Word2Vec

Introduction to the Word2Vec

BOW, TFIDF - Problems

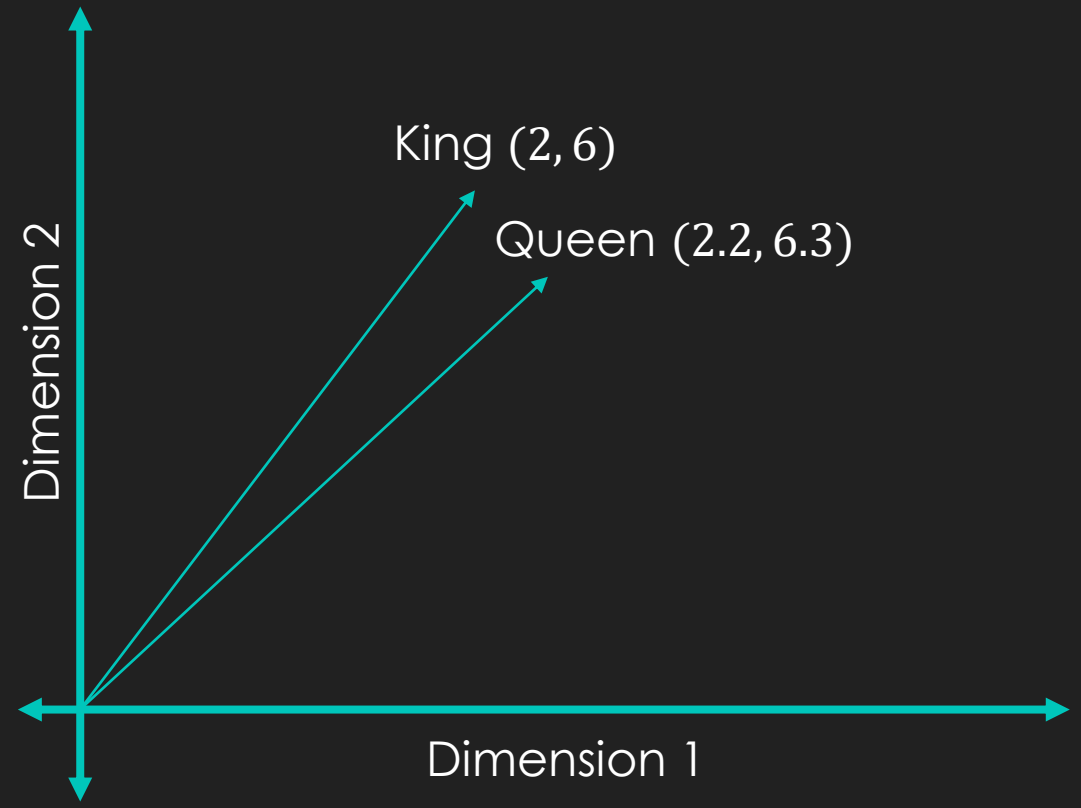
- Semantic information of the words is not stored. Even in TF-IDF model we only give more importance to the uncommon words.
- There's a chance of overfitting the model. Overfitting a scenario when model performs very well with your dataset but fails miserably when applied to any new dataset.

Word2Vec – The solution

- In this model, each word is represented as vector of 32 or more dimension instead of a single number.
- Relation between different words is preserved.

