EARNING CALL TRANSCRIPTS

Mentor: Manpreet Makkad

Presented by:

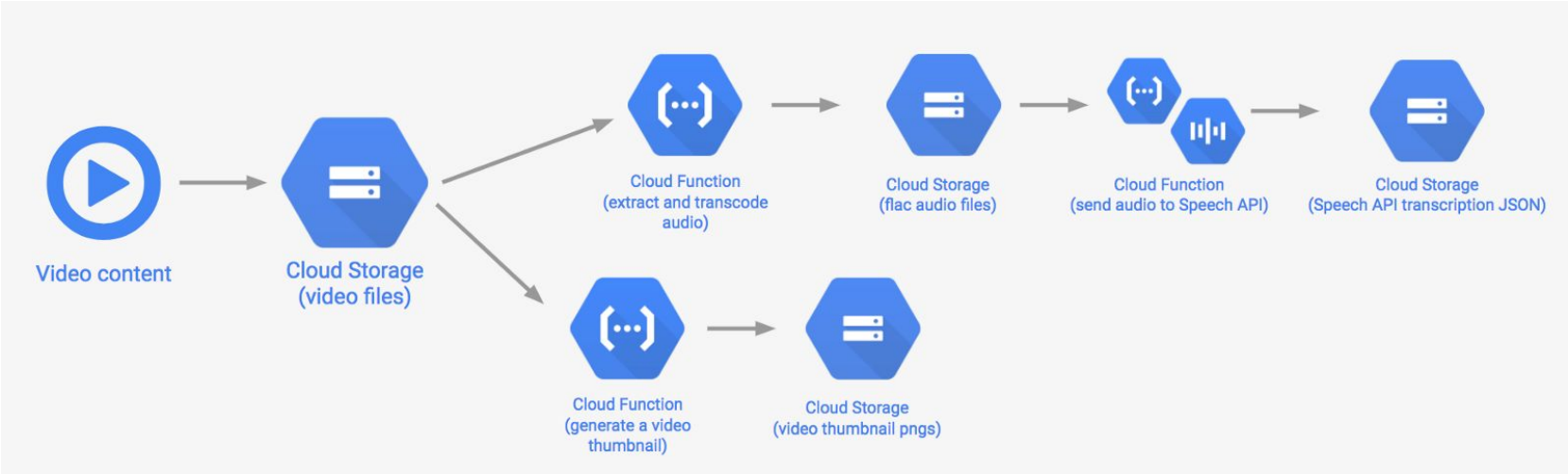
	GEID
Nimit Shah	1011139629
Rahasya Barkur	1011139547
Kancharapu Anil Kumar	1011139542
Annlin Chacko	1011139585

TASK 0: CONVERTING THE AUDIO CALL TO TEXT

Method 1: Using python's inbuilt speech recognition library : SpeechRecognition

Issue: Does not work on larger audio files so it is unusable for our purpose

Method 2: Using google's Speech Recognition API



An input of a ~1 hour long audio file was converted in around 25-30 minutes

Issues :

Conversion is inaccurate. Many words are missing and a few misspelt. For an input audio consisting of ~8800 words, the output obtained contained only ~4500 words.

Initial setup is complex

*Alternate method of web scraping could be used to extract earning call transcripts. This method was used to extract data from Seeking Alpha’s site for our project.

TASK 0: EXTRACT EARNING CALL TRANSCRIPT

Using Web Scraping techniques to extract information (Earning call transcripts) from seeking Alpha's site.

WORKFLOW

1. Send Request and Load the webpage
2. Parse the content for desired data.--using Beautiful Soup library
3. Store the data as text file

BeautifulSoup is a Python library used for parsing documents (i.e. mostly HTML or XML files). Using Requests to obtain the HTML of a page and then parsing whichever information you are looking for with BeautifulSoup from the raw HTML is the quasi-standard web scraping.

TASK 1: SUMMARIZATION OF THE CALLS

Broadly categorized into 2 types :

1. Extractive Summarization : Attempts to summarize an article by selecting a subset of words that retain the most important points
2. Abstractive Summarization : Selects words based on semantic understanding. Aims at producing important material in a new way.

Which method is better ?

