Rahul Shah (NLP Assignment - I)
9262240022

Q1 Zipf's law :-

It states that frequency of words at rank r is inversely proportional to rank of word in corpus.

$$f(r) = \frac{f(1)}{r}$$

$f(1)$ = Frequency of most frequent words (rank 1)

i) If rank of word is doubled. what should be expected frequency of word.

$r_2 = 2r_1$ $\quad r_1 = 1$

$$f(r_1) = \frac{f(1)}{r_1}$$

$$f(r_2) = \frac{f(1)}{r_2}$$

$$= \frac{f(1)}{2r_1} \Rightarrow \frac{f(1)}{2} \Rightarrow \frac{500}{2} = 250$$

ii) $N = 1000$ words; $f(1) = 50$; $f(10) = ?$

$$f(r_2) = \frac{f(1)}{r_2}$$

$$f(10) = \frac{50}{10} = 5$$

VISION

2) Min edit Distance

		S	A	T	U	R	D	A	Y
	S	0	1	2	3	4	5	6	7
	U	1	1	2	3	4	5	6	7
	N	2	2	2	3	4	5	6	7
	D	3	3	3	3	4	3	4	5
	A	4	3	4	5	5	4	3	4
	Y	5	4	4	6	6	5	4	3

We know formula; If $S[i-1] = T[j-1]$

$d(i,j) = d[i-1][j] + 1$ (Delete) [D]

$d[i][j-1] + 1$ (Insert) [I]

$d[i-1][j-1] + 1$ Substitute [S]

So;

		D	I	S
i)	S → S	0	0	0
	S → SA	0	1	0
	S → SAT	0	2	0
	S → SATU	0	3	0
	S → SATUR	0	4	0
	S → SATURD	0	5	0
	S → SATURDA	0	6	0
	S → SATURDAY	0	7	0

		D	I	S
ii)	SU → S	0	1	0
	SU → SA	0	0	1
	SU → SAT	0	1	1
	SU → SATU	0	2	1
	SU → SATUR	0	3	1
	SU → SATURD	0	4	1
	SU → SATURDA	0	5	1

	D	I	S
iii) SUN → S	2	0	0
SUN → SA	1	0	1
SUN → SAT	0	0	2
SUN → SATU	0	1	2
SUN → SATUR	0	2	2
SUN → SATURD	0	3	2
SUN → SATURDA	0	4	2

	D	I	S
iv) SUND → S	0	3	0
SUND → SA	2	0	1
SUND → SAT	1	0	2
SUND → SATU	2	2	3
SUND → SATUR	1	2	3
SUND → SATURD	0	2	1
SUND → SATURDA	0	3	1
SUND → SATURDAY	0	4	1

	D	I	S
v) SUNDA → S	4	0	0
SUNDA → SA	3	0	0
SUNDA → SAT	3	1	0
SUNDA → SATU	3	2	0
SUNDA → SATUR	2	2	1
SUNDA → SATURD	1	2	1
SUNDA → SATURDA	0	2	1
SUNDA → SATURDAY	0	3	1

	D	I	S
vi) SUNDAY → S	5	0	0
SUNDAY → SA	4	0	0
SUNDAY → SAT	3	0	1
SUNDAY → SATU	4	2	0
SUNDAY → SATUR	3	2	1
SUNDAY → SATURD	2	2	1
SUNDAY → SATURDA	1	2	1
SUNDAY → SATURDAY	0	2	1

Q.3

Given Sentence :-

S1 : "The cat sat on the mat"
S2 : "The dog jumped over the fence."
S3 : The cat and the dog played together."

Vocab

Tokenize :-

S1 : "the", "Cat", "sat", "on", "the", "mat"
S2 : "the", "dog", "jumped", "over", "the", "fence"
S3 : "The", "Cat", "and", "the", "dog", "played", "together" !

BOW :→

Vocab : [and, the, jumped, mat, on, played, cat, dog, together, sat, over, fence]

Words	doc 1	doc 2	doc 3
and			
the			
jumped			
mat			
On			
played			
cat			
dog			
together			
sat			

BOW : (Bag of words.

