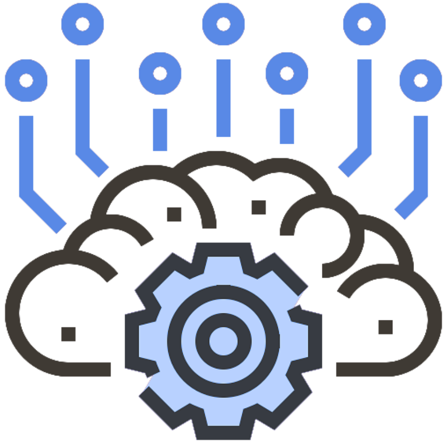
INTRODUCTION TO
NATURAL LANGUAGE PROCESSING

Shilpa Shaju

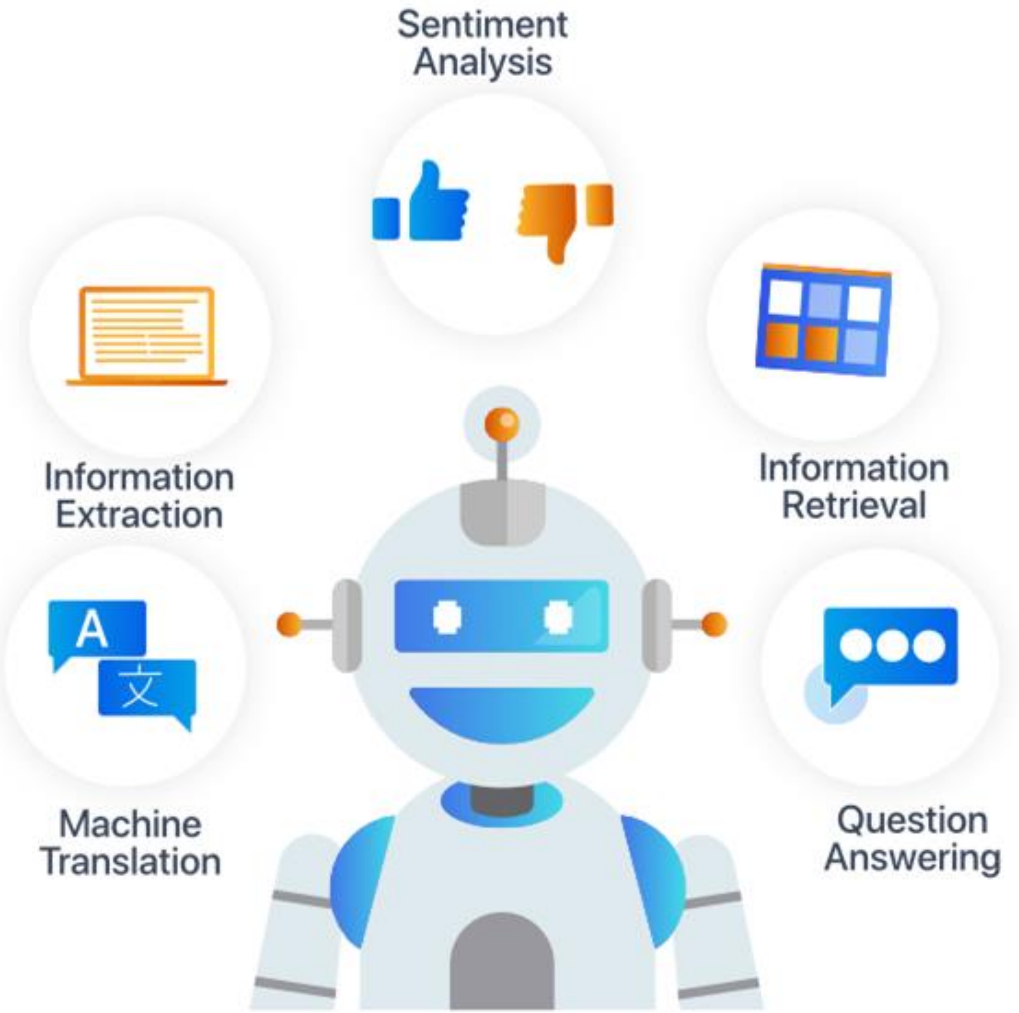


What is NLP??

Natural language processing (NLP)- A branch of Artificial Intelligence that gives machines the ability to understand natural human langauge.

billions of text data being generated every day and most of them are unstructured.