Abstractive summarization works better for our cause, as it is observed to work better on larger texts but it requires huge dataset and is a scope of improvement

Step 1: Text Preprocessing - Cleaning the data

1.Removing punctuation

2.Remove stopwords

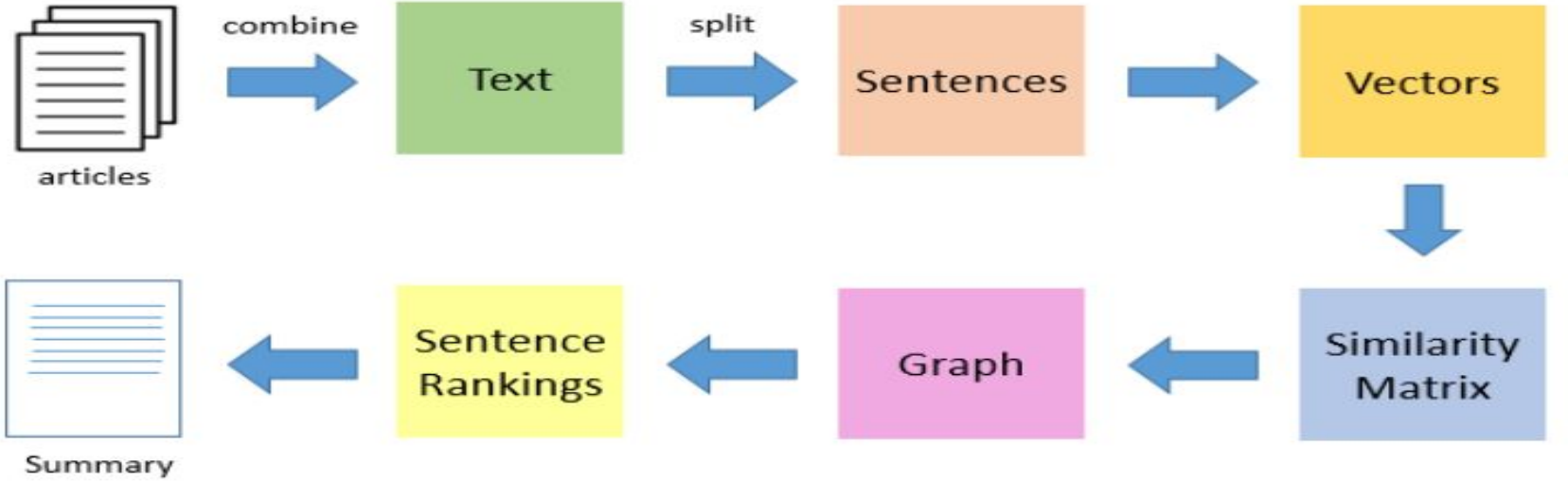
3.Remove additional spaces and digits

4.And lemmatise the Text

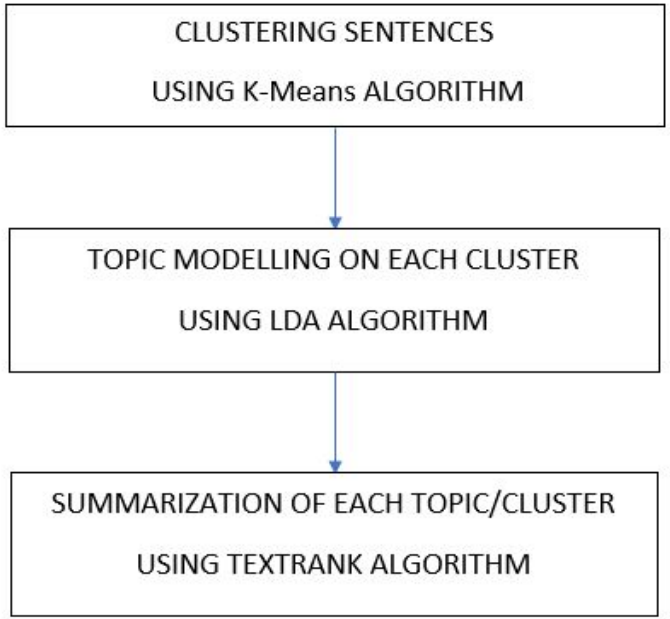
5.Returns cleaned list.