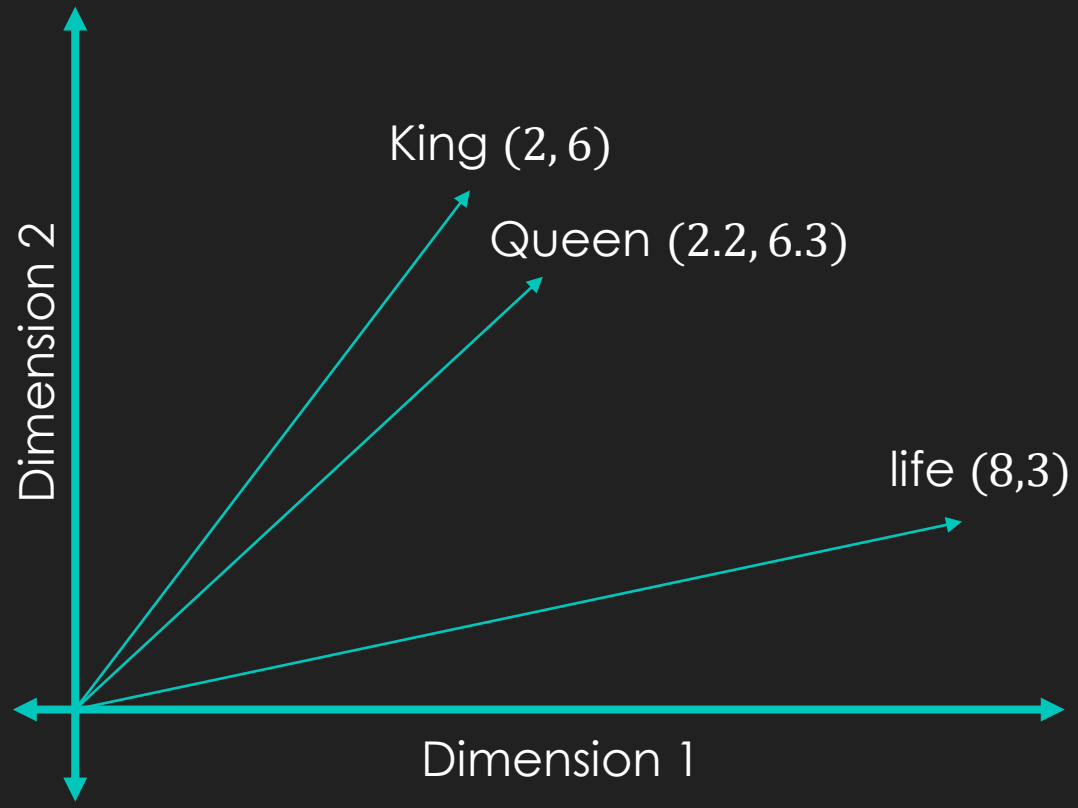
Word2Vec – Graphical Representation

2 dimensional



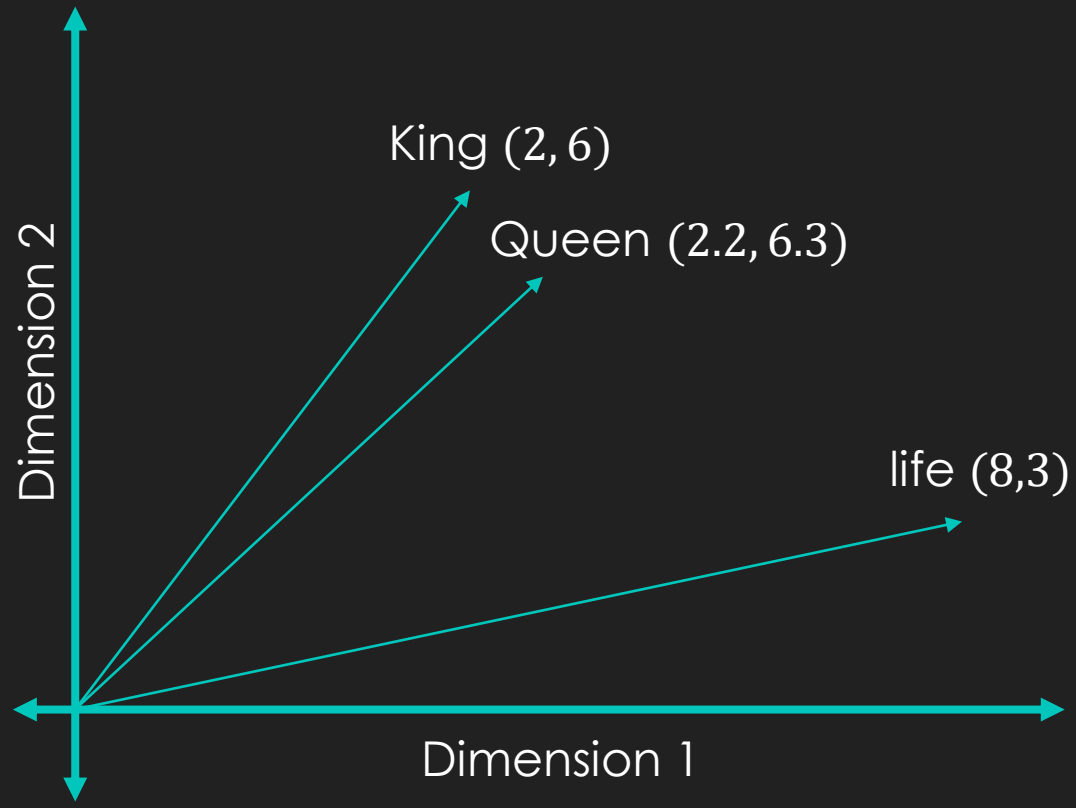
Word2Vec – Graphical Representation

2 dimensional



Word2Vec – Graphical Representation

2 dimensional



King – Man + Woman = Queen

Word2Vec – Extracting sentence meaning

"Sachin Tendulkar is the Roger Federer of Cricket"

Word2Vec – Extracting sentence meaning

"Sachin Tendulkar is the Roger Federer of Cricket"

Roger Federer – tennis + cricket = Sachin Tendulkar

Word2Vec – Steps to build the model

- Scrape through a **huge** dataset like the whole Wikipedia.
- Create a matrix with all the unique words in the dataset. The matrix represents the occurrence relation between the words.
- Split the matrix into two thin matrices.
- We have the model.

Word2Vec – Sample Dataset

going
to
today
i
am
it
is
rain
not
outside

"it is going to rain today"

"today i am not going outside"

"i am going to watch the season premiere"

Word2Vec – Steps to build the model

Words	going	to	today	i	am	it	is	rain	not	outside
going										
to										
today										
i										
am										
it										
is										
rain										
not										
outside										

Word2Vec – Sample Dataset

going
to
today
i
am
it
is
rain
not
outside

"it is going to rain today"

"today i am not going outside"

"i am going to watch the season premiere"

Word2Vec – Steps to build the model

Words	going	to	today	i	am	it	is	rain	not	outside
going	3									
to										
today										
i										
am										
it										
is										
rain										
