Problem Statement

Topic Modelling Project (Natural Language Processing)

Introduction:

Topic modelling is recognizing the words from the topics present in the document or the corpus of data. This is useful because extracting the words from a document takes more time and is much more complex than extracting them from topics present in the document.
The aim of this project was to identify the relationships between the entities across the given json files which contain the news articles in it and come up with a solution to provide the named entity recognition.

About this Data

In the real live project, the xml files were taken as source. But for now, the xml files been converted to JSON files. A total of 2000 JSON files been provided within the dataset folder.

Description

Domain: Natural Language Processing

Problem Statement:

Topic modelling is recognizing the words from the topics present in the document or the corpus of data. This is useful because extracting the words from a document takes more time and is much more complex than extracting them from topics present in the document. For example, there are 1000 documents and 500 words in each document. So to process this it requires 500*1000 = 500000 threads. So when you divide the document containing certain topics then if there are 5 topics present in it, the processing is just 5*500 words = 2500 threads.

This looks simple than processing the entire document and this is how topic modelling has come up to solve the problem and also visualizing things better.

Some of the important points or topics which makes text processing easier in NLP:

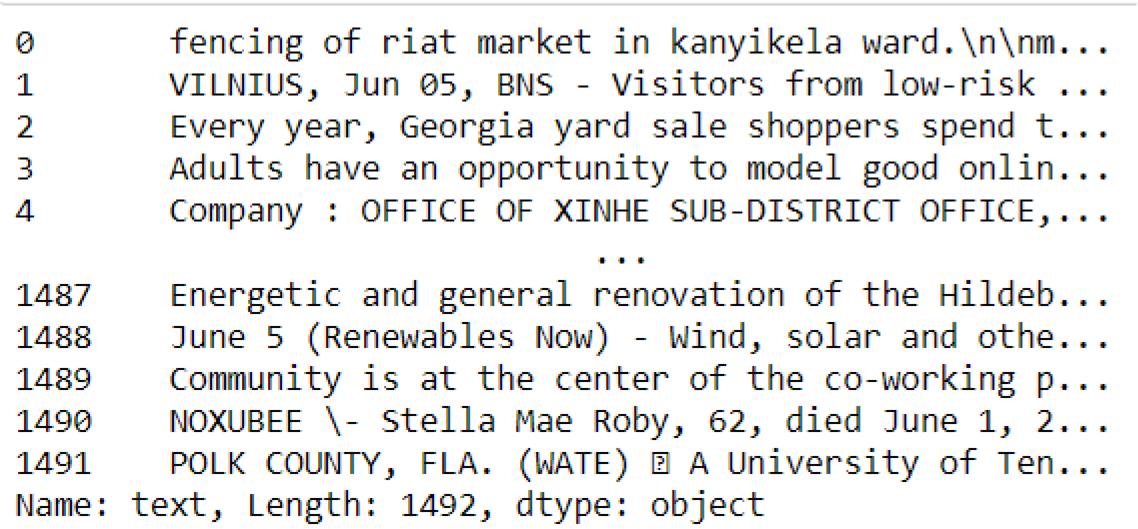
- Removing stopwords and punctuation marks
- Stemming
- Lemmatization
- Encoding them to ML language using Countvectorizer or Tfidf vectorizer

For this project we will pre-process the data and explore various models that help us accomplish this tasks

Data Type: We have provided you a set of Json Files.

Task to Do:

1. Import the required libraries and read the Set of Json Files from folder provided and extract the text from the files. We recommend that you save the data into a csv file for further processing. Here is a glimpse of the data.



```
0       fencing of riat market in kanyikela ward.\n\nm...
1       VILNIUS, Jun 05, BNS - Visitors from low-risk ...
2       Every year, Georgia yard sale shoppers spend t...
3       Adults have an opportunity to model good onlin...
4       Company : OFFICE OF XINHE SUB-DISTRICT OFFICE,...
                                ...
1487    Energetic and general renovation of the Hildeb...
1488    June 5 (Renewables Now) - Wind, solar and othe...
1489    Community is at the center of the co-working p...
1490    NOXUBEE \- Stella Mae Roby, 62, died June 1, 2...
1491    POLK COUNTY, FLA. (WATE) ⍰ A University of Ten...
Name: text, Length: 1492, dtype: object
```

2. Data pre-processing - Removing stopwords and punctuation marks.

   We remove stopwords like the, is, you, me etc and punctuation marks. As these will effect the model performance.