Method1:Summarizing the whole Text file using Text Rank Algorithm

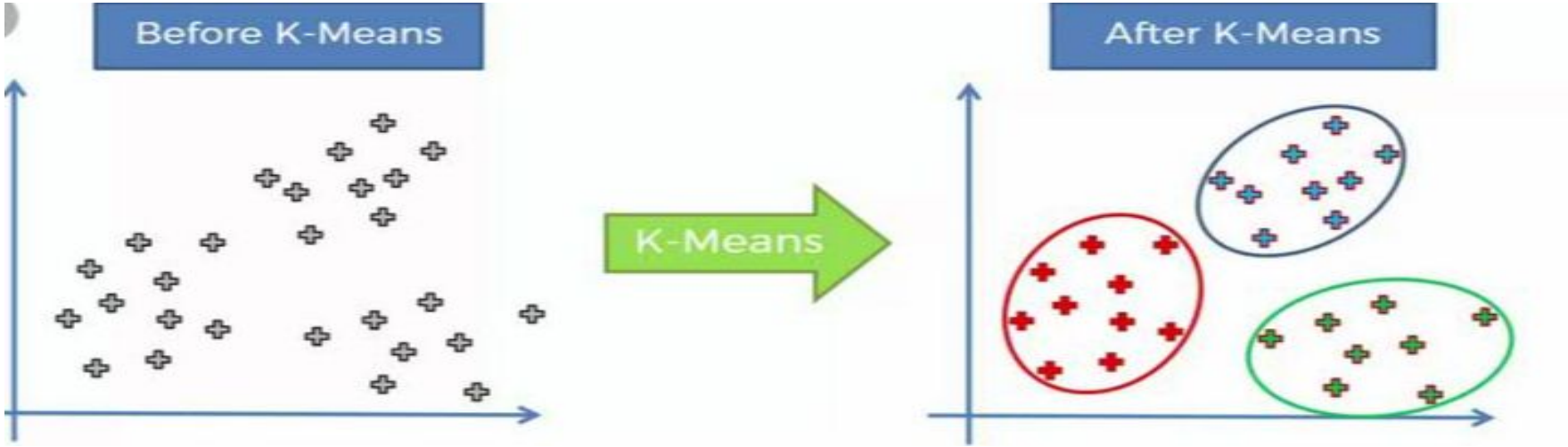
Text Rank Algorithm: TextRank is an extractive and unsupervised text summarization technique.



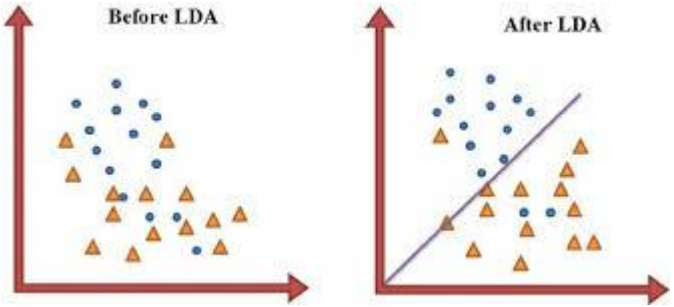
Method 2: Clustering and Topic Modelling Before Summarization



K-Means Algorithm:Grouping similar data points together and discover underlying patterns. K-Means looks for a fixed number (k) of clusters in a dataset and identifies k number of centroids and allocates every data point to the nearest cluster.



Latent Dirichlet Algorithm (LDA): It is a “generative probabilistic model” that allows sets of observation to be explained by unobserved groups that explain why some parts of the data are similar. Here,observations are words collected into documents,it posits each document is a mixture of a small number of topics and that each word’s presence is attributable to one of the document’s topics.



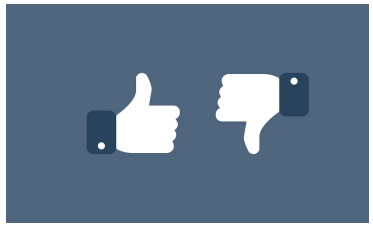
Method 2 gave better results due to the following reasons:

1. By Clustering, it was made sure that the important but less frequent topics were also included which was not the case with Method 1.

2. By Topic Modelling, Important topics related to the cluster was found out which made it easier to understand what the cluster is about.