Applications of Natural Language Processing in Different Domains



Sentiment Analysis

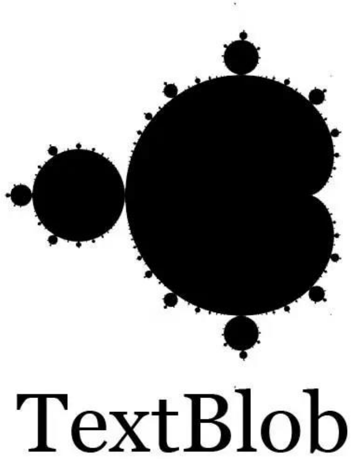
Information Extraction

Information Retrieval

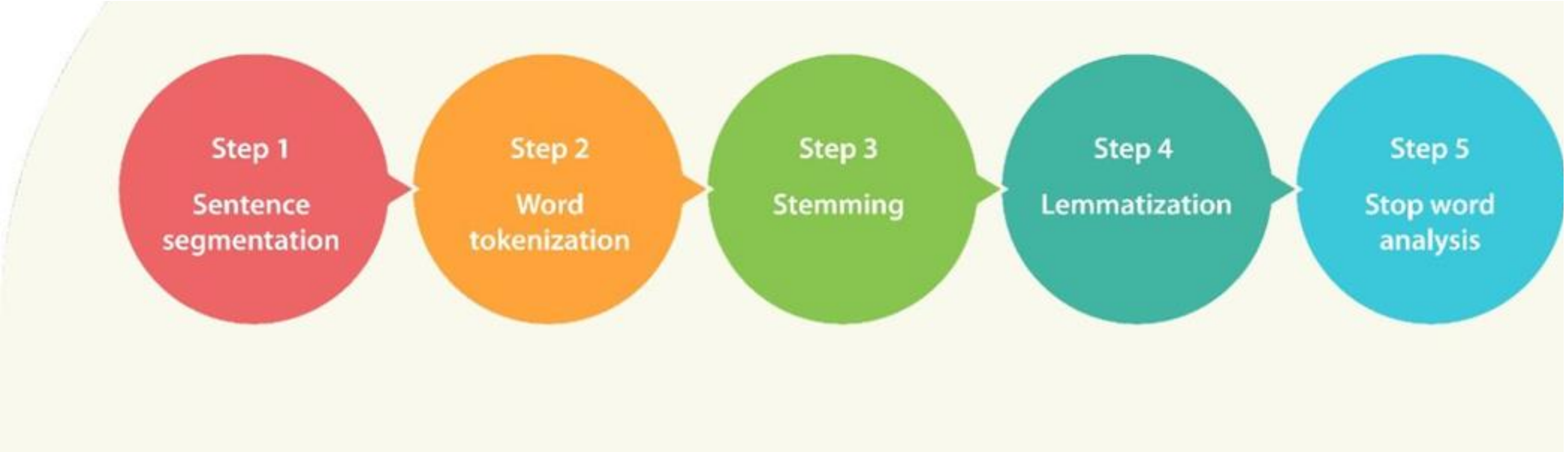
Machine Translation

Question Answering

Python Libraries for NLP

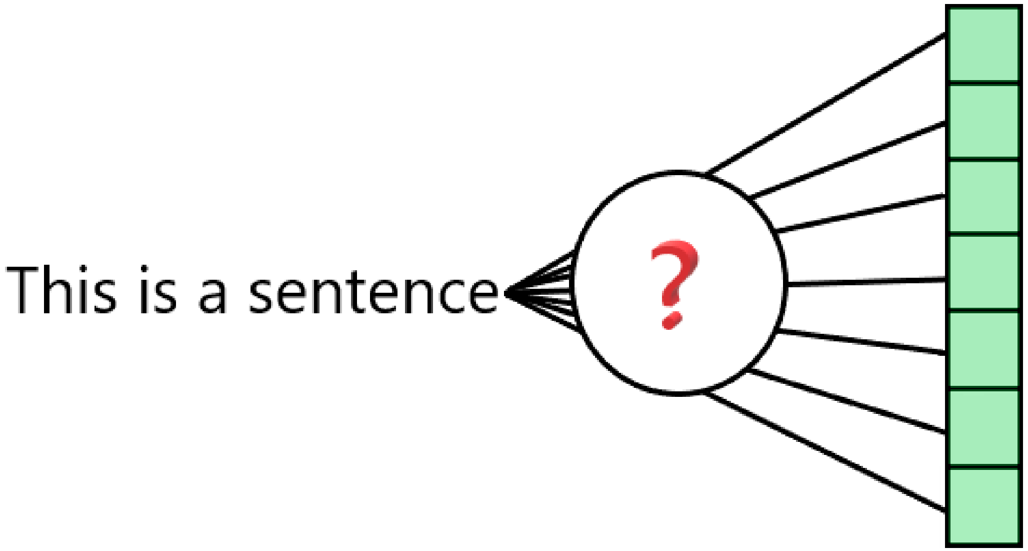


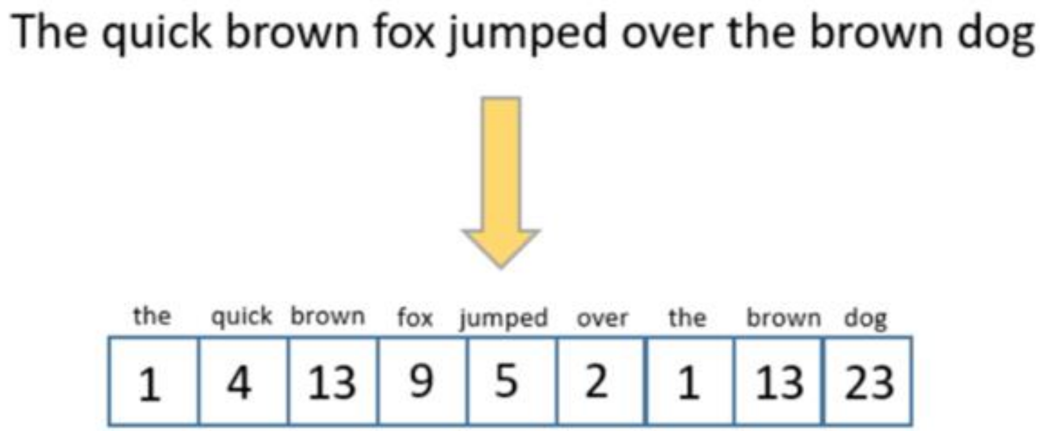
Common Terminologies



Source