3. Data pre-processing - Stemming.

   When Stemming is applied to the words in the corpus the word gives the base for that particular word. It is like from a tree with branches you are removing the branches till their stem. Eg: fix, fixing, fixed gives fix when stemming is applied. There are different types through which Stemming can be performed. Some of the popular ones which are being used are:
   1. Porter Stemmer
   2. Lancaster Stemmer
   3. Snowball Stemmer

4. Data pre-processing – Lemmatization

   Lemmatization also does the same task as Stemming which brings a shorter word or base word. There is a slight difference between them is Lemmatization cuts the word to gets its lemma word meaning it gets a much more meaningful form than what stemming does. The output we get after Lemmatization is called ‘lemma’.

   When stemming is used it might give ‘hav’ cutting its affixes whereas lemmatization gives ‘have’. There are many methods through which lemma can get obtained and lemmatization can be performed. Some of them are WordNet Lemmatization, TextBlob, Spacy, Tree Tagger, Pattern, Genism, and Stanford CoreNLP lemmatization. Lemmatization can be applied from the mentioned libraries.

5. Data Pre-Processing – Tokenization

   Tokenization breaks the raw text into words, sentences called tokens. These tokens help in understanding the context or developing the model for the NLP. If the text is split into words using some separation technique it is called word tokenization and same separation done for sentences is called sentence tokenization

6. We can now save the pre-processed data as a separate file and use it for model building.

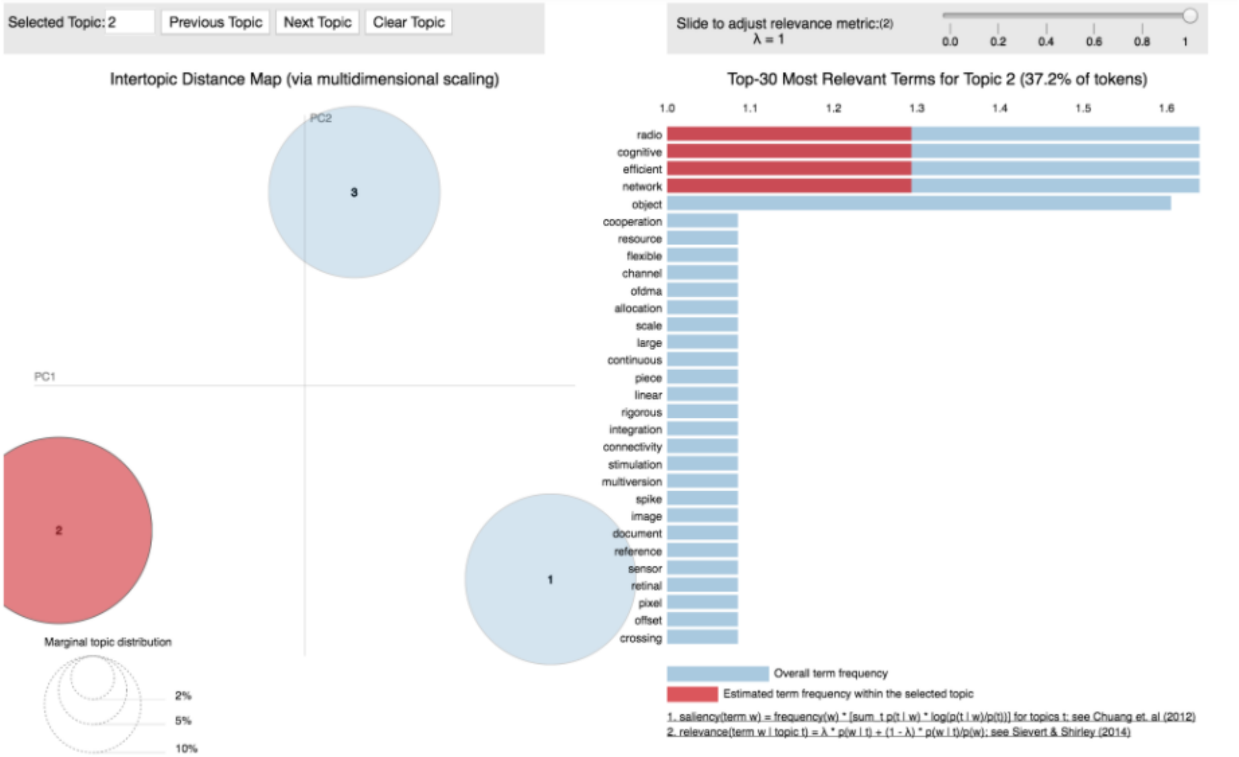
7. Model Building –

   We can use Various models now to accomplish our required task.

   Eg: Topic modelling is done using LDA(Latent Dirichlet Allocation). Topic modelling refers to the task of identifying topics that best describes a set of documents. These topics will only emerge during the topic modelling process (therefore called latent). And one popular topic modelling technique is known as Latent Dirichlet Allocation (LDA).

8. Model Evaluation –

   Visualize the topics and their importance.

