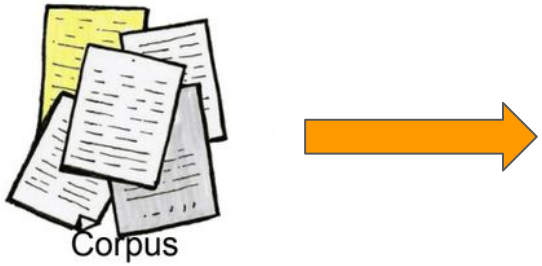
Text Representation

What is Text Representation?

What is Text Representation?

Converting text data from its raw form into numerical format.



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									

Analytics Vidhya

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									
I love skiing in winters									

Analytics Vidhya

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									
I love skiing in winters									

Analytics Vidhya

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

Analytics Vidhya

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

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 occuring words

Reducing Vocabulary Size

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

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

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 occurance of a term in a document

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

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)

$$idf_t = \log_{10} (N/df_t)$$

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)

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

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