Text Features Extraction



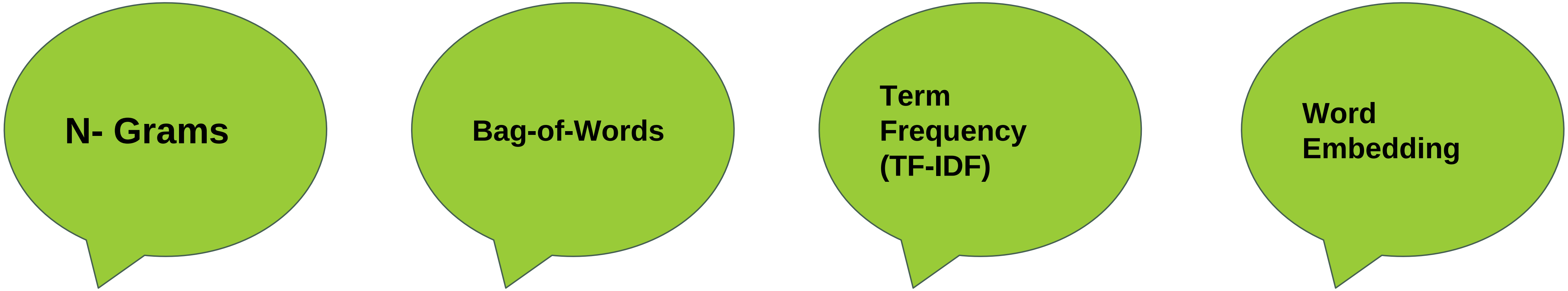

The quick brown fox jumped over the brown dog

the	quick	brown	fox	jumped	over	the	brown	dog
1	4	13	9	5	2	1	13	23

Turning text into vectors that can be then fed to machine learning models in a classical way

Types

Machine learning algorithms cannot work with raw text directly; the text must be converted into numbers. Specifically, vectors of numbers.