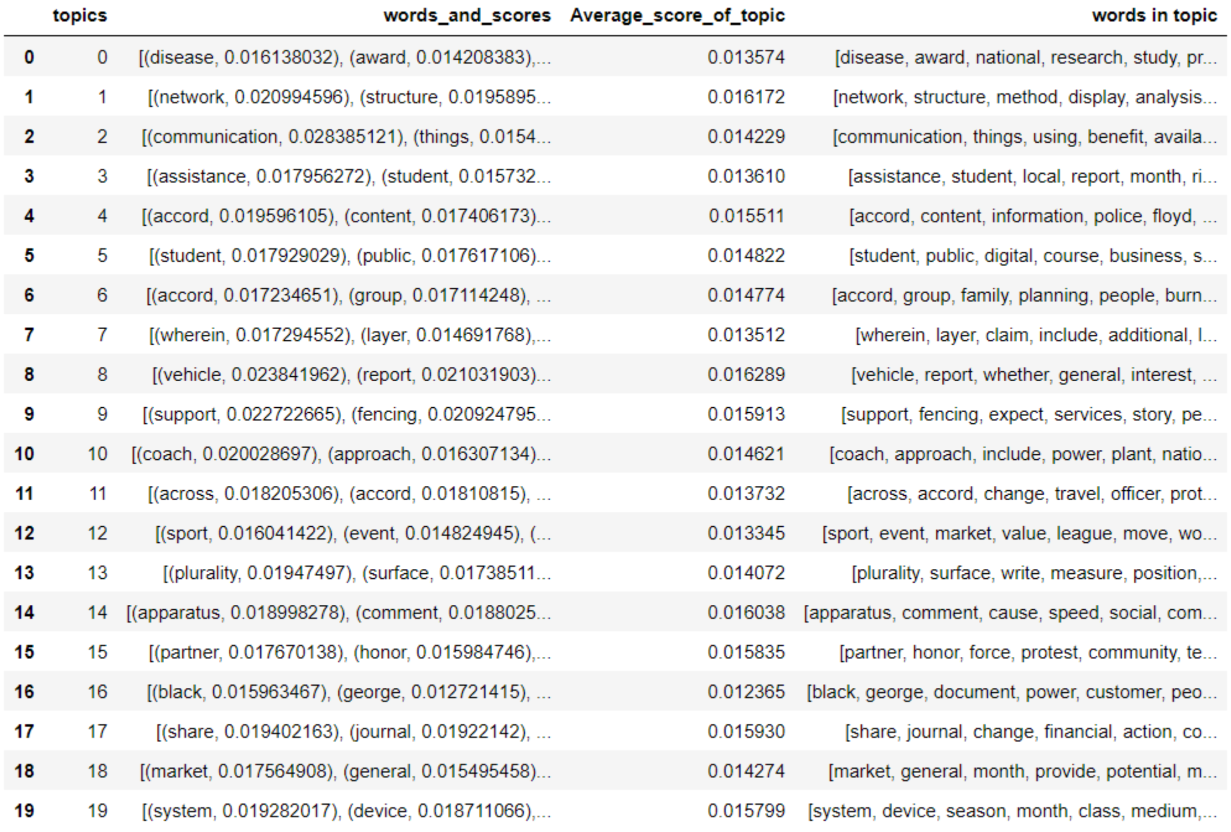


Selected Topic: 2

Slide to adjust relevance metric:(2)
λ = 1

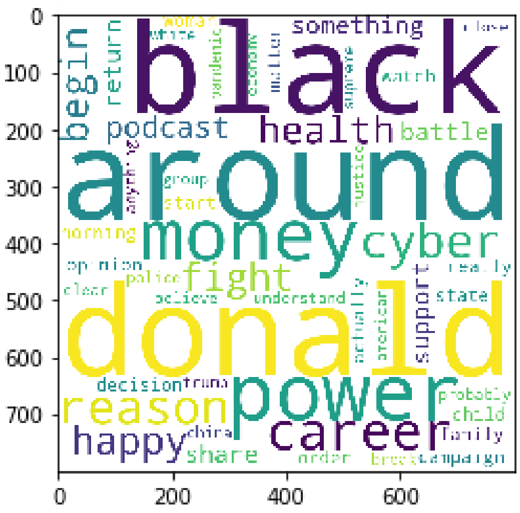
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

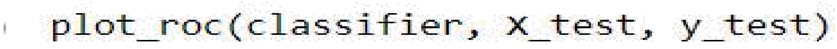
9. Publish your final result in form of a data frame as follows



