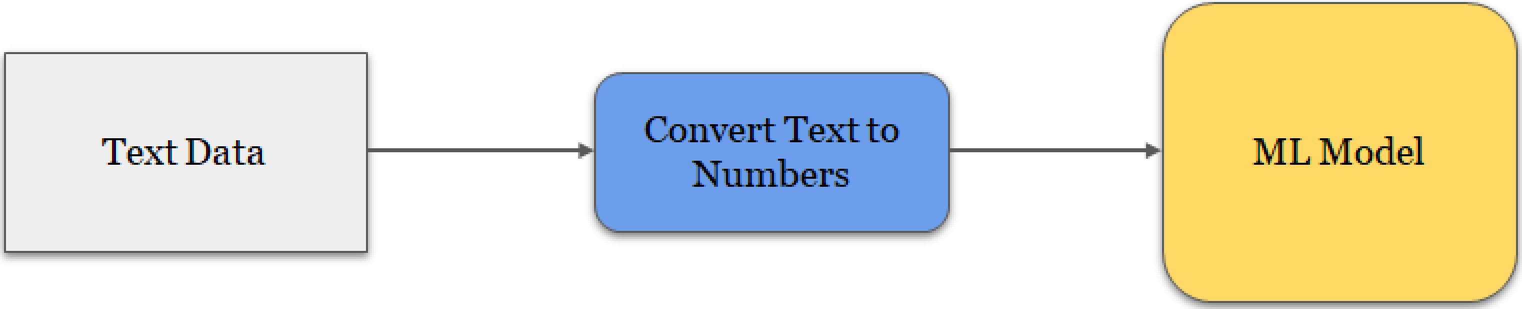
Statistical NLP

S. No	Topic	Scope	Objective
1	Bag of words	Creating features using Bag of words, Countvectorizer	Understand feature creation using Bag-Of-Words
2	Tf-IDF Model	Term frequency, idf	Understand difference between count-vectorizer and tf-idf vectorizer
3	Text classification using ML	How to use ML algorithm to do text classification	Understand use cases of text classification
4	VADER Sentiment Analysis	sentiment Analysis with VADER	Brief idea on sentiment analysis

What after text pre-processing?



Word Representations

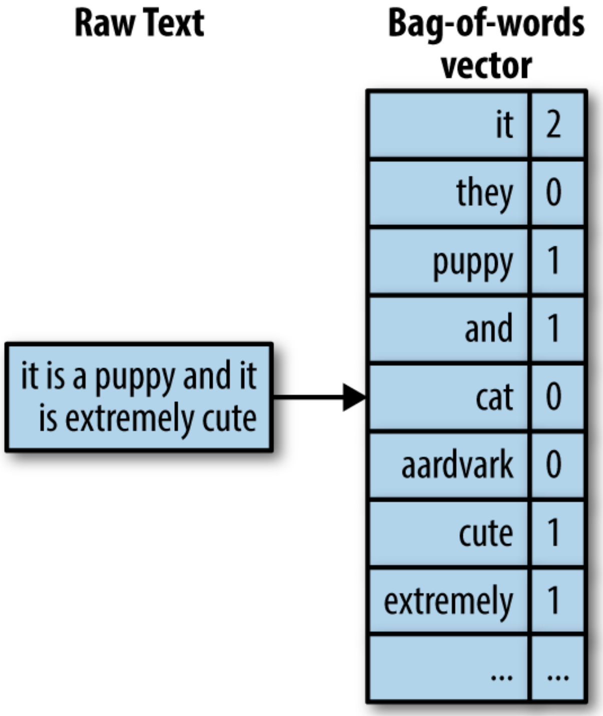
- Sparse
  - Term-document matrix or Document - Term matrix
    - Given a fixed vocabulary, we count the number of times each word occurs in a document for all documents. This matrix is the term - document matrix.
    - We count the number of times a each word pair occurs in the document for a given vocabulary, resulting matrix is the term -term matrix.
  - TF-IDF

Vectorization

- Tf-Idf vectors
  - Tf-idf is similar to term -document matrix with each word occurrence count divided by inverse document matrix.
- One-hot encoding of words.
- Above representations of documents are sparse since most of the elements in the matrix will be zero.
- These representations do not take into account individual word relationships.

Bag of Words

- Feature extraction approach in NLP
- In this model, a text (such as a sentence or a document) is represented as the bag of its words, disregarding grammar and even word order but keeping multiplicity.
- We use the tokenized words for each observation and find out the frequency of each token.



Raw Text

it is a puppy and it is extremely cute

Bag-of-words vector

it	2
they	0
puppy	1
and	1
cat	0
aardvark	0
cute	1
extremely	1
...	...

Bag of words

- We define the vocabulary of corpus as all the unique words in the corpus above and below some certain threshold of frequency.
- Each sentence or document is defined by a vector of same dimension as vocabulary containing the frequency of each word of the vocabulary in the sentence.
- The bag-of-words model is commonly used in methods of document classification where the (frequency of) occurrence of each word is used as a feature for training a classifier.