not										
outside										

Word2Vec – Sample Dataset

going
to
today
i
am
it
is
rain
not
outside

"it is going to rain today"

"today i am not going outside"

"i am going to watch the season premiere"

Word2Vec – Word Matrix Formation

Words	going	to	today	i	am	it	is	rain	not	outside
going	3	2								
to										
today										
i										
am										
it										
is										
rain										
not										
outside										

Word2Vec – Sample Dataset

going
to
today
i
am
it
is
rain
not
outside

"it is going to rain today"

"today i am not going outside"

"i am going to watch the season premiere"

Word2Vec – Word Matrix Formation

Words	going	to	today	i	am	it	is	rain	not	outside
going	3	2	2							
to										
today										
i										
am										
it										
is										
rain										
not										
outside										

Word2Vec – Word Matrix Formation

Words	going	to	today	i	am	it	is	rain	not	outside
going	3	2	2	2	2	1	1	1	1	1
to										
today										
i										
am										
it										
is										
rain										
not										
outside										

Word2Vec – Word Matrix Formation

Words	going	to	today	i	am	it	is	rain	not	outside
going	3	2	2	2	2	1	1	1	1	1
to	2	2	1	1	1	1	1	1	0	0
today										
i										
am										
it										
is										
rain										
not										
outside										

Word2Vec – Word Matrix Formation

Words	going	to	today	i	am	it	is	rain	not	outside
going	3	2	2	2	2	1	1	1	1	1
to	2	2	1	1	1	1	1	1	0	0
today	2	1	2	1	1	1	1	1	1	1
i										
am										
it										
is										
rain										
not										
outside										

