Text Visualization

Creating TF-IDF Model

Bag Of Words – Problems

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

Bag Of Words – Problems

- All words have the same importance.
- No semantic information preserved.

TF-IDF Model – The Solution

- Some semantic information is preserved as uncommon words are given more importance than common words.

  Example: ‘She is beautiful’. Here ‘beautiful’ will have more importance than ‘She’ or ‘is’.

TF-IDF Model – Sample Corpus

"It is going to rain today"

"Today I am not going outside"

"I am going to watch the season premiere"

TF-IDF Model – Lower the sentences

"it is going to rain today"

"today i am not going outside"

"i am going to watch the season premiere"

TF-IDF Model – Tokenization

Sentence 1

it

is

going

to

rain

today

Sentence 2

today

i

am

not

going

outside

Sentence 3

i

am

going

to

watch

the

season

premiere

TF-IDF Model – Intuition

TF = Term Frequency

IDF = Inverse Document Frequency

TF-TDF = TF * IDF

TF-IDF Model – Term Frequency

Formula

$$\frac{(Number\ of\ occurrences\ of\ a\ word\ in\ a\ document)}{(Number\ of\ words\ in\ that\ document)}$$

TF-IDF Model – Term Frequency

$$\frac{(Number\ of\ occurrences\ of\ a\ word\ in\ a\ document)}{(Number\ of\ words\ in\ that\ document)}$$

TF-IDF Model – Term Frequency

$$\frac{(Number\ of\ occurrences\ of\ a\ word\ in\ a\ document)}{(Number\ of\ words\ in\ that\ document)}$$

"to be or not to be"

TF-IDF Model – Term Frequency

$$\frac{(Number\ of\ occurrences\ of\ a\ word\ in\ a\ document)}{(Number\ of\ words\ in\ that\ document)}$$

"to be or not to be"

to = $\frac{1+1}{}$

TF-IDF Model – Term Frequency

$$\frac{(Number\ of\ occurrences\ of\ a\ word\ in\ a\ document)}{(Number\ of\ words\ in\ that\ document)}$$

"to be or not to be"

to = $\frac{1+1}{6}$

TF-IDF Model – Term Frequency

$$\frac{(Number\ of\ occurrences\ of\ a\ word\ in\ a\ document)}{(Number\ of\ words\ in\ that\ document)}$$

"to be or not to be"

to = $\frac{1+1}{6}$

to = 0.33

TF-IDF Model – Term Frequency

$$\frac{(Number\ of\ occurrences\ of\ a\ word\ in\ a\ document)}{(Number\ of\ words\ in\ that\ document)}$$

"to be or not to be"

to = $\frac{1+1}{6}$

to = 0.33

be = 0.33

TF-IDF Model – Term Frequency

$$\frac{(Number\ of\ occurrences\ of\ a\ word\ in\ a\ document)}{(Number\ of\ words\ in\ that\ document)}$$

"to be or not to be"

to = $\frac{1+1}{6}$

to = 0.33

be = 0.33

or = 0.16

TF-IDF Model – Term Frequency

Sentence 1	Sentence 2	Sentence 3
it	today	i
is	i	am
going	am	going
to	not	to
rain	going	watch
today	outside	the
		season
		premiere

TF-IDF Model – Term Frequency

Word	Count
going	3
to	2
today	2
i	2
am	2
it	1
is	1
rain	1

TF-IDF Model – Term Frequency

Words/ Documents	Document 1	Document 2	Document 3
going			
to			
today			
i			
am			
it			
is			
rain			

TF-IDF Model – Term Frequency

Words/ Documents	Document 1	Document 2	Document 3
going	1/6		
to			
today			
i			
am			
it			
is			
rain			

Sentence 1

it

is

going

to

rain

today

TF-IDF Model – Term Frequency

Words/ Documents	Document 1	Document 2	Document 3
going	0.16		
to			
today			
i			
am			
it			
is			
rain			

Sentence 1

it

is

going

to

rain

today

TF-IDF Model – Term Frequency

Words/ Documents	Document 1	Document 2	Document 3
going	0.16		
to	0.16		
today			
i			
am			
it			
is			
rain			