Word	Doc 1	Doc 2	Doc 3
and	0	0	1
cat	1	0	1
dog	0	1	1
fence	0	1	0
jumped	0	1	0
mat	1	0	0
on	1	0	0
over	0	1	0
played	0	0	1
sat	1	0	0
the	2	2	2
together	0	0	1

Doc 1 : [0 1 0 0 0 1 1 0 0 1 2 0]
Doc 2 : [0 0 1 1 1 0 0 1 0 0 2 0]
Doc 3 : [1 1 1 0 0 0 0 0 1 0 2 1]

Now;

ii) TF-IDF

We know that;

TF → Term frequency

$$TF = \frac{\text{No of times term } t \text{ in document } D}{\text{Total no of terms in Document } D}$$

word	D1	D2	D3
and	0	0	1/7
eat	1/6	0	1/7
dog	0	1/6	1/7
fence	0	1/6	0
jumped	0	1/6	0
mat	1/6	0	0
on	1/6	0	0
over	0	1/6	0
played	0	0	1/7
sat	1/6	0	0
the	2/6	2/6	2/7
together	0	0	1/7

IDF

$$IDF = \lg\left(\frac{N}{1+DF(t)}\right) = \frac{\text{Total no of documents}}{\text{No of documents containing term } t}$$

Document Frequency (DF):

Word	DF(t)	IDF	IDF	TF-IDF		
				D_1	D_2	D_3
and	1	$\lg(3/2)$	0.405	0	0	0.057
cat	2	$\lg(3/3)$	0	0	0	0
dog	2	$\lg(3/3)$	0	0	0	0
fence	1	$\lg(3/2)$	0.405	0	0.067	0
jumped	1	$\lg(3/2)$	0.405	0	0.067	0
mat	1	$\lg(3/2)$	0.405	0.067	0	0
on	1	$\lg(3/2)$	0.405	0.067	0	0
over	1	$\lg(3/2)$	0.405	0	0.067	0
played	1	$\lg(3/2)$	0.405	0	0	0.057
sat	1	$\lg(3/2)$	0.405	0.067	0	0
the	3	$\lg(3/3)$	-0.287	-0.095	-0.095	-0.082
together	1	$\lg(3/2)$	0.405	0	0	0.057

$D_1 = [0\ 0\ 0\ 0\ 0\ 1\ 1\ 0\ 0\ 1\ 0\ 0]$

$D_2 = [0\ 0\ 0\ 1\ 1\ 0\ 0\ 1\ 0\ 0\ 0\ 0]$

$D_3 = [1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1]$

Similarity is 0 for all combinations

D_1 & D_2 : 0

D_1 & D_3 : 0

D_2 & D_3 : 0

Similarity of BOW :-,

~~Dooor~~:

Magnitude of Doc 1 : $\sqrt{1+1+1+1+4} = \sqrt{8}$

Magnitude of Doc 2 : $\sqrt{1+1+1+1+4} = \sqrt{8}$

Magnitude of Doc 3 : $\sqrt{1+1+1+1+1+4} = \sqrt{9} = 3$

Doc 1 & Doc 2 :

Sim 1 : $\frac{4}{\sqrt{8}\sqrt{8}} = \frac{4}{8} = \frac{1}{2} = 0.5 \approx 0.5$

Doc 1 & Doc 3 :

Sim 2 : $\frac{(1+4)}{\sqrt{8}\sqrt{9}} = \frac{5}{\sqrt{8}*3} = 0.589 \approx 0.6$

Doc 2 & Doc 3 :

Sim 3 : $\frac{(1+4)}{\sqrt{8}\sqrt{9}} = \frac{5}{\sqrt{8}*3} = 0.589 \approx 0.6$

Q.3 Doc 1 : " The quick brown fox jumps over the lazy dog" 9
Doc 2 : "The lazy cat sleeps on the sunny mat". 8
Doc 3 : "The quick brown fox likes to jump high". 8