Word2Vec – Word Matrix Formation

Words	going	to	today	i	am	it	is	rain	not	outside
going	3	2	2	2	2	1	1	1	1	1
to	2	2	1	1	1	1	1	1	0	0
today	2	1	2	1	1	1	1	1	1	1
i	2	1	1	2	2	0	0	0	1	1
am										
it										
is										
rain										
not										
outside										

Word2Vec – Word Matrix Formation

Words	going	to	today	i	am	it	is	rain	not	outside
going	3	2	2	2	2	1	1	1	1	1
to	2	2	1	1	1	1	1	1	0	0
today	2	1	2	1	1	1	1	1	1	1
i	2	1	1	2	2	0	0	0	1	1
am	2	1	1	2	2	0	0	0	1	1
it										
is										
rain										
not										
outside										

Word2Vec – Word Matrix Formation

Words	going	to	today	i	am	it	is	rain	not	outside
going	3	2	2	2	2	1	1	1	1	1
to	2	2	1	1	1	1	1	1	0	0
today	2	1	2	1	1	1	1	1	1	1
i	2	1	1	2	2	0	0	0	1	1
am	2	1	1	2	2	0	0	0	1	1
it	1	1	1	0	0	1	1	1	0	0
is										
rain										
not										
outside										

Word2Vec – Word Matrix Formation

Words	going	to	today	i	am	it	is	rain	not	outside
going	3	2	2	2	2	1	1	1	1	1
to	2	2	1	1	1	1	1	1	0	0
today	2	1	2	1	1	1	1	1	1	1
i	2	1	1	2	2	0	0	0	1	1
am	2	1	1	2	2	0	0	0	1	1
it	1	1	1	0	0	1	1	1	0	0
is	1	1	1	0	0	1	1	1	0	0
rain										
not										
outside										

Word2Vec – Word Matrix Formation

Words	going	to	today	i	am	it	is	rain	not	outside
going	3	2	2	2	2	1	1	1	1	1
to	2	2	1	1	1	1	1	1	0	0
today	2	1	2	1	1	1	1	1	1	1
i	2	1	1	2	2	0	0	0	1	1
am	2	1	1	2	2	0	0	0	1	1
it	1	1	1	0	0	1	1	1	0	0
is	1	1	1	0	0	1	1	1	0	0
rain	1	1	1	0	0	1	1	1	0	0
not										
outside										

Word2Vec – Word Matrix Formation

Words	going	to	today	i	am	it	is	rain	not	outside
going	3	2	2	2	2	1	1	1	1	1
to	2	2	1	1	1	1	1	1	0	0
today	2	1	2	1	1	1	1	1	1	1
i	2	1	1	2	2	0	0	0	1	1
am	2	1	1	2	2	0	0	0	1	1
it	1	1	1	0	0	1	1	1	0	0
is	1	1	1	0	0	1	1	1	0	0
rain	1	1	1	0	0	1	1	1	0	0
not	1	0	1	1	1	0	0	0	1	1
outside										

Word2Vec – Word Matrix Formation

Words	going	to	today	i	am	it	is	rain	not	outside
going	3	2	2	2	2	1	1	1	1	1
to	2	2	1	1	1	1	1	1	0	0
today	2	1	2	1	1	1	1	1	1	1
i	2	1	1	2	2	0	0	0	1	1
am	2	1	1	2	2	0	0	0	1	1
it	1	1	1	0	0	1	1	1	0	0
is	1	1	1	0	0	1	1	1	0	0
rain	1	1	1	0	0	1	1	1	0	0
not	1	0	1	1	1	0	0	0	1	1
outside	1	0	1	1	1	0	0	0	1	1

Word2Vec – Steps to build the model

- Scrape through a **huge** dataset like the whole Wikipedia.
- Create a matrix with all the unique words in the dataset. The matrix represents the occurrence relation between the words.
- Split the matrix into two thin matrices.
- We have the model.