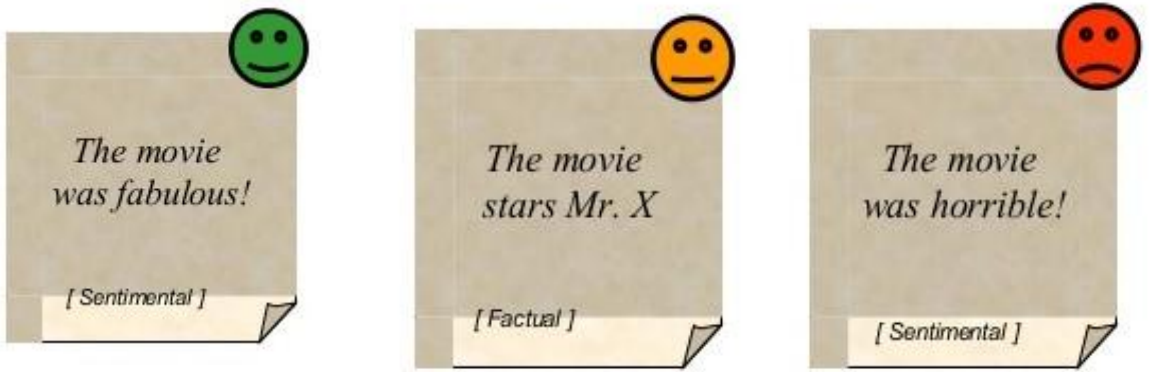
SENTIMENT ANALYSIS

TASK - 2

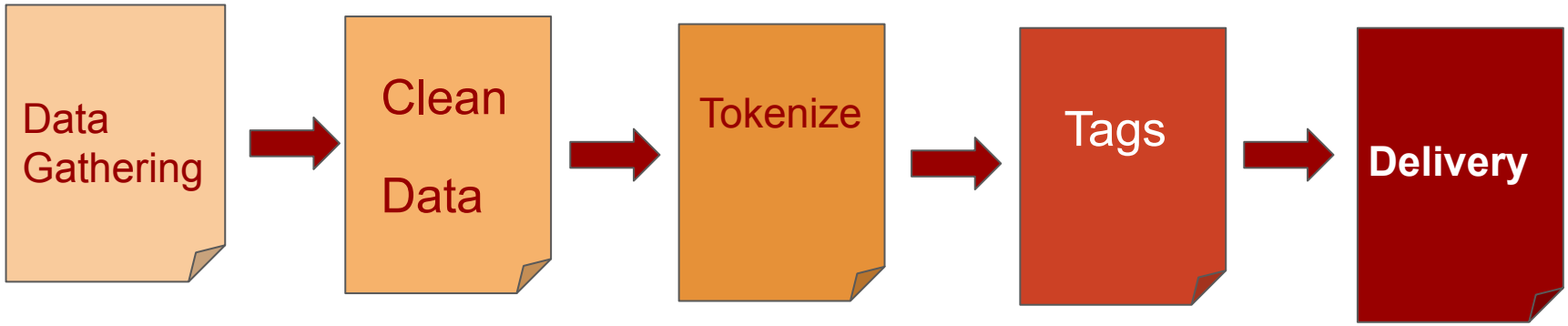


What is Sentiment Analysis?

Identify the orientation of opinion in a piece of text, in other words to determine if a sentence or a document expresses positive, negative, neutral sentiment towards some object.



Process Flow



Read from Text File.

Remove punctuation, convert to lower case

Group the text by sentences

Tags sentences based on polarity

Displays output thru graphical representation

TAGS

METHODS:

1. TEXT BLOB

2. VADER Sentiment Analysis

TEXT BLOB

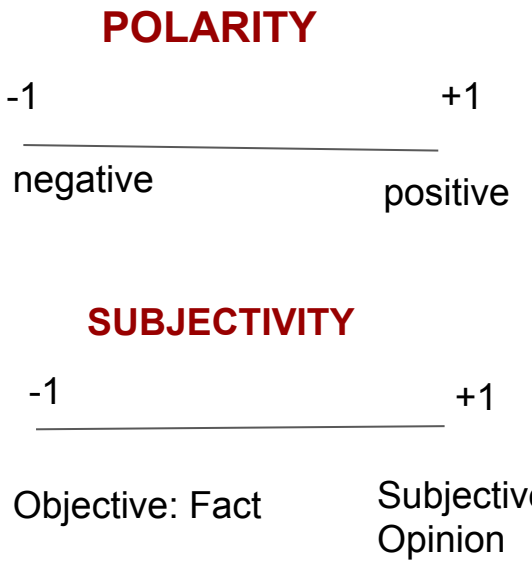
- Is a Python library that is build on top of **nltk** (natural language toolkit)
- Provides the polarity and subjectivity of a sentence

```
from textblob import TextBlob
TextBlob("Today is Monday").sentiment
```

Output: `Sentiment(polarity=0.0, subjectivity=0.0)`

```
from textblob import TextBlob
TextBlob("I love nlp").sentiment
```

Output: `Sentiment(polarity=0.5, subjectivity=0.6)`



TEXT BLOB

```
TextBlob("great").sentiment
```

```
Sentiment(polarity=0.8,
subjectivity=0.75)
```

```
TextBlob(" not great").sentiment
```

```
Sentiment(polarity=-0.4,
subjectivity=0.75)
```

```
TextBlob(" very great").sentiment
```

```
Sentiment(polarity=1.0,
subjectivity=0.9750000000000001)
```

After removing stop words from a sentence each word is assigned a polarity score and a subjectivity score. Which is then averaged for the sentence.

