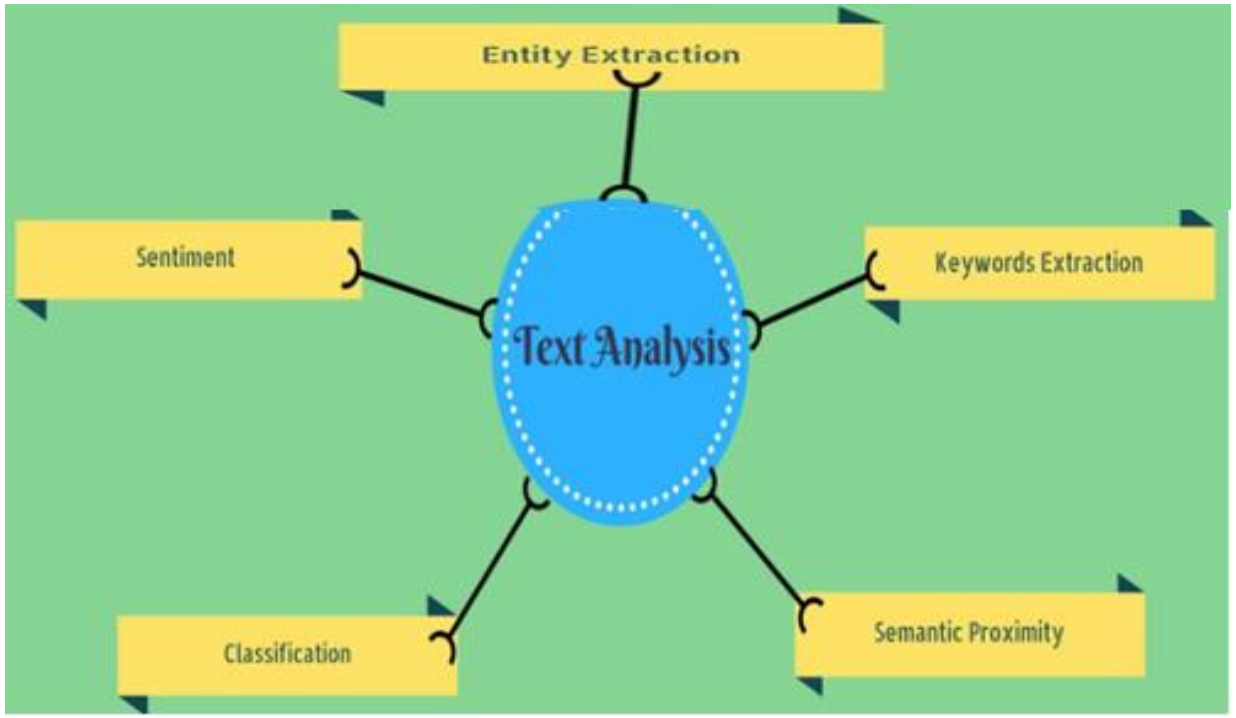
Text Analytics

Objective

- What is text analytics?
- Bag of Words approach
- Tf-Idf
- Example : Spam-Ham Detection

Text Analytics

- Text analytics is the process of derivation of high-end information.
- Make use of established patterns and trends in a piece of text.
- Used to inspect and classify documents.
- Used to extract the sentiment and text semantics



Reference : https://images.app.goo.gl/w6psDYmrqWcKehYt9

Bag of Words Approach

- The simplest form of text representation in numbers.
- Example:
- Review 1: This movie is very scary and long
- Review 2: This movie is not scary and is slow
- Review 3: This movie is spooky and good

	1 This	2 movie	3 is	4 very	5 scary	6 and	7 long	8 not	9 slow	10 spooky	11 good	Length of the review(in words)
Review 1	1	1	1	1	1	1	1	0	0	0	0	7
Review 2	1	1	2	0	0	1	1	0	1	0	0	8
Review 3	1	1	1	0	0	0	1	0	0	1	1	6

Reference : https://cdn.analyticsvidhya.com/wp-content/uploads/2020/02/BoWBag-of-Words-model-2.png

Term Frequency-Inverse Document Frequency (TF-IDF)

- Term Frequency (TF): It is a measure of how frequently a term t appears in a document d.
- Fraction of repeating of single term present in a given document.
- To account most repeating term.

Term	Review 1	Review 2	Review 3	TF (Review 1)	TF (Review 2)	TF (Review 3)
This	1	1	1	1/7	1/8	1/6
movie	1	1	1	1/7	1/8	1/6
is	1	2	1	1/7	1/4	1/6
very	1	0	0	1/7	0	0
scary	1	1	0	1/7	1/8	0
and	1	1	1	1/7	1/8	1/6
long	1	0	0	1/7	0	0
not	0	1	0	0	1/8	0
slow	0	1	0	0	1/8	0
spooky	0	0	1	0	0	1/6
good	0	0	1	0	0	1/6

Reference : Term Frequency Table

Inverse Document Frequency (IDF)

- Assign weightage to terms on behalf of repetition.
- Inverse Logarithmic fraction of documents having specific term.
- Log (no. of documents/ no. of documents having term t)
- Review 1: **This** movie is very scary and long
- Review 2: **This** movie is not scary and is slow
- Review 3: **This** movie is spooky and good
- IDF('**This**') = log(number of documents/number of documents containing the word 'this') = log(3/3) = log(1) = 0

Spam Ham Demonstration

- Spam Messages, Spam emails or messages belong to the broad category of unsolicited messages received by a user.
- Bag-of-words which is a natural language processing (NLP) algorithm can be used to classify messages as ham or spam.

Reference : SPAM-HAM

Lexicons for Semantic Analytics

- Sentiment analysis is a process by which information is analyzed using natural language processing (NLP).
- Lexicons calculate the sentiment from the semantic orientation of word or phrases that occur in a text.
- Sentiment present in sentence may be positive, negative or neutral.
- Mostly dependent of presence of intention specific terms.
- Sentiment Analytic Libraries: Affin, Textblob, VADER etc.

Afinn

- Lexicons used for sentiment analysis.
- Developed by Finn Årup Nielsen.
- It contains 3300+ words with a polarity score associated with each word.
- Initialize afinn sentiment analyzer.

  from afinn import Afinn

  af = Afinn()
- Identifies the sentiment specific term in sentence.

TextBlob

- A simple python library that offers API access to different NLP tasks such as sentiment analysis, spelling correction etc.
- Polarity: float, lies between [-1,1]

  -1 symbols negative sentiment

  +1 symbols positive sentiment.
- Subjectivity: float, lies in the range of [0,1].
- Generally refer to personal opinion, emotion, or judgment.

```
from textblob import TextBlob

testimonial = TextBlob("The food was great!")
print(testimonial.sentiment)
```

```
Sentiment(polarity=1.0, subjectivity=0.75)
```

VADER

- Valence aware dictionary for sentiment reasoning
- Returns the probability of a given input sentence to be Positive, Negative, and Neutral.
- In day to day life, we experience text with complex and fuzzy sentiments.
- Mostly used for social media texts.

```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer


analyzer = SentimentIntensityAnalyzer()
sentence = "The food was great!"
vs = analyzer.polarity_scores(sentence)
print("{:-<65} {}".format(sentence, str(vs)))
```

```
{'compound': 0.6588, 'neg': 0.0, 'neu': 0.406, 'pos': 0.594}
```

Hands On

Thank You