NATURAL LANGUAGE PROCESSING – TEXT ANALYTICS

Robert Frost Poems

NOVEMBER 6, 2022

SUBHASHINI NATARAJAN

OBJECTIVE

The objective of the project is to perform Natural Language Processing, text analytics on selected poems by Robert Frost in pdf. The analysis involves in processing the unstructured, text heavy data to help understand the most used words, the sentiments reflected, and topics used by the poet.

BACKGROUND/HISTORY

Natural language processing helps business processes in efficiently processing unstructured data. NLP uses techniques syntax and semantic analysis for text analytics. Syntax technique is used to assess the meaning from a language based on grammatical rules. This project includes in performing text analytics on selected poems by Robert Frost.

DATA

The data for the project has been obtained from poem hunter website. The link for the file is given below.

https://www.poemhunter.com/robert-frost/ebooks/?ebook=0&filename=robert_frost_2004_9.pdf

The link has some of the selected poems of Robert Frost.

METHODS & PROCESSES

1. **Text Preprocessing:**

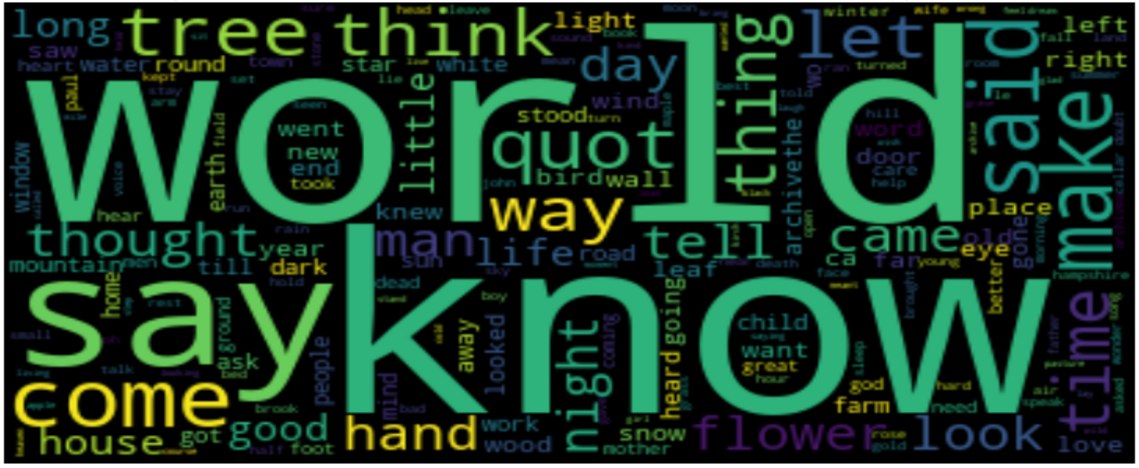
The pdf file is read into a variable, and all the lines in the pdf text are extracted using textract method. The number of pages in the document is 308.

All the blank spaces are removed. Auto correct method is used to correct typos and all the text in the data is converted into lower case. All the punctuations and special characters are removed from the text. The words are lemmatized using wordnetlemmatizer method. The cleansed data is stored in a data frame.

2. Word cloud with most frequent words:

In the cleansed data, the stop words are excluded. Using countvectorizer method, the number of occurrences of tokens are calculated and built a sparse matrix of documents. In total, 5303 unique words excluding blank spaces are in the document. The top 100 words are extracted into a dictionary. The words, 'Robert', 'Frost' and 'poetry' are excluded from the dictionary. Using wordcloud method the top hundred words are displayed as a word cloud. The size of the word is based on the frequency of the word, larger the size, more frequent the word is. The word cloud is as given below.

Top most common 100 words from Robert Frost poems



3. Sentiment Analysis:

Sentiment analysis identifies the polarity of a given text. There are different flavors of sentiment analysis, but one of the most widely used techniques labels data into positive, negative, and neutral.

VADER is a rule-based feeling analysis technique. VADER utilizes a mix of words that are, for the most part, marked by their semantic direction as one or the other positive or negative. Thus, VADER not only talks about the Polarity score yet, in addition, it tells us concerning how positive or negative a conclusion is. Textblob method is applied on cleansed data and the sentiment results are observed to be,

```
Sentiment (polarity=0.06836453311688115,
   subjectivity=0.45382644761381175)
```

The polarity of 0.06 which means that the document is **neutral,** and **0.45** subjectivity refers self-opinion in the document rather than public opinions, and beliefs.

4. Text Summarization:

Text summarization includes in extracting a concise summary from a lengthy text. The introduction mentioned in preface, is three page long and the text is summarized using spaCy statistical models - **en_core_web_sm.** The three-page, document is summarized as below –

```
         'Frosts poems are critiqued in the Anthology of Modern Ameri
can Poetry, Oxford University Press, where it is mentioned that behi
nd a sometimes charmingly familiar and rural façade, Frosts poetry f
requently presents pessimistic and menacing undertones which often a
re not recognized nor analyzed. grandfather Frost had, shortly befor
e his death, purchased a farm for the young couple in Derry, New Ham
```

```
pshire; and Robert worked the farm for nine years, while writing ear
ly in the mornings and producing many of the poems that would later b
ecome famous.'
```

The above is a perfect summarization of the three-page text.

5. Topic Modeling:

Topic modeling is a type of statistical modeling to discover the abstract 'topics' that are presented in a set of documents or a single document. A topic is a collection of prevalent keywords that are typical representatives. It’s through keywords in which one determines the topic.

Latent Dirichlet Allocation(LDA) is a popular algorithm for topic modeling with excellent implementations in the Python’s gensimpackage. LDA is used to classify text in our document to a particular topic. LDA model is applied on the data.