(without removing stopwords)

Tokenized :

Doc 1 : ['the', 'quick', brown, fox, jumps, over, the, lazy, dog]
Da 2 : ['the', 'lazy', 'cat', 'sleeps', 'on', 'the', 'sunny', 'mat']
DOC 3 : ['the', 'quick', 'brown', 'fox', 'likes', 'to', 'jump', 'high']

Vocab :

Vin O : [the, quick, brown, fox, jumps, over, lazy, dog, cat, sleeps, on, sunny, mat, likes, to, jump, high]
O : [brown, cat, dog, fox, high, jump/jumps, lazy, likes, mat, on, over, quick, sleeps, sunny, the, to]

Bag of words vector :

D_1 : [1 0 0 1 0 1 1 0 0 0 1 1 0 0 2 0]
D_2 : [0 1 0 0 0 0 1 0 1 1 0 0 1 1 2 0]
D_3 : [1 0 0 1 1 1 0 1 0 0 0 1 0 0 1 1]

Mag of D_1 : $\sqrt{1+1+1+1+1+1+1+4} = \sqrt{11}$
Mag of D_2 : $\sqrt{1+1+1+1+1+1+4} = \sqrt{10}$
Mag of D_3 : $\sqrt{1+1+1+1+1+1+1+1} = \sqrt{8}$

VISION

Doc 1 & Doc 2:

$$Sim\ 1 = \frac{5}{\sqrt{11}\sqrt{10}} = 0.476$$

Doc 1 & Doc 3:

$$Sim\ 2 = \frac{6}{\sqrt{11}\sqrt{8}} = 0.639$$

Doc 2 & Doc 3:

$$Sim\ 3 = \frac{2}{\sqrt{10}\sqrt{8}} = 0.2236$$

TF-IDF :

Words	TF			DF	IDF	IDF	TF-IDF		
	D_1	D_2	D_3				D_1	D_2	D_3
brown	1/9	0	1/8	2	$\lg(3/3)$	0	0	0	0
cat	0	1/8	0	1	$\lg(3/2)$	0.176	0	0.022	0
dog	1/9	0	0	1	$\lg(3/2)$	0.176	0.019	0	0
for	1/9	0	1/8	2	$\lg(3/3)$	0	0	0	0
high	0	0	1/8	1	$\lg(3/2)$	0.176	0	0	0.022
jumps	1/9	0	1/8	2	$\lg(3/3)$	0	0	0	0
lazy	1/9	1/8	0	2	$\lg(3/3)$	0	0	0	0
likes	0	0	1/8	1	$\lg(3/2)$	0.176	0	0	0.022
mat	0	1/8	0	1	$\lg(3/2)$	0.176	0	0.022	0
on	0	1/8	0	1	$\lg(3/2)$	0.176	0	0.022	0
over	1/9	0	0	1	$\lg(3/2)$	0.176	0.019	0	0
quick	1/9	0	1/8	2	$\lg(3/3)$	0	0	0	0
sleeps	0	1/8	0	1	$\lg(3/2)$	0.176	0	0.022	0
sunny	0	1/8	0	1	$\lg(3/2)$	0.176	0	0.022	0
the	2/9	2/8	1/8	3	$\lg(3/4)$	-0.124	-0.027	-0.031	-0.015
to	0	0	1/8	1	$\lg(3/2)$	0.176	0	0	0.022

Similarity,

D_1 & D_2

$$\text{Sim 1}: \frac{0.00041}{(0.038)(0.058)} = 0.186$$

D_1 & D_3

$$\text{Sim 2}: \frac{0.00041}{(0.038)(0.0409)} = 0.263$$

D_2 & D_3

$$\text{Sim 3}: \frac{0.00047}{(0.058)(0.0409)} = 0.198$$

~~$D_1 < D_2$~~ =

$|D_1| = 0.038$

$|D_2| = 0.058$

$|D_3| = 0.0409$