



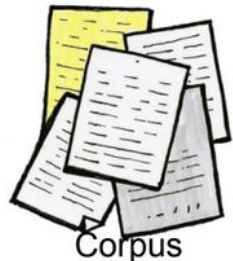
Text Representation

What is Text Representation?



What is Text Representation?

Converting text data from its raw form into numerical format.



```
100110100001100100001110101100110011000011001000011101011001
00110101011000 0011010101010 00110101011000 0011010101010
100110100001100100001110101100110011000011001000011101011001
00110101011000 001101 1010100000110101011000 001101 10101000
101011001111100001000111010000101100111100001000111010000
1010100000010101010100001110101010000010101010100001
1000000111000101010110 000111000000111000101010110 0001
110010001101 01101000110010 001110010001101 01101000110010 00
100000011 0011010101110000011100000011 0011010101110000001
00110101011100000011010101010 100110101011100000011010101010
00010101010111000 0011010101000000110101011000 0110101010
0100001110101100111100001000100100001101011001111000010001
11101000010101000000110001001111010000101010000011100110
10011010000110010000111010110011001100100011001000011101011001
00110101011000 0011010101010 00110101011000 0011010101010
10011010000110010000111010110011001100100011001000011101011001
00110101011000 001101 101010000011010111000 001101 10101000
10101100111100001000111010000101100111100001000111010000
10101100000101010111000011101011000001010101011000011
1000000111000101010110 000111000000111000101010110 0001
1100100001101 01101000110010 001110010001101 01101000110010 00
100000011 0011010101110000011100000011 0011010101110000001
00110101011100000011010101010 100110101011100000011010101010
0001101010111000 0011010101000000110101011000 0011010101010
01000011101011001111000010001001000011101011001111000010001
1110100000101010000011100010111101000001010100000111000110
10111000000110101010000011011011100000011010101000001010
100110100001100100011101011001100110011001000011101011001
00110101011000 0011010101010 00110101011000 0011010101010
1001101000011001000011101011001110011001000011101011001
00110101011000 0011010101010 00110101011000 0011010101010
1001101000011001000011101011001110011001000011101011001
```

How to achieve Text Representation?



How to achieve Text Representation?

1. Bag of Words (BoW)



How to achieve Text Representation?

1. Bag of Words (BoW)
2. TF - IDF



How to achieve Text Representation?

1. Bag of Words (BoW)



Bag of Words (BoW)

- I like summers
- I love monsoon and love summers
- I love skiing in winters



Bag of Words (BoW)

- I like summers
- I love monsoon and love summers
- I love skiing in winters



“I”, “like”, “summers”, “love”, “monsoon”, “and”, “skiing”, “in”, “winters”

Bag of Words (BoW)

Text	I	like	summers	love	monsoon	and	skiing	in	winters
I like summers									
I love monsoon and love summers									
I love skiing in winters									

Bag of Words (BoW)

Text	I	like	summers	love	monsoon	and	skiing	in	winters
I like summers	1	1	1						
I love monsoon and love summers				1	1	1			
I love skiing in winters							1	1	

Bag of Words (BoW)

Text	I	like	summers	love	monsoon	and	skiing	in	winters
I like summers	1	1	1	0	0	0	0	0	0
I love monsoon and love summers	1	0	1	1	1	0	0	0	0
I love skiing in winters	1	0	0	1	0	0	1	1	0

Bag of Words (BoW)

Text	I	like	summers	love	monsoon	and	skiing	in	winters
I like summers	1	1	1	0	0	0	0	0	0
I love monsoon and love summers	1	0	1	1	1	1	0	0	0
I love skiing in winters	1	0	0	1	0	0	1	1	1

Bag of Words (BoW)

- I like summers = [1 1 1 0 0 0 0 0 0]
- I love monsoon and love summers = [1 0 1 1 1 1 0 0 0]
- I love skiing in winters = [1 0 0 1 0 0 1 1 1]



Bag of Words (BoW)

Text	I	like	summers	love	monsoon	and	skiing	in	winters
I like summers	1	1	1	0	0	0	0	0	0
I love monsoon and love summers	1	0	1	2	1	1	0	0	0
I love skiing in winters	1	0	0	1	0	0	1	1	1

Bag of Words (BoW)

- I like summers = [1 1 1 0 0 0 0 0 0]
- I love monsoon and love summers = [1 0 1 2 1 1 0 0 0]
- I love skiing in winters = [1 0 0 1 0 0 1 1 1]



Drawback of Bag of Words (BoW)

- Sparse matrix



Drawback of Bag of Words (BoW)

- Sparse matrix

Analytics

Text	I	like	summers	love	monsoon	and	skiing	in	winters
I like summers	1	1	1	0	0	0	0	0	0
I love monsoon and love summers	1	0	1	2	1	1	0	0	0
I love skiing in winters	1	0	0	1	0	0	1	1	1


Drawback of Bag of Words (BoW)

- Sparse matrix
- Word order not captured



Drawback of Bag of Words (BoW)

- Sparse matrix
- Word order not captured



Text	I	like	summers	love	monsoon	and	skiing	in	winters
I like summers	1	1	1	0	0	0	0	0	0
I love monsoon and love summers	1	0	1	2	1	1	0	0	0
I love skiing in winters	1	0	0	1	0	0	1	1	1

Reducing Vocabulary Size

- Lowercase words



Reducing Vocabulary Size

- Lowercase words
- Remove punctuations



Reducing Vocabulary Size

- Lowercase words
- Remove punctuations
- Lemmatization



Reducing Vocabulary Size

- Lowercase words
- Remove punctuations
- Lemmatization
- Keeping only the top n frequently occurring words

Analytics
Vidhya

Reducing Vocabulary Size

- Lowercase words
- Remove punctuations
- Lemmatization
- Keeping only the top n frequently occurring words
- And so on...