Sentence 1

it

is

going

to

rain

today

TF-IDF Model – Term Frequency

Words/ Documents	Document 1	Document 2	Document 3
going	0.16		
to	0.16		
today	0.16		
i			
am			
it			
is			
rain			

Sentence 1

it

is

going

to

rain

today

TF-IDF Model – Term Frequency

Words/ Documents	Document 1	Document 2	Document 3
going	0.16		
to	0.16		
today	0.16		
i	0/6		
am			
it			
is			
rain			

Sentence 1

it

is

going

to

rain

today

TF-IDF Model – Term Frequency

Words/ Documents	Document 1	Document 2	Document 3
going	0.16		
to	0.16		
today	0.16		
i	0		
am			
it			
is			
rain			

Sentence 1

it

is

going

to

rain

today

TF-IDF Model – Term Frequency

Words/ Documents	Document 1	Document 2	Document 3
going	0.16		
to	0.16		
today	0.16		
i	0		
am	0		
it			
is			
rain			

Sentence 1

it

is

going

to

rain

today

TF-IDF Model – Term Frequency

Words/ Documents	Document 1	Document 2	Document 3
going	0.16		
to	0.16		
today	0.16		
i	0		
am	0		
it	0.16		
is			
rain			

Sentence 1

it

is

going

to

rain

today

TF-IDF Model – Term Frequency

Words/ Documents	Document 1	Document 2	Document 3
going	0.16		
to	0.16		
today	0.16		
i	0		
am	0		
it	0.16		
is	0.16		
rain			

Sentence 1

it

is

going

to

rain

today

TF-IDF Model – Term Frequency

Words/ Documents	Document 1	Document 2	Document 3
going	0.16		
to	0.16		
today	0.16		
i	0		
am	0		
it	0.16		
is	0.16		
rain	0.16		

Sentence 1

it

is

going

to

rain

today

TF-IDF Model – Term Frequency

Words/ Documents	Document 1	Document 2	Document 3
going	0.16		
to	0.16		
today	0.16		
i	0		
am	0		
it	0.16		
is	0.16		
rain	0.16		

Sentence 2

today

i

am

not

going

outside

TF-IDF Model – Term Frequency

Words/ Documents	Document 1	Document 2	Document 3
going	0.16	0.16	
to	0.16		
today	0.16		
i	0		
am	0		
it	0.16		
is	0.16		
rain	0.16		

Sentence 2

today

i

am

not

going

outside

TF-IDF Model – Term Frequency

Words/ Documents	Document 1	Document 2	Document 3
going	0.16	0.16	
to	0.16	0	
today	0.16		
i	0		
am	0		
it	0.16		
is	0.16		
rain	0.16		

Sentence 2

today

i

am

not

going

outside

TF-IDF Model – Term Frequency

Words/ Documents	Document 1	Document 2	Document 3
going	0.16	0.16	
to	0.16	0	
today	0.16	0.16	
i	0		
am	0		
it	0.16		
is	0.16		
rain	0.16		

Sentence 2

today

i

am

not

going

outside

TF-IDF Model – Term Frequency

Words/ Documents	Document 1	Document 2	Document 3
going	0.16	0.16	
to	0.16	0	
today	0.16	0.16	
i	0	0.16	
am	0		
it	0.16		
is	0.16		
rain	0.16		

Sentence 2

today

i

am

not

going

outside

TF-IDF Model – Term Frequency

Words/ Documents	Document 1	Document 2	Document 3
going	0.16	0.16	
to	0.16	0	
today	0.16	0.16	
i	0	0.16	
am	0	0.16	
it	0.16		
is	0.16		
rain	0.16		

Sentence 2

today

i

am

not

going

outside

TF-IDF Model – Term Frequency

Words/ Documents	Document 1	Document 2	Document 3
going	0.16	0.16	
to	0.16	0	
today	0.16	0.16	
i	0	0.16	
am	0	0.16	
it	0.16	0	
is	0.16		
rain	0.16		

Sentence 2

today

i

am

not

going

outside

TF-IDF Model – Term Frequency

Words/ Documents	Document 1	Document 2	Document 3
going	0.16	0.16	
to	0.16	0	
today	0.16	0.16	
i	0	0.16	
am	0	0.16	
it	0.16	0	
is	0.16	0	
rain	0.16		