Tf-idf Vector

- TF-IDF (term frequency times inverse document frequency) is a scheme to weight individual tokens.
- One of the advantage of TF-IDF is reduce the impact of tokens that occur very frequently, hence offering little to none in terms of information.

$$TFIDF\ score\ for\ term\ i\ in\ document\ j = TF(i,j) * IDF(i)$$

$where$

$IDF = Inverse\ Document\ Frequency$

$TF = Term\ Frequency$

$$TF(i,j) = \frac{\text{Term i frequency in document j}}{\text{Total words in document j}}$$

$$IDF(i) = \log_2\left(\frac{\text{Total documents}}{\text{documents with term i}}\right)$$

and

$t = Term$

$j = Document$

Tf-idf Vector

Document #1 -

He is a good boy. She is also good.

He	1
is	2
a	1
good	2
boy	1
she	1
also	1
Total	**9**

$$TF = \frac{Frequency\ of\ the\ word\ in\ a\ Doc}{Total\ number\ of\ words\ in\ the\ Doc}$$

TF(He, doc#1) = 1/9 = 0.11

TF(good, doc#1) = 2/9 = 0.22

TF captures how important a word is to the document (without looking at other documents in the dataset)

Tf-idf Vector

Document #2 - Radhika is a good person.

Radhika	1
is	1
a	1
good	1
person	1
Total	**5**

$$TF = \frac{Frequency\ of\ the\ word\ in\ a\ Doc}{Total\ number\ of\ words\ in\ the\ Doc}$$

TF(He, doc#2) = 0/5 = 0

TF(good, doc#2) = 1/5 = 0.2

Tf-idf Vector

Document #1

He is a good boy. She is also good.

Document #1

Radhika is a good person.

He	1
is	2
a	1
good	2
boy	1
she	1
also	1
Total	**9**

Radhika	1
is	1
a	1
good	1
person	1
Total	**5**

$$IDF = log(\frac{Num\ of\ Docs}{Word\ in\ Num\ of\ Docs})$$

IDF(He) = log(2/1) = 0.301
IDF(good) = log(2/2) = 0

IDF tells us if a word (feature) can be used to distinguish documents. If a word appears in majority of the documents then IDF will be close to '0' i.e. give low weightage to that feature.

Tf-idf Vector

greatlearning
Learning for Life

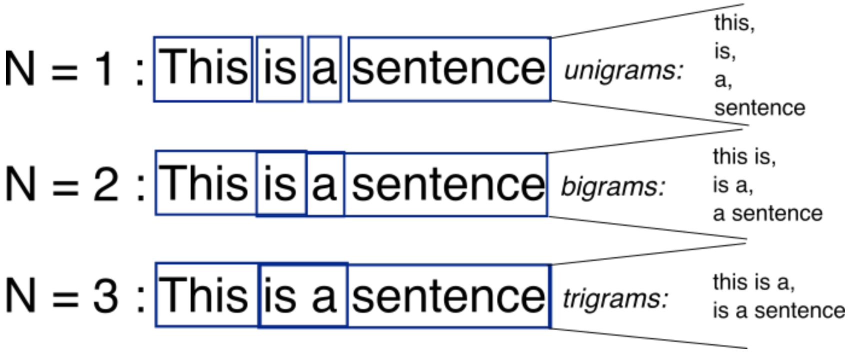
TF-IDF(He, doc#1) = 0.11 * 0.301 = 0.03311
TF-IDF(good, doc#1) = 0.22 * 0 = 0
TF-IDF(He, doc#2) = 0 * 0.301 = 0
TF-IDF(good, doc#2) = 0.2 * 0 = 0

IDF tells us if a word (feature) can be used to distinguish documents . If a word appears in majority of the documents then IDF will be close to ‘0’ i.e. give low weightage to that feature.

	a	also	boy	good	He	Is	person	She	Radhika
Index	0	1	2	3	4	5	6	7	8
Document #1				0	0.03311				
Document #2				0	0				

N-gram

- It’s a sequence of N-words.
- Bi-gram is a special case of N-grams where we consider only the sequence of two words.
- In N-gram models we calculate the probability of Nth words give the sequence of N -1 words. We do this by calculating the relative frequency of the sequence occurring in the text corpus.



Bigram approximation

$$P(w_1^n) = \prod_{k=1}^{n} P(w_k \mid w_{k-1})$$

N-gram approximation

$$P(w_1^n) = \prod_{k-1}^{n} P(w_k \mid w_{k-N-1}^{k-1})$$

Text classifications using ML

We can assign a category to different forms of text i.e. Articles, news, paragraphs, books, web pages etc.

Applications of Text Classification

Text classification has applications like spam filtering, sentiment analysis, document classification

Text based analysis in Marketing, Product Management

Sentiment Analysis

- It’s the content based subjective information retrieval obtained from monitoring online conversations
- Used extensively to get overall end user feedback
- Used for making text classification based automated mailing applications
- Used for gaining insights on areas of improvement in product management

VADER

- VADER(Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.
- VADER uses a combination of sentiment lexicon which is a list of lexical features (e.g., words) which are generally labelled according to their semantic orientation as either positive or negative.

- Like VADER sentiment analysis can be also be done using - NLTK , TextBlob

Sentiment Analysis-Example

- Sentiment analysis – positive or negative
  - “This is a ridiculously priced toothbrush. Seriously, no way to get around it. It is absurdly priced and I'm almost embarrassed to be admitting that I bought it. With that said... Wow, this thing is amazing.”
  - “These pens make me feel so feminine and desirable. I can barely keep the men away when I'm holding one of these in my dainty hand. My husband has started to take fencing lessons just to keep the men away.”

greatlearning
Learning for Life

Thank you!

Happy Learning :)