VADER Sentiment Analysis

Follows the same logic as TextBlob but uses a different Lexicon (dictionary).

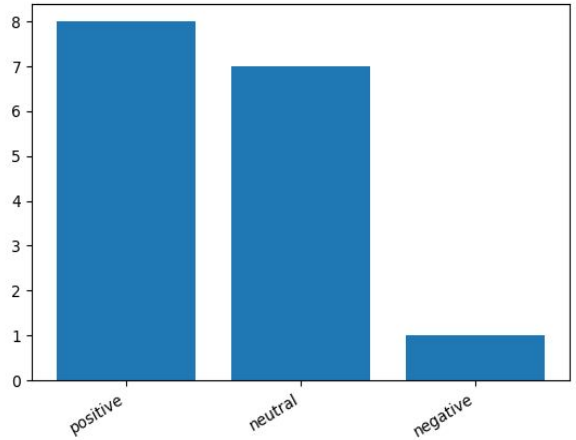
Is a Python library that is build on top of **nltk**

Library Used:

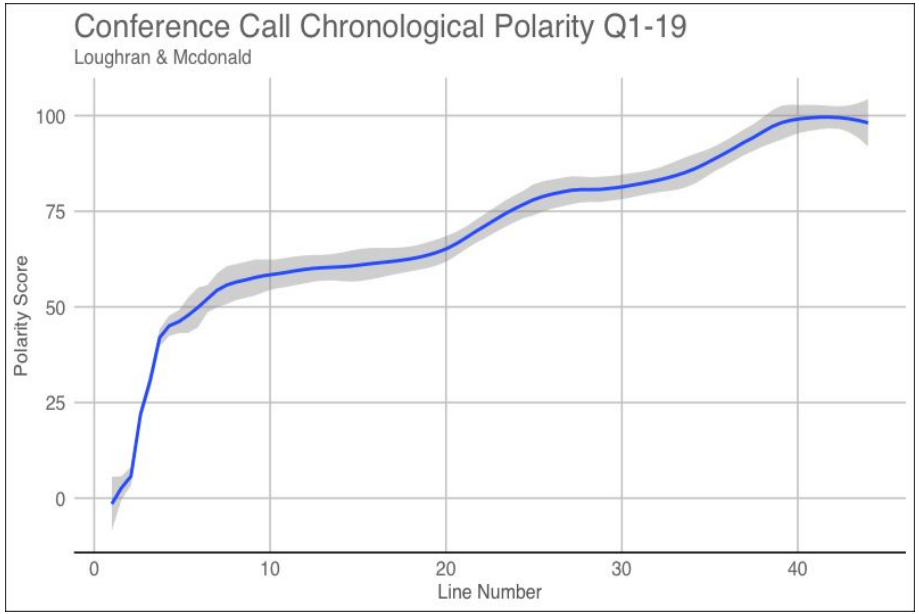
```
nltk.download('vader_lexicon')
from nltk.sentiment.vader import SentimentIntensityAnalyzer
score=SentimentIntensityAnalyzer().polarity_scores("great")
print(score)
```

Output: `{'neg': 0.0, 'neu': 0.0, 'pos': 1.0, 'compound': 0.6249}`

VISUALISATION



Count of Positive, Negative and Neutral sentences.



TASK 3:SCORING LINGUISTIC COMPLEXITY

textstat : Python package to calculate statistics from text to determine readability,complexity and grade level of a particular corpus.The following scores out of the mentioned can be calculated using these packages

a)The Dale-Chall formula: It compares text to a known repository of words to understand if the word in the given text is complex or not.

b)The Gunning fog formula:estimates an index representing the years of formal education a person needs to understand the text on the first reading

c)Fry Readability graph: is calculated by the average number of sentences and syllables per hundred words.

d)McLaughlin’s SMOG formula:similar to gunning formula, estimates the years of education but widely used for checking health messages

e)The FORCAST formula: it is useful for texts without complete sentence, uses only vocabulary element but used for army-job reading materials

f)Readability and Newspaper readership: ease with which readers can understand

g)Flesch scores: calculates the average length of the sentence and the average number of syllables per word to measure reading ease.