N- Grams

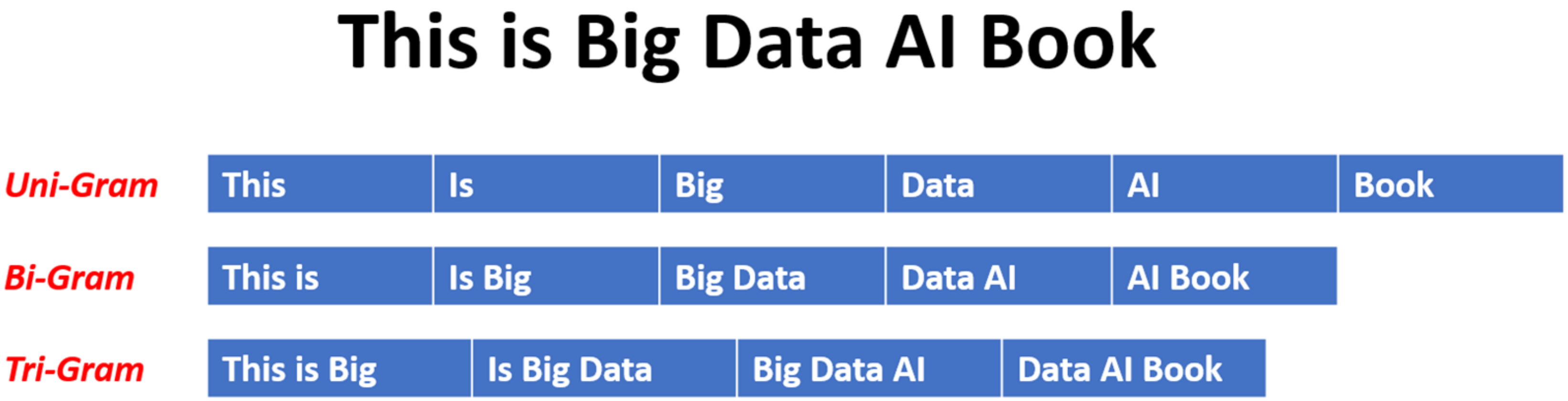
Bag-of-Words

Term Frequency (TF-IDF)

Word Embedding

N-grams

N-grams are the combination of multiple words used together. Ngrams with N=1 are called unigrams. Similarly, bigrams (N=2), trigrams (N=3) and so on can also be used.

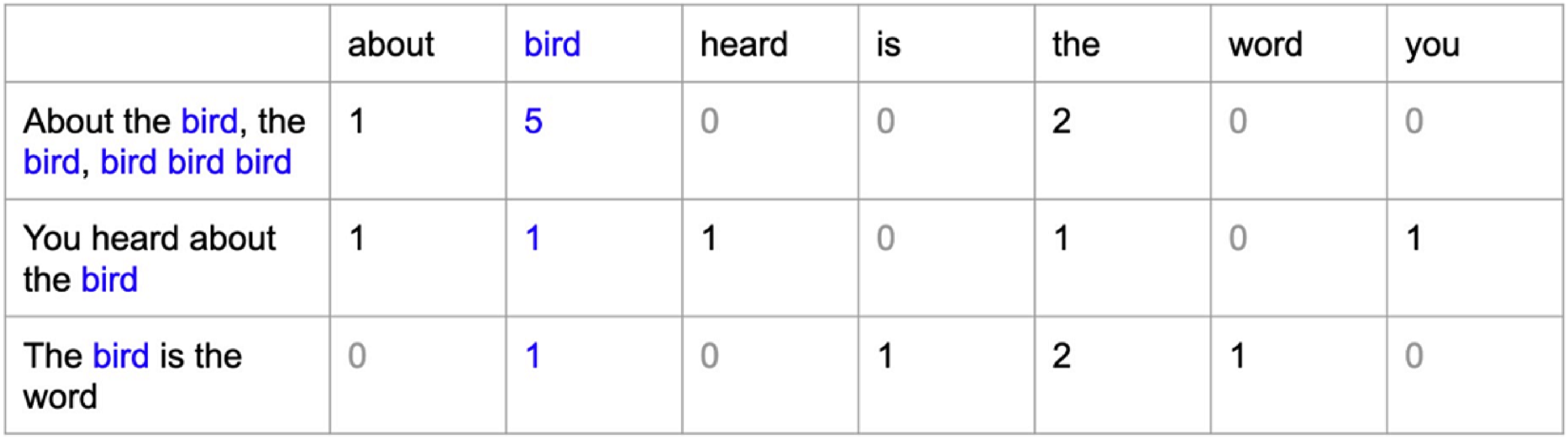


This is Big Data AI Book

Uni-Gram	This	Is	Big	Data	AI	Book
Bi-Gram	This is	Is Big	Big Data	Data AI	AI Book	
Tri-Gram	This is Big	Is Big Data	Big Data AI	Data AI Book		

