_2'. She was also involved in negotiating $ 750 million a year worth of grain handling and storage contracts , and in co-ordinating AWB 's response to the UN 's Volcker inquiry
- Text ID is '2031_0'. The number of grain-fed cattle in Australia fell by nearly 60,000 head in the last quarter , due to skyrocketing grain prices and the worsening drought

Problems faced during implementation:

- Cleaning the corpus file which had a lot of Unicode characters.
- Head Word extraction from a sentence was a major hurdle as Stanford Dependency parser is very slow and heavy application.
- Solr setup and indexing was tough. Had to learn Solr from scratch.
- Indexing takes a total of 5-6 hours because of the slow dependency parser.
- Coming up with the queries for which result change for good and bad in internal comparison of tasks 2, 3 & 4 was another challenge that we faced.
- Extracting and dealing with features that are null for many the words was another problem.
- Stanford Dependency parser return Articles, Determiners and other un-useful words as the sentence headword which worsens the final result.

Pending Issues:

- For the features - Hypernym, Holonym, Meronym etc, the top Synset has been used. Other Synsets could also be explored.
- Instead of using one synonym, we can also use other synonyms for a given word.
- We can also use other dependency parsers to perfectly extract the head word of a given sentence.
- Parse and Dependency trees can also be added as a feature to the index to get better results.

Potential Improvements:

- Custom similarity functions can be created to further improve the results of the retrieval.
- A Term Frequency, Frequency matrix can be used to retrieve the best results.
- Experiments such as stop words removal, Punctuations removal can be done to further improve the query results.
- Removal of duplicate words in the whole document might help in the retrieval for which experiment needs to be done.