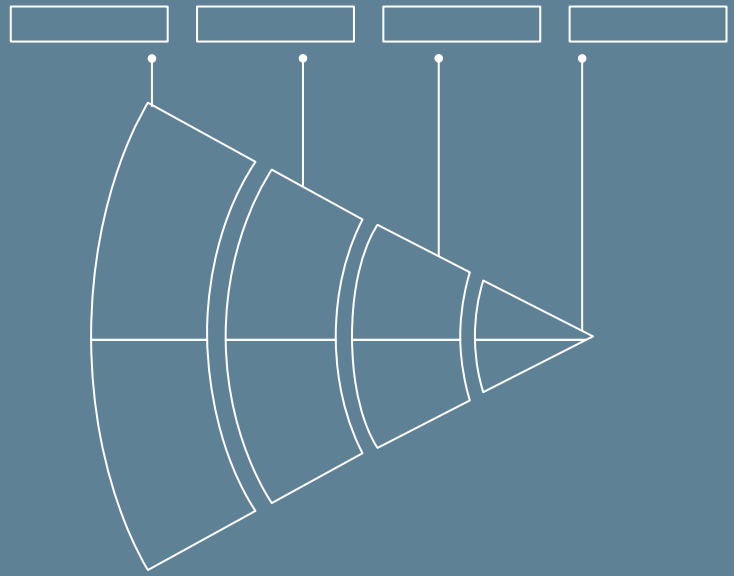
Teaching Computers to Read (Part 2)

Classification with Bag of Words and TF-IDF

NLP Tasks

Text data is difficult to work with. It requires a lot of

Pre Processing



Regular Expressions

Before

<h1>Donald J. Trump - @realDonaldTrump</h1>
Despite the constant negative press covfefe

After

Donald J. Trump
Despite the constant negative press covfefe

Tokenization

Before

The greatest wealth is to live content with little

After

['The',
'greatest',
'wealth',
'is',
'to',
'live',
'content',
'with',
'little']

Stop-Words

Before

Be kind, for everyone you meet is fighting a hard battle.

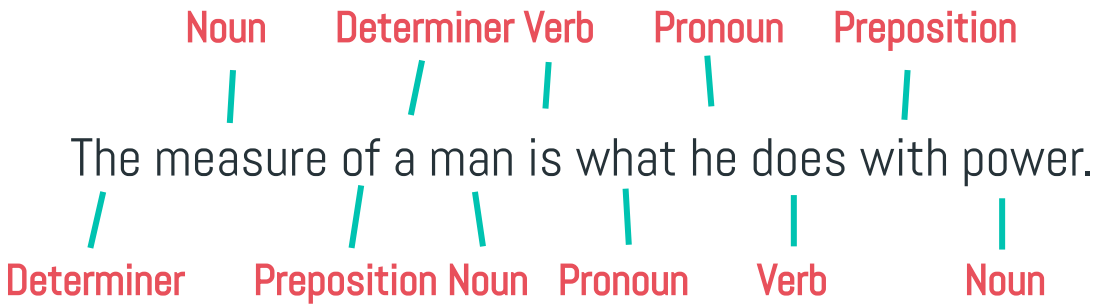
After

Be kind, everyone meet fighting hard battle.

Stemming

Before	After
Magical, Magician, Magically	Magic
Hunters, Hunting, Hunted	Hunt
Airline, Airliner, Airlines	Airlin

Part of Speech (PoS) Tagging



Lemmatization

Before	After
Was, is, are	Be
Meeting, meet, met (verb)	Meet
Meeting (noun)	Meeting

N-grams

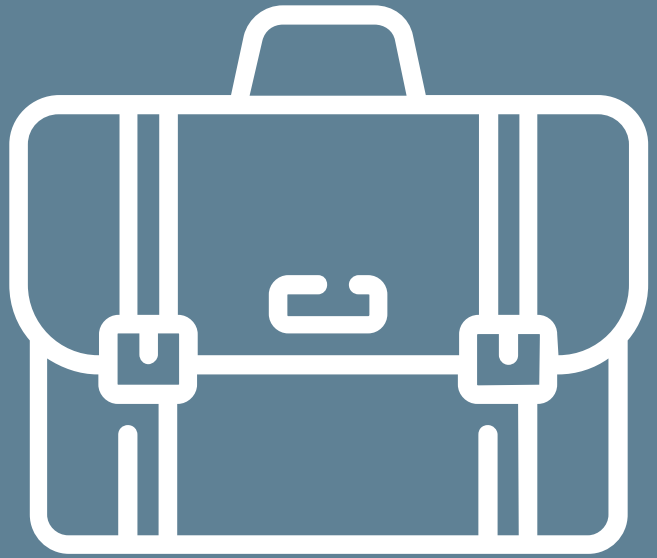
[Only[the[dead[have]seen]the end of war.