Bag of Words (BoW)

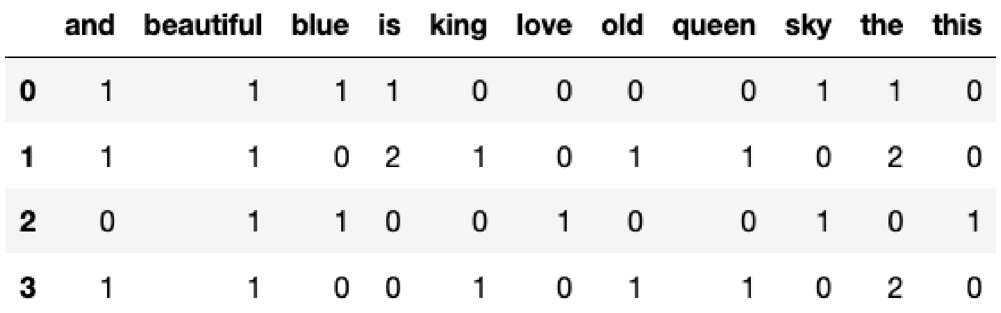
- used to analyze text and documents based on **word count.**
- model does not account for word order within a document.



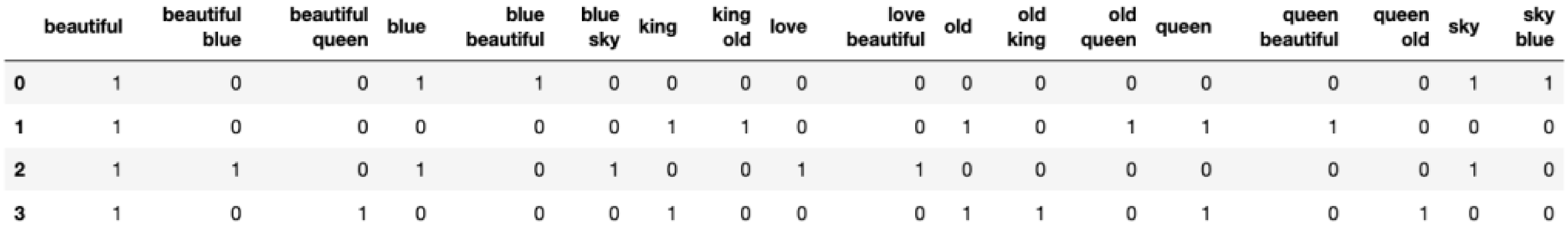
	about	bird	heard	is	the	word	you
About the bird, the bird, bird bird bird	1	5	0	0	2	0	0
You heard about the bird	1	1	1	0	1	0	1
The bird is the word	0	1	0	1	2	1	0

Bag of Words(BOW) Limitation

'The sky is blue and beautiful',
'The king is old and the queen is beautiful',
'Love this beautiful blue sky',
'The beautiful queen and the old king']



	and	beautiful	blue	is	king	love	old	queen	sky	the	this
0	1	1	1	1	0	0	0	0	1	1	0
1	1	1	0	2	1	0	1	1	0	2	0
2	0	1	1	0	0	1	0	0	1	0	1
3	1	1	0	0	1	0	1	1	0	2	0



	beautiful	beautiful blue	beautiful queen	blue	blue beautiful	blue sky	king	king old	love	love beautiful	old	old king	old queen	queen	queen beautiful	queen old	sky	sky blue
0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1
1	1	0	0	0	0	0	1	1	0	0	1	0	1	1	1	0	0	0
2	1	1	0	1	0	1	0	0	1	1	0	0	0	0	0	0	1	0
3	1	0	1	0	0	0	1	0	0	0	1	1	0	1	0	1	0	0