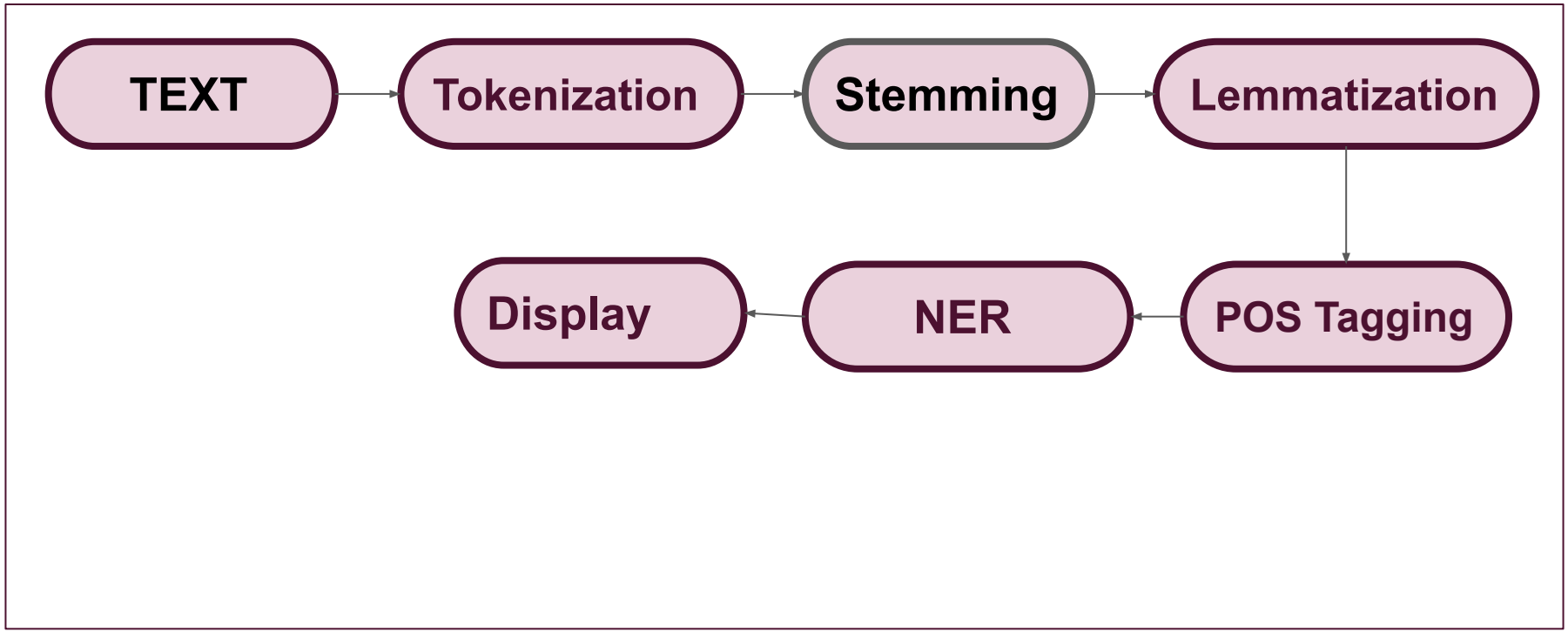
Conclusion: The Dale-Chall formula and Flesch scores seems to fulfill our purpose

Task 4: NAMED ENTITY RECOGNITION(NER)

Named-entity recognition is a subtask of information extraction that seeks to locate and classify named entity mentioned in unstructured text into predefined categories such as:person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages,

