CHRIS SUNNY THALIYATH
QUIZ NLP

Data :

- doc1, doc2, doc 3, doc4, doc5
  these are CORPUS

- W1, W2, W3, W4, W5, W6, W7
  words

- W6, W7 — Stop words

Represent corpus

↳ 1 Binary Matrix
   ↳ where
      ↳ Row : document is represented
      ↳ column : Represtents words

"1 if word present in document, else its 0"

[TF-IDF(t,d) = TF(t,d) x IDF (t)]

IDF-TF (Inverse Docoment Frequency)

↳ Histogram of words in all docoments ie the count of different words

CBOW (continous Bag of words / Skip-Gram)

↳ We Create pair of words (Target words & Contexct words)

↳ Skip Gram : predicts the contexct based on current word & surrounding words

Binary Matrix

Document	W1	W2	W3	W4	W5	W6	W7
doc 1	1	1	1	0	0	0	1
doc 2	1	1	1	0	0	0	1
doc3	1	0	0	0	1	0	1
doc 4	0	0	0	1	1	1	0
do 5	0	0	0	0	0	1	0

$IDF(t) = \log (N / d (f(t)))$

N -: Total # of documents

df (t) : : # of documents with term t.

df (w 1) = 3

IDF (w1 = log (5/3)

TF - IDF (W1, doc1) = 1* IDF (W1)

=

df (w2) = 2

IDF (w 2) =

Calculations

	TF(doc 1	TF(doc2)	TF(doc 3	TF(doc4)	TF(doc5)	DF	IDF
W1	1	1	1	0	0	3	log(5/3)
W2	1	1	0	0	0	2	log(5/2)
W3	1	1	0	0	0	2	log(5/2
W4	0	0	0	1	0	1	log(5/1)
W5	0	0	1	1	0	2	log(5/2)
W6	0	0	0	1	1	2	log(5/2)
W7	1	1	1	0	0	3	log(5/3)

TF - IDF

Document	W1	W2	W3	W4	W5	W6	W7
doc 1	0.51	0.90	0.90	0	0	0	0.51
doc2	0.51	0.90	0.90	0	0	0	0.51
" 3	0.51	0	0	0	0.90	0	0.51
" 4	0	0	0	1.61	0.90	0.90	0
" 5	0	0	0	0	0	0.90	0

CBOW - Skip Gram

a) Tokenization : Break document into words
Remove stop words

(b) Windowing : window size (2-3 words)

CBOW (Continous Bag of words)
↳ input : Words within the window
↳ output : Target words
↳ Model : Surrounding Context

Skip Gram :

- input : target word
- output : Context words

Example with doc2 with window size 2

Document 2: (w1, w2, w3, w7)

Training Examples

(w1, w2) : Predict w2 given w1

(w2, w3) : "

(w2 w3) : "

(w2, w7) : "

(w3, w7) : "