Term Document – Inverse Document Frequnecy Matrix

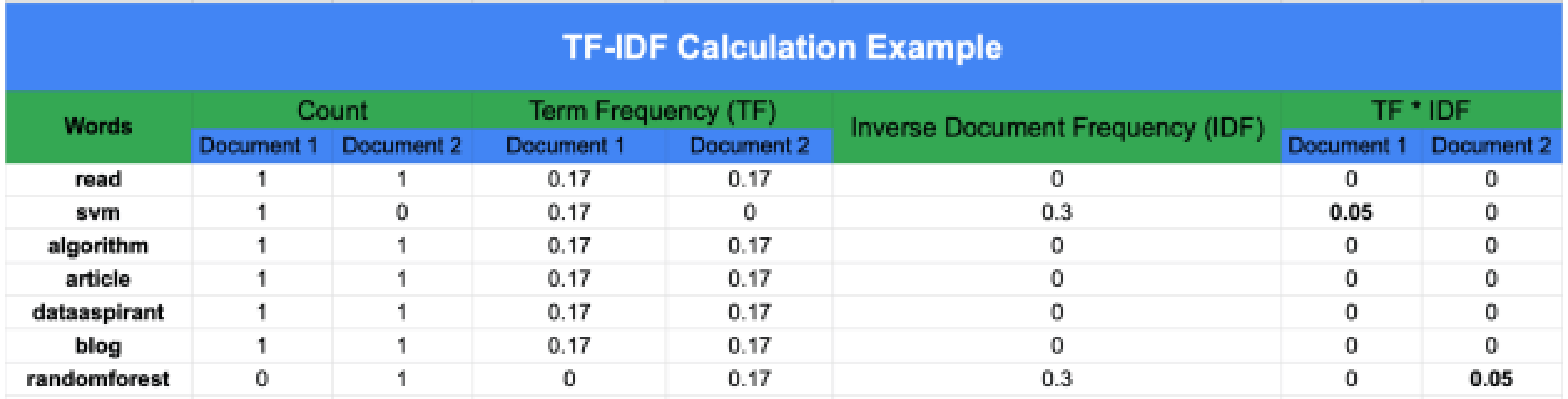
The *term frequency* is a ratio of the count of a word's occurrence in a document and the number of words in the document

Let us show the count of word *i* in document *j* by tf_{ij}

Let us represent document frequency for word *i* by df_i. With *N* as the number of documents in the corpus, the tf-idf weight w_{ij} for word *i* in document *j* is computed by the following formula:

The *document frequency* of word *i* represents the number of documents in the corpus with word *i* in them