(Only, the)
(the, dead)
(dead, have)
(have, seen)
(seen, the)
(the, end)
(end, of)
(of, war)

Bag of Words

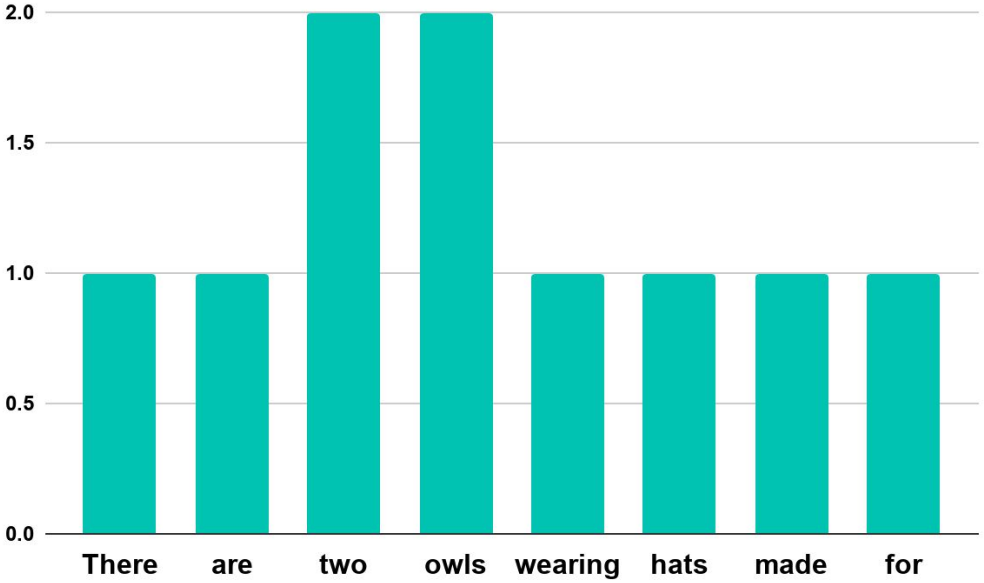
Sometimes called the **Unigram Model**, bag of words is interested in

Word Frequency



Bag of Words

There are two owls wearing two hats made for owls



Bag of Words

{ “There”: 1,

“Are”: 1,

“Two”: 2,

“Owls”: 2,

“Wearing”: 1,

“Hats”: 1,

“Made”: 1,

“For”: 1 }

Bag of Words

Sentence 1: There are two owls wearing two hats made for owls

Sentence 2: The owls were having great fun in their hats

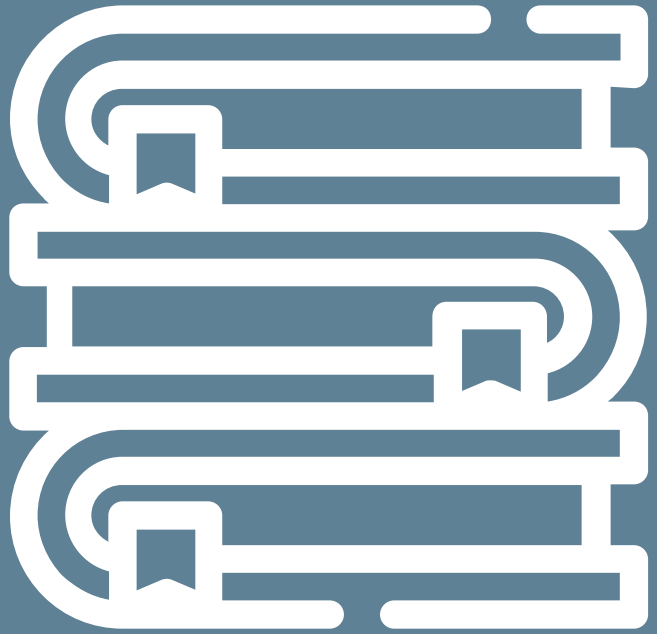
Sentence 3: Owls love wearing hats

Bag of Words

	There	are	two	owls	wearing	hats	made	for	the	were	having	great	fun	in	their	love
Sentence 1	1	1	2	2	1	1	1	1	0	0	0	0	0	0	0	0
Sentence 2	0	0	0	1	0	1	0	0	1	1	1	1	1	1	1	0
Sentence 3	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	1

TF – IDF

This model uses
Term Frequency
and
Inverse Document Frequency



TF-IDF

Term Frequency

How often each term (or word) occurs within a document

Document Frequency

How often each term (or word) appears across all documents in the corpus

Inverse Document Frequency

$$\text{Log}_e\left(\frac{\text{Number of documents}}{\text{Number of documents that include the term}}\right)$$

Combining TF and IDF

Term Frequency * Inverse Document Frequency

TF-IDF Example

	the warm water	the cold water	the murky water
the	1	1	1
warm	1.693	0	0
water	1	1	1
cold	0	1.693	0
murky	0	0	1.693

How do I get started?

Watson NLU (Natural Language Understanding)

IBM Developer - developer.ibm.com

Thanks!

Do you have any questions?

edmundshee@uk.ibm.com

@ukcloudman