Sentence 2

today

i

am

not

going

outside

TF-IDF Model – Term Frequency

Words/ Documents	Document 1	Document 2	Document 3
going	0.16	0.16	
to	0.16	0	
today	0.16	0.16	
i	0	0.16	
am	0	0.16	
it	0.16	0	
is	0.16	0	
rain	0.16	0	

Sentence 2

today

i

am

not

going

outside

TF-IDF Model – Term Frequency

Words/ Documents	Document 1	Document 2	Document 3
going	0.16	0.16	1/8
to	0.16	0	
today	0.16	0.16	
i	0	0.16	
am	0	0.16	
it	0.16	0	
is	0.16	0	
rain	0.16	0	

Sentence 3

i
am
going
to
watch
the
season
premiere

TF-IDF Model – Term Frequency

Words/ Documents	Document 1	Document 2	Document 3
going	0.16	0.16	0.12
to	0.16	0	
today	0.16	0.16	
i	0	0.16	
am	0	0.16	
it	0.16	0	
is	0.16	0	
rain	0.16	0	

Sentence 3

i
am
going
to
watch
the
season
premiere

TF-IDF Model – Term Frequency

Words/ Documents	Document 1	Document 2	Document 3
going	0.16	0.16	0.12
to	0.16	0	0.12
today	0.16	0.16	0
i	0	0.16	0.12
am	0	0.16	0.12
it	0.16	0	0
is	0.16	0	0
rain	0.16	0	0

Sentence 3

i
am
going
to
watch
the
season
premiere

TF-IDF Model – Intuition

TF = Term Frequency

IDF = Inverse Document Frequency

TF-TDF = TF * IDF

TF-IDF Model – Inverse Document Frequency

Formula

$$log(\frac{(Number\ of\ documents)}{(Number\ of\ documents\ containing\ word)})$$

TF-IDF Model – Inverse Document Frequency

$$log(\frac{(Number\ of\ documents)}{(Number\ of\ documents\ containing\ word)})$$

TF-IDF Model – Inverse Document Frequency

$$log(\frac{(Number\ of\ documents)}{(Number\ of\ documents\ containing\ word)})$$

"to be or not to be"
"i have to be"
"you got to be"

TF-IDF Model – Inverse Document Frequency

$$log(\frac{(Number\ of\ documents)}{(Number\ of\ documents\ containing\ word)})$$

"to be or not to be"
"i have to be"
"you got to be"

$$\text{to} = \log\left(\frac{3}{3}\right)$$

TF-IDF Model – Inverse Document Frequency

$$log(\frac{(Number\ of\ documents)}{(Number\ of\ documents\ containing\ word)})$$

"to be or not to be"
"i have to be"
"you got to be"

$$\text{to} = \log\left(\frac{3}{3}\right)$$

$$\text{to} = 0$$

TF-IDF Model – Inverse Document Frequency

$$log(\frac{(Number\ of\ documents)}{(Number\ of\ documents\ containing\ word)})$$

"to be or not to be"
"i have to be"
"you got to be"

to = $\log\left(\frac{3}{1}\right)$

to = 0

be = $\log(\frac{3}{3})$

TF-IDF Model – Inverse Document Frequency

$$log(\frac{(Number\ of\ documents)}{(Number\ of\ documents\ containing\ word)})$$

"to be or not to be"
"i have to be"
"you got to be"

to = $\log\left(\frac{3}{3}\right)$

to = 0

be = $\log(\frac{3}{3})$

be = 0

TF-IDF Model – Inverse Document Frequency

$$log(\frac{(Number\ of\ documents)}{(Number\ of\ documents\ containing\ word)})$$