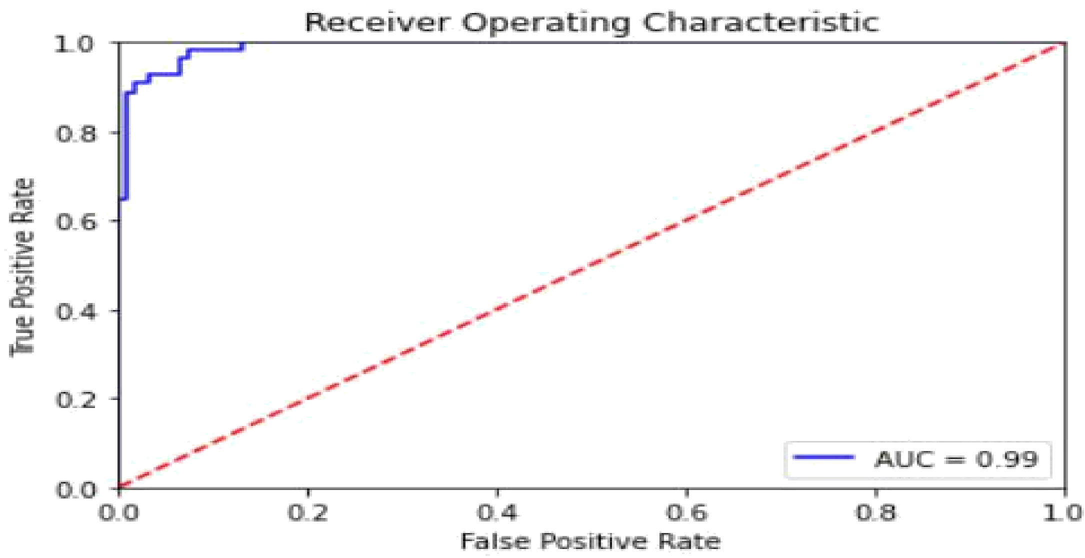
	topics	words_and_scores	Average_score_of_topic	words in topic
0	0	[(disease, 0.016138032), (award, 0.014208383),...	0.013574	[disease, award, national, research, study, pr...
1	1	[(network, 0.020994596), (structure, 0.0195895...	0.016172	[network, structure, method, display, analysis...
2	2	[(communication, 0.028385121), (things, 0.0154...	0.014229	[communication, things, using, benefit, availa...
3	3	[(assistance, 0.017956272), (student, 0.015732...	0.013610	[assistance, student, local, report, month, ri...
4	4	[(accord, 0.019596105), (content, 0.017406173)...	0.015511	[accord, content, information, police, floyd, ...
5	5	[(student, 0.017929029), (public, 0.017617106)...	0.014822	[student, public, digital, course, business, s...
6	6	[(accord, 0.017234651), (group, 0.017114248), ...	0.014774	[accord, group, family, planning, people, burn...
7	7	[(wherein, 0.017294552), (layer, 0.014691768),...	0.013512	[wherein, layer, claim, include, additional, l...
8	8	[(vehicle, 0.023841962), (report, 0.021031903)...	0.016289	[vehicle, report, whether, general, interest, ...
9	9	[(support, 0.022722665), (fencing, 0.020924795...	0.015913	[support, fencing, expect, services, story, pe...
10	10	[(coach, 0.020028697), (approach, 0.016307134)...	0.014621	[coach, approach, include, power, plant, natio...
11	11	[(across, 0.018205306), (accord, 0.01810815), ...	0.013732	[across, accord, change, travel, officer, prot...
12	12	[(sport, 0.016041422), (event, 0.014824945), (...	0.013345	[sport, event, market, value, league, move, wo...
13	13	[(plurality, 0.01947497), (surface, 0.01738511...	0.014072	[plurality, surface, write, measure, position,...
14	14	[(apparatus, 0.018998278), (comment, 0.0188025...	0.016038	[apparatus, comment, cause, speed, social, com...
15	15	[(partner, 0.017670138), (honor, 0.015984746),...	0.015835	[partner, honor, force, protest, community, te...
16	16	[(black, 0.015963467), (george, 0.012721415), ...	0.012365	[black, george, document, power, customer, peo...
17	17	[(share, 0.019402163), (journal, 0.01922142), ...	0.015930	[share, journal, change, financial, action, co...
18	18	[(market, 0.017564908), (general, 0.015495458)...	0.014274	[market, general, month, provide, potential, m...
19	19	[(system, 0.019282017), (device, 0.018711066),...	0.015799	[system, device, season, month, class, medium,...

10. Create a word cloud of top 10 words.


black
around
donald
money
power
career
reason
happy
cyber
fight
health
podcast
begin
return
battle
something
woman
white
pandemic
economy
million
supreme
watch
close
group
start
justice
anything
morning
opinion
police
really
clear
believe
understand
actually
american
support
state
decision
trump
probably
child
family
china
share
order
break
campaign
0
100
200
300
400
500
600
700
0
200
400
600



```
plot_roc(classifier, X_test, y_test)
```


Receiver Operating Characteristic
True Positive Rate
False Positive Rate
AUC = 0.99