$$w_{i,j} = tf_{i,j} \times log(\frac{N}{df_i})$$

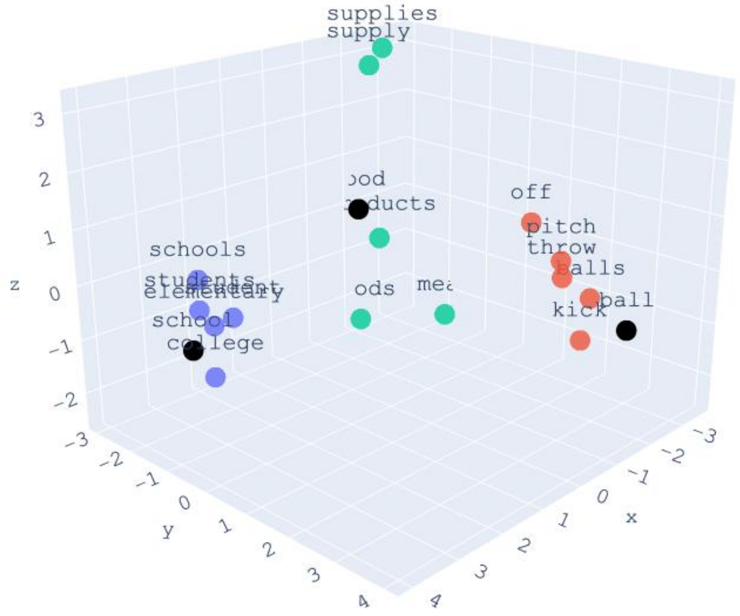


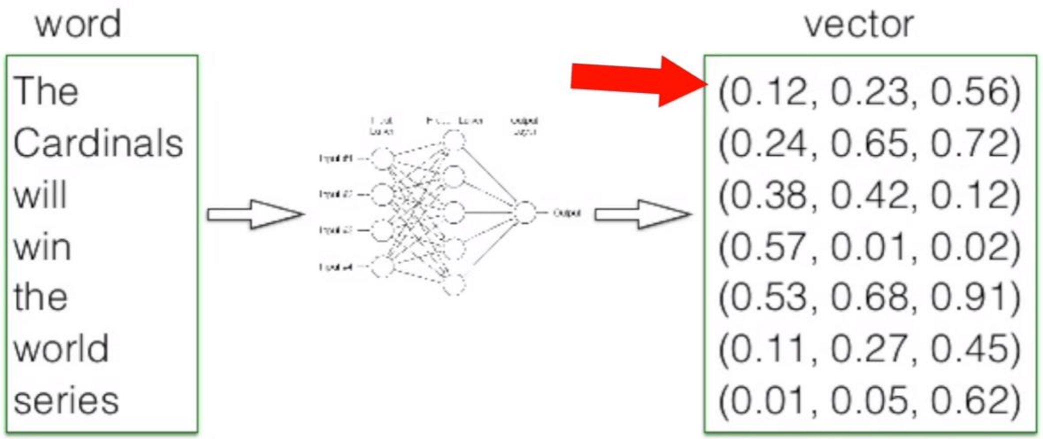
TF-IDF Calculation Example

Words	Count		Term Frequency (TF)		Inverse Document Frequency (IDF)	TF * IDF	
	Document 1	Document 2	Document 1	Document 2		Document 1	Document 2
read	1	1	0.17	0.17	0	0	0
svm	1	0	0.17	0	0.3	**0.05**	0
algorithm	1	1	0.17	0.17	0	0	0
article	1	1	0.17	0.17	0	0	0
dataaspirant	1	1	0.17	0.17	0	0	0
blog	1	1	0.17	0.17	0	0	0
randomforest	0	1	0	0.17	0.3	0	**0.05**

Text Embedding

Word Embedding is the representation of text in the form of vectors. The underlying idea here is that similar words will have a minimum distance between their vectors.





word

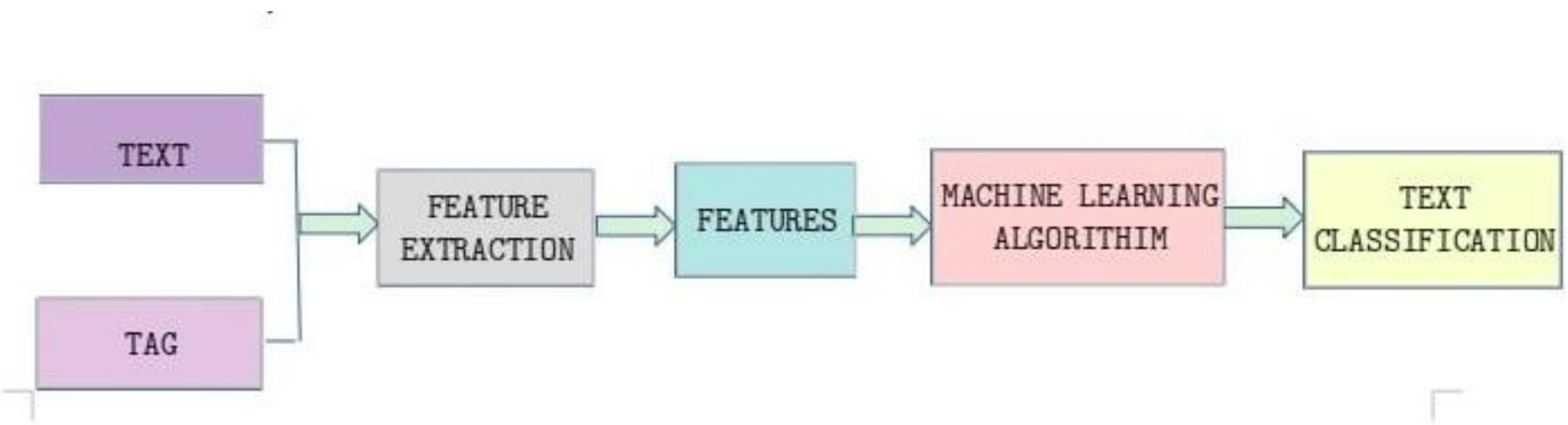
The
Cardinals
will
win
the
world
series

vector

(0.12, 0.23, 0.56)
(0.24, 0.65, 0.72)
(0.38, 0.42, 0.12)
(0.57, 0.01, 0.02)
(0.53, 0.68, 0.91)
(0.11, 0.27, 0.45)
(0.01, 0.05, 0.62)

- Word2Vec
- Doc2Vec

Text Classification -Pipeline



Assignment Question

Thank You

ANY QUESTION