"to be or not to be"
"i have to be"
"you got to be"

to = $\log\left(\frac{3}{3}\right)$

to = 0

be = $\log(\frac{3}{3})$

be = 0

have = $\log(\frac{3}{1})$

TF-IDF Model – Inverse Document Frequency

Sentence 1

it

is

going

to

rain

today

Sentence 2

today

i

am

not

going

outside

Sentence 3

i

am

going

to

watch

the

season

premiere

TF-IDF Model – Inverse Document Frequency

Word	Count
going	3
to	2
today	2
i	2
am	2
it	1
is	1
rain	1

TF-IDF Model – Inverse Document Frequency

Words	IDF Value
going	
to	
today	
i	
am	
It	
is	
rain	

TF-IDF Model – Inverse Document Frequency

Words	IDF Value
going	$\log(3/3)$
to	$\log(3/2)$
today	$\log(3/2)$
i	$\log(3/2)$
am	$\log(3/2)$
It	$\log(3/1)$
is	$\log(3/1)$
rain	$\log(3/1)$

"it is going to rain today"

"today i am not going outside"

"i am going to watch the season premiere"

TF-IDF Model – Inverse Document Frequency

Words	IDF Value
going	0
to	0.41
today	0.41
i	0.41
am	0.41
It	1.09
is	1.09
rain	1.09

"it is going to rain today"

"today i am not going outside"

"i am going to watch the season premiere"

TF-IDF Model – Building Model

Words	IDF Value
going	0
to	0.41
today	0.41
i	0.41
am	0.41
It	1.09
is	1.09
rain	1.09

Words/ Documents	Document 1	Document 2	Document 3
going	0.16	0.16	0.12
to	0.16	0	0.12
today	0.16	0.16	0
i	0	0.16	0.12
am	0	0.16	0.12
it	0.16	0	0
is	0.16	0	0
rain	0.16	0	0

TF-IDF Model – Building Model

Words/ Documents	going	to	today	i	am	it	is	rain
Document 1								
Document 2								
Document 3								

$$TFIDF(Word) = TF(Document, Word) * IDF(Word)$$

TF-IDF Model – Building Model

Words	IDF Value
going	0
to	0.41
today	0.41
i	0.41
am	0.41
It	1.09
is	1.09
rain	1.09

Words/ Documents	Document 1	Document 2	Document 3
going	0.16	0.16	0.12
to	0.16	0	0.12
today	0.16	0.16	0
i	0	0.16	0.12
am	0	0.16	0.12
it	0.16	0	0
is	0.16	0	0
rain	0.16	0	0

TF-IDF Model – Building Model

Words/ Documents	going	to	today	i	am	it	is	rain
Document 1	0*0.16 = 0							
Document 2								
Document 3								

$$TFIDF(Word) = TF(Document, Word) * IDF(Word)$$

TF-IDF Model – Building Model

Words	IDF Value
going	0
to	0.41
today	0.41
i	0.41
am	0.41
It	1.09
is	1.09
rain	1.09

Words/ Documents	Document 1	Document 2	Document 3
going	0.16	0.16	0.12
to	0.16	0	0.12
today	0.16	0.16	0
i	0	0.16	0.12
am	0	0.16	0.12
it	0.16	0	0
is	0.16	0	0
rain	0.16	0	0

TF-IDF Model – Building Model

Words/ Documents	going	to	today	i	am	it	is	rain
Document 1	0*0.16	0.41* 0.16						
Document 2								
Document 3								

$$TFIDF(Word) = TF(Document, Word) * IDF(Word)$$

TF-IDF Model – Building Model

Words	IDF Value
going	0
to	0.41
today	0.41
i	0.41
am	0.41
It	1.09
is	1.09
rain	1.09

Words/ Documents	Document 1	Document 2	Document 3
going	0.16	0.16	0.12
to	0.16	0	0.12
today	0.16	0.16	0
i	0	0.16	0.12
am	0	0.16	0.12
it	0.16	0	0
is	0.16	0	0
rain	0.16	0	0

TF-IDF Model – Building Model

Words/ Documents	going	to	today	i	am	it	is	rain
Document 1	0*0.16	0.41*0.16	0.41*0.16					
Document 2								
Document 3								

$$TFIDF(Word) = TF(Document, Word) * IDF(Word)$$

TF-IDF Model – Building Model

Words/ Documents	going	to	today	i	am	it	is	rain
Document 1	0	0.07	0.07	0	0	0.17	0.17	0.17
Document 2	0	0	0.07	0.07	0.07	0	0	0
Document 3	0	0.05	0	0.05	0.05	0	0	0

$$TFIDF(Word) = TF(Document, Word) * IDF(Word)$$