Figure 1: An example of NER application on an example text

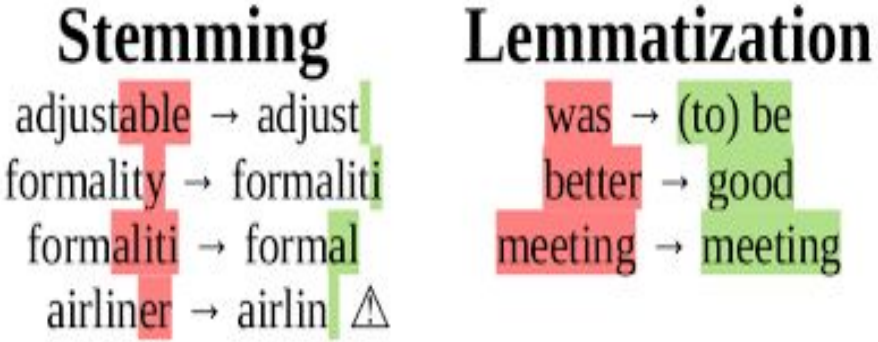
Process Flow

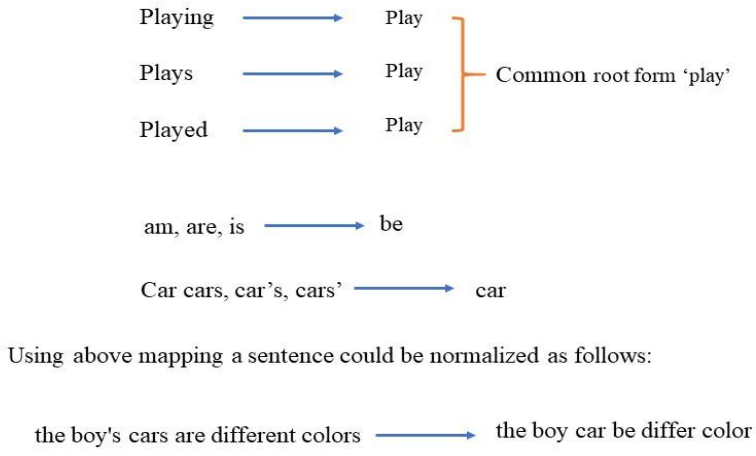


Stemming and Lemmatization

Stemming: Remove ‘ing’ ‘al’

Lemmatization: Producing dictionary form of word

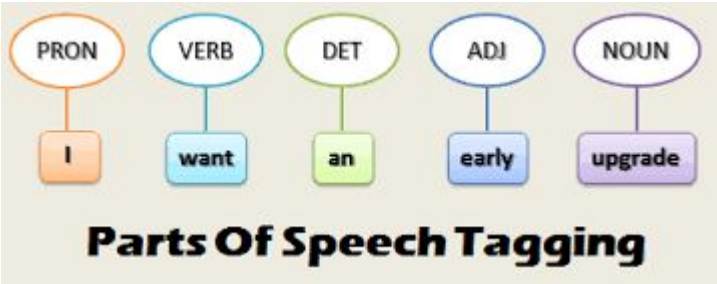




NER

IN POS tagging each word is assigned a fine grained parts of speech tag using

- Spacy library and loading en_core_web_sm english library from spacy.



NER

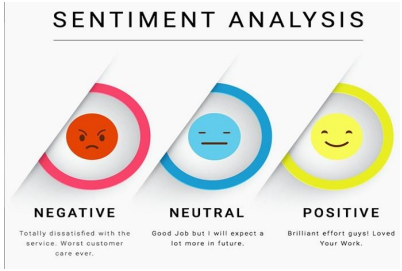
Doc.ents from Spacy library is used to extract named entities. Predefined entities are shown below. From spacy.matcher using PhraseMatcher and Entity Ruler new entity list can be created or could be added to the existing list. It's then displayed using displacy.render.

TYPE	DESCRIPTION
PERSON	People, including fictional
NORP	Nationalities or religious or political groups
FACILITY	Buildings, airports, highways, bridges, etc
ORG	Companies, agencies, institutions, etc
GPE	Countries, cities, states
LOC	Non-GPE locations, mountain ranges, bodies of water
PRODUCT	Objects, vehicles, foods, etc (Not services)
EVENT	Named hurricanes, battles, wars, sports events, etc
WORK_OF_ART	Titles of books, songs, etc
LAW	Named documents made into laws
LANGUAGE	Any named language
DATE	Absolute or relative dates or periods.
TIME	Times smaller than a day
PERCENT	Percentage, including "%".
MONEY	Monetary values, including unit
QUANTITY	Measurements, as of weight or distance
ORDINAL	"first", "second", etc
CARDINAL	Numerals that do not fall under another type

Task 6: CLUSTERING OF COMPANIES

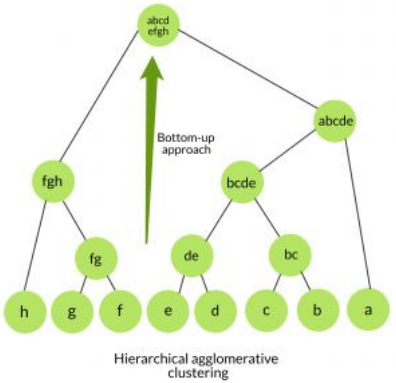
Clustering of Companies/Trade Ideas: There are various methods to cluster the companies. Hierarchical based and Partitioning are the most commonly methods used for clustering.