Word2Vec – Word Matrix Formation

Words	going	to	today	i	am	it	is	rain	not	outside
going	3	2	2	2	2	1	1	1	1	1
to	2	2	1	1	1	1	1	1	0	0
today	2	1	2	1	1	1	1	1	1	1
i	2	1	1	2	2	0	0	0	1	1
am	2	1	1	2	2	0	0	0	1	1
it	1	1	1	0	0	1	1	1	0	0
is	1	1	1	0	0	1	1	1	0	0
rain	1	1	1	0	0	1	1	1	0	0
not	1	0	1	1	1	0	0	0	1	1
outside	1	0	1	1	1	0	0	0	1	1

Word2Vec – Splitting into smaller matrices

Words	Dimension 1	Dimension 2
going		
to		
today		
i		
am		
it		
is		
rain		
not		
outside		

Words	going	to	today	i	am	it	is	rain	not	outside
Dimension 1										
Dimension 2										

Word2Vec – Word Matrix Formation

Words	going	to	today	i	am	it	is	rain	not	outside
going	3	2	2	2	2	1	1	1	1	1
to	2	2	1	1	1	1	1	1	0	0
today	2	1	2	1	1	1	1	1	1	1
i	2	1	1	2	2	0	0	0	1	1
am	2	1	1	2	2	0	0	0	1	1
it	1	1	1	0	0	1	1	1	0	0
is	1	1	1	0	0	1	1	1	0	0
rain	1	1	1	0	0	1	1	1	0	0
not	1	0	1	1	1	0	0	0	1	1
outside	1	0	1	1	1	0	0	0	1	1

Word2Vec – Splitting into smaller matrices

Words	Dimension 1	Dimension 2
going		
to		
today		
i		
am		
it		
is		
rain		
not		
outside		

Words	going	to	today	i	am	it	is	rain	not	outside
Dimension 1										
Dimension 2										

Word2Vec – Splitting into smaller matrices

Words	Dimension 1	Dimension 2
going		
to		
today		
i		
am		
it		
is		
rain		
not		
outside		

Words	going	to	today	i	am	it	is	rain	not	outside
Dimension 1										
Dimension 2										

A

Word2Vec – Splitting into smaller matrices

Words	Dimension 1	Dimension 2
going		
to		
today		
i		
am		
it		
is		
rain		
not		
outside		

Words	going	to	today	i	am	it	is	rain	not	outside
Dimension 1										
Dimension 2										

$A * A^T$

Word2Vec – Word Matrix Formation

Words	going	to	today	i	am	it	is	rain	not	outside
going	3	2	2	2	2	1	1	1	1	1
to	2	2	1	1	1	1	1	1	0	0
today	2	1	2	1	1	1	1	1	1	1
i	2	1	1	2	2	0	0	0	1	1
am	2	1	1	2	2	0	0	0	1	1
it	1	1	1	0	0	1	1	1	0	0
is	1	1	1	0	0	1	1	1	0	0
rain	1	1	1	0	0	1	1	1	0	0
not	1	0	1	1	1	0	0	0	1	1
outside	1	0	1	1	1	0	0	0	1	1

Word2Vec – Word Vectors

Words	Dimension 1	Dimension 2
going		
to		
today		
i		
am		
it		
is		
rain		
not		
outside		

Word2Vec – Word Vectors

Words	Dimension 1	Dimension 2
going	X_{1going}	X_{2going}
to	X_{1to}	X_{2to}
today	X_{1today}	X_{2today}
i	X_{1i}	X_{2i}
am	X_{1am}	X_{2am}
it	X_{1it}	X_{2it}
is	X_{1is}	X_{2is}
rain	X_{1rain}	X_{2rain}
not	X_{1not}	X_{2not}
outside	$X_{1outside}$	$X_{2outside}$

Word2Vec – Word Vectors

$$\text{going} = (X_{1going}, X_{2going}, \dots, X_{300going})$$

Word2Vec – Additional Read

Efficient Estimation of Word Representations in Vector Space
https://arxiv.org/pdf/1301.3781.pdf