Analytics
Vidhya

N-gram Bag of Words



N-gram Bag of Words

I like summers



N-gram Bag of Words

I like summers

- 2-gram
 - I like
 - like summers



N-gram Bag of Words

I like summers

- 3-gram
 - I like summers



N-gram Bag of Words

- I like summers
- I love monsoon and love summers
- I love skiing in winters

2 grams:

I like, like summers, I love, love monsoon, monsoon and, and love, love summers, love skiing, skiing in, in winters

N-gram Bag of Words

Analytics

Text	I like	like summers	I love	love monsoon	monsoon and	and love	love summers	love skiing	skiing in	in winters
I like summers	1	1	1	0	0	0	0	0	0	0
I love monsoon and love summers	0	0	1	1	1	1	1	0	0	0
I love skiing in winters	0	0	1	0	0	0	0	1	1	1

How to achieve Text Representation?

1. Bag of Words (BoW)



How to achieve Text Representation?

2. TF - IDF



Term Frequency - Inverse Document Frequency (TF-IDF)

Term Frequency–Inverse Document Frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

Raw Term-Frequency (TF)

- Count of occurrence of a term in a document

$tf_{t,d}$ = Number of times, term t occurs in document d

Challenges: Raw Term-Frequency (TF)

- Relevance of a document increases with the term frequency



Log-normalized Term-Frequency (TF)

$$W_{t,d} = \begin{cases} 1 + \log_{10}(tf_{t,d}) & , \text{ if } tf_{t,d} > 0 \\ 0 & , \text{ Otherwise} \end{cases}$$

Log-normalized Term-Frequency (TF)

Text	I	like	summers	love	monsoon	and	skiing	in	winters
I like summers	$1+\log(1)$	$1+\log(1)$	$1+\log(1)$	0	0	0	0	0	0
I love monsoon and love summers	$1+\log(1)$	0	$1+\log(1)$	$1+\log(2)$	$1+\log(1)$	$1+\log(1)$	0	0	0
I love skiing in winters	$1+\log(1)$	0	0	$1+\log(1)$	0	0	$1+\log(1)$	$1+\log(1)$	$1+\log(1)$

Log-normalized Term-Frequency (TF)

Text	I	like	summers	love	monsoon	and	skiing	in	winters
I like summers	1	1	1	0	0	0	0	0	0
I love monsoon and love summers	1	0	1	1.30	1	1	0	0	0
I love skiing in winters	1	0	0	1	0	0	1	1	1

Inverse Document Frequency (IDF)

IDF is a measure of how important a term is.



Inverse Document Frequency (IDF)

$$\text{idf}_t = \log_{10} (N/\text{df}_t)$$

N is Number of documents in the dataset

df_t is Number of documents containing the terms t

Inverse Document Frequency (IDF)

Text	I	like	summers	love	monsoon	and	skiing	in	winters
IDF	$\log(3/3)$	$\log(3/1)$	$\log(3/2)$	$\log(3/2)$	$\log(3/1)$	$\log(3/1)$	$\log(3/1)$	$\log(3/1)$	$\log(3/1)$

Vidhaya

Inverse Document Frequency (IDF)

Text	I	like	summers	love	monsoon	and	skiing	in	winters
IDF	0	0.48	0.18	0.18	0.48	0.48	0.48	0.48	0.48

Vidhaya

TF-IDF score

$$W_{t,d} = \begin{cases} (1 + \log_{10}(tf_{t,d})) * \log_{10}(N/df_t) & , \text{ if } tf_{t,d} > 0 \\ 0 & , \text{ Otherwise} \end{cases}$$

TF-IDF score

Text	I	like	summers	love	monsoon	and	skiing	in	winters
I like summers	1	1	1	0	0	0	0	0	0
I love monsoon and love summers	1	0	1	1.30	1	1	0	0	0
I love skiing in winters	1	0	0	1	0	0	1	1	1

TF-IDF score

Text	I	like	summers	love	monsoon	and	skiing	in	winters
I like summers	1	1	1	0	0	0	0	0	0
I love monsoon and love summers	1	0	1	1.30	1	1	0	0	0
I love skiing in winters	1	0	0	1	0	0	1	1	1

Text	I	like	summers	love	monsoon	and	skiing	in	winters
IDF	1	0.48	0.18	0.18	0.48	0.48	0.48	0.48	0.48

TF-IDF score

Text	I	like	summers	love	monsoon	and	skiing	in	winters
I like summers	0	0.48	0.18	0	0	0	0	0	0
I love monsoon and summers	0	0	0.18	0.23	0.48	0.48	0	0	0
I love skiing in winters	0	0	0	0.18	0	0	0.48	0.48	0.48

Text	I	like	summers	love	monsoon	and	skiing	in	winters
I like summers	1	1	1	0	0	0	0	0	0
I love monsoon and love summers	1	0	1	2	1	1	0	0	0
I love skiing in winters	1	0	0	1	0	0	1	1	1

TF-IDF score

Text	I	like	summers	love	monsoon	and	skiing	in	winters
I like summers	0	0.48	0.18	0	0	0	0	0	0
I love monsoon and summers	0	0	0.18	0.23	0.48	0.48	0	0	0
I love skiing in winters	0	0	0	0.18	0	0	0.48	0.48	0.48

Text	I	like	summers	love	monsoon	and	skiing	in	winters
I like summers	1	1	1	0	0	0	0	0	0
I love monsoon and love summers	1	0	1	2	1	1	0	0	0
I love skiing in winters	1	0	0	1	0	0	1	1	1



Thank You