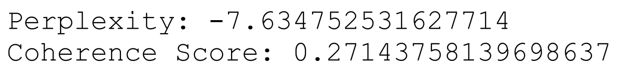
The top ten keywords, include,

```
  Know, go, see, come, say, make, world, thing, look, leave
```

The topic model is evaluated by computing the perplexity and topic coherence.

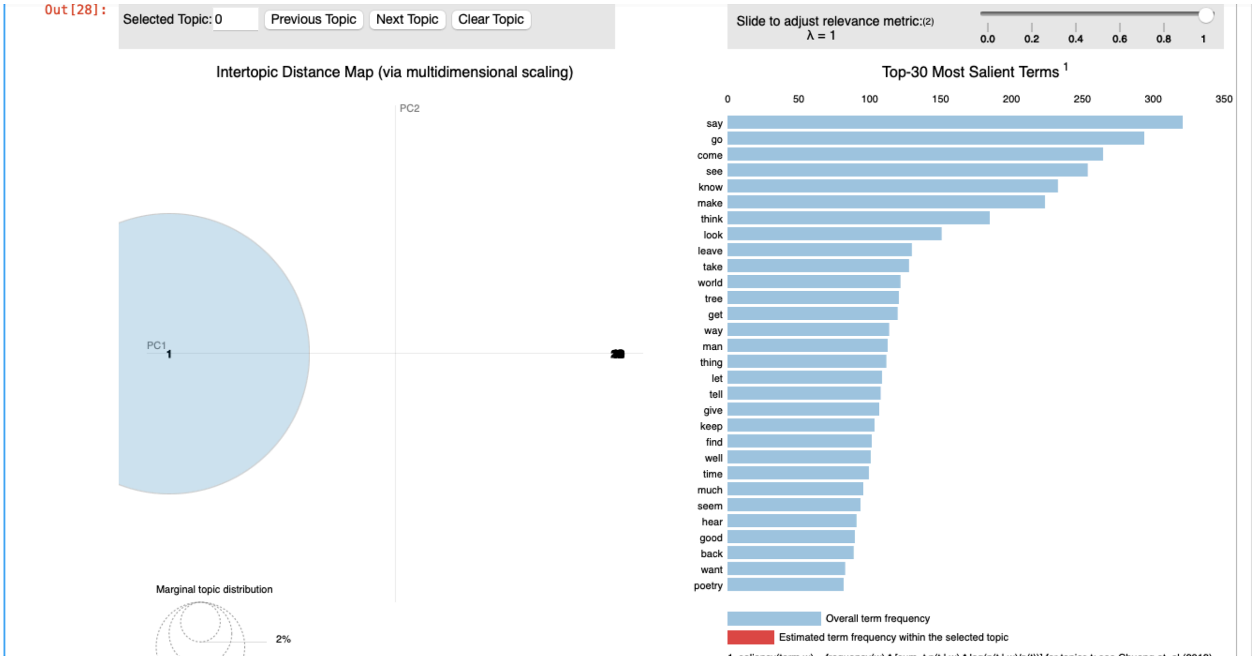
Perplexity is an evaluation metric on how probable (predictive likelihood) new unseen data is given the model that was learned earlier.

Topic **coherence** measures a score on a single topic through measuring the degree of semantic similarity of all the high scoring words in a topic. Both model perplexity and topic score provide a convenient measure to judge how good a given topic model is.



```
Perplexity: -7.634752531627714
Coherence Score: 0.27143758139698637
```

Topic modeling is visualized using pyLDAvis –


Out[28]:
Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
Marginal topic distribution
2%
Top-30 Most Salient Terms 1
0
50
100
150
200
250
300
350
say
go
come
see
know
make
think
look
leave
take
world
tree
get
way
man
thing
let
tell
give
keep
find
well
time
much
seem
hear
good
back
want
poetry
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

CONCLUSION

The document is successfully analyzed, by following data cleansing. The top 100 words in the data are identified and presented as word cloud and “World” is identified as the most frequent word. For sentiment analysis, Frost’s poems are identified to be neutral. The summarization produced a concise summary of the preface document. Topic modelling was successful, in which the document is focused on world, followed by tree, excluding the verbs.

CHALLENGES

I initially attempted to prepare the text summarization for the entire list of poems assuming there can be a common theme. The summarization didn’t produce the meaningful result despite refining the code. However, the same code could meaningfully summarize the text data in the preface.

LIMITATIONS

The verbs in the text like ‘say’, ‘go’ are not removed from the data set and hence the topic modeling includes the verbs as well.

ETHICAL CONSIDERATIONS

The subject of analysis in the project is to understand the literature that has no societal context but only about nature. It is also not used for any decision making or make societal judgment. In general NLP is a branch that is formed at the intersection of artificial intelligence, computational linguistics, and computer science. It aims to give computers the ability to understand texts and spoken words in a similar way to that of human beings. NLP analysis can face historical bias, representation bias, ambiguity, errors in text and speech, usage of a specific slang. The analysis should carefully evaluate for such biases in the analysis.

FUTURE USES/ADDITIONAL APPLICATIONS

The analysis performed can be applied to the companies for understanding business documents like legal documents, contract agreements. In the field of education, can be applied for understanding research papers.

RECOMMENDATIONS

The text contains verbs like ‘say’ and it will be further useful to remove the verbs and analysis only on the objects that the poet refers to.

IMPLEMENTATION PLAN

The machine learning models are deployed and scheduled to production in batch mode or online mode. This being a text analytics can be run one time or on demand basis.

REFERENCES

https://www.analyticsvidhya.com/blog/2020/03/spacy-tutorial-learn-natural-language-processing/

https://towardsdatascience.com/a-beginners-guide-to-sentiment-analysis-in-python-95e354ea84f6

https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24