Input Parameters: Sentiment Analysis, Linguistic Complexity and Products

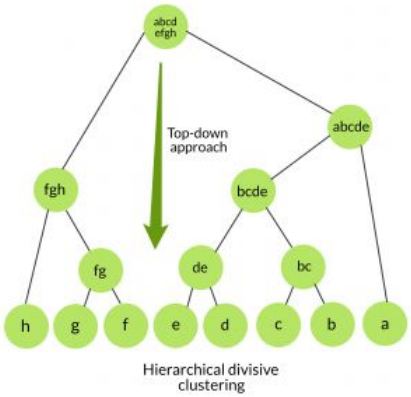


Score	Notes
4.9 or lower	easily understood by an average 4th-grade student or lower
5.0–5.9	easily understood by an average 5th or 6th-grade student
6.0–6.9	easily understood by an average 7th or 8th-grade student
7.0–7.9	easily understood by an average 9th or 10th-grade student
8.0–8.9	easily understood by an average 11th or 12th-grade student
9.0–9.9	easily understood by an average 13th to 15th-grade (college) student
10.0 or higher	easily understood by an average college graduate

Hierarchical Based Method: The clusters formed in this method forms a tree-type structure based on hierarchy. New clusters will form using previously formed ones.



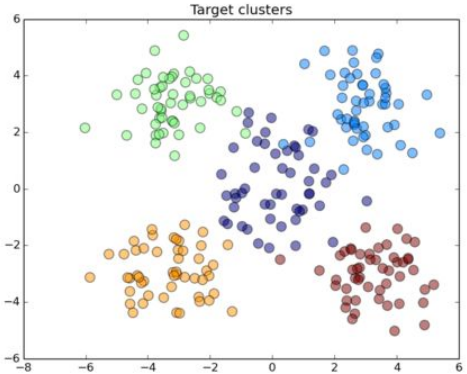
Individual to single cluster



Single cluster to individual

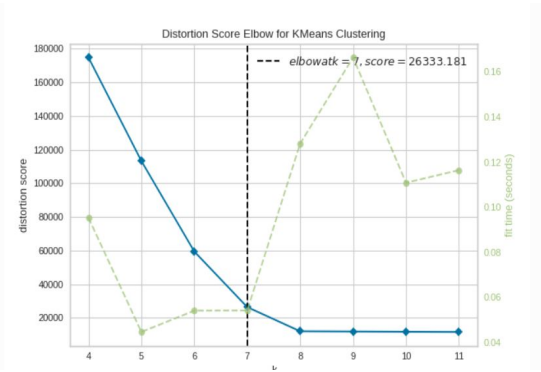
Partitioning Methods: These methods partition the object into k clusters and each partition forms one cluster. Distance parameter is used.

K-means: K-Means looks for a fixed number (k) of clusters in a dataset and identifies k number of centroids and allocates every data point to the nearest cluster. It is a unsupervised Machine Learning Algorithm.



How to find K in K-means?

Elbow Method: The “elbow” method select the optimal number of clusters by fitting the model with a range of values for K. If the line chart resembles an arm, then the “elbow”(the point of inflection on the curve) is a good indication that the underlying model fits best at that point.



X-axis : Number of clusters

Y-axis : Distortion Score

SCOPE FOR IMPROVEMENT

- Using RNN, LSTM - supervised methods to increase the accuracy of text summarisation.
- Similarly sentiment Analysis can be performed using LSTM
- Sentiment Analysis plotted for different quarters within a company
- Sentiment Analysis distinguished between management and analyst during each call.
- Visualise if there is any correlation between market performance and sentiment of earning call.
- Applying KNN method for clustering of companies.
- Writing outputs to csv files.

SCOPE FOR IMPROVEMENT

- Improve accuracy of Name entity recognition using input data set of wide range of values.
- Using supervised models to increase accuracy
- Generate insights and trade ideas from earning call
- Plot performance of various products of a company for different quarters using count frequency.