NAME :- NAGESWARARAO CLASS ID - 18

① Stopwords :-

Stopwords are words which are not contain important significance to be used in search queries. usually these words are filtered out from search queries because they return vast amount of unnessary information.

Removal of Stopwords and punctuation :-

DOC1 :- After removing stopwords and punctuation the output will be as follows.

O/P :- researchers focus computational phenotyping produce disease prediction models machine learning statistical tools.

DOC2 :- researchers develop tools bayesian statistical information generate ~~com~~ casual models large complex phenotyping datasets

DOC3 :- researchers build computational information engine uses machine learning combine gene function gene interaction information distributed genetic data sources

N-gram

N-gram is a contiguous sequence of n items from a given sequence of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. The n-grams typically are collected from a text or speech corpus. When the items are words, n-grams may also be called shingles.

An n-gram of size 1 is referred to as a "unigram"; size 2 is a bigram, size 3 is a trigram. Larger sizes are sometimes referred to by the value of n in modern language. eg: four-gram "five gram" and so on.

Ex: After applying N-gram the result of doc is as follows:

The researchers will
researchers will focus
will focus on
focus on computational
on computational phenotyping
phenotyping and will
and will produce
.
.
.

The output for N=3 after removal stopwords for Doc 1 is :-

N=3

research focus computational
focus computational phenotype
computational phenotype produce
phenotype produce disease
disease prediction models
prediction models machine
:

Doc 2 :-

research develop tools
develop tools Bayesian
tools Bayesian statistical

} N=3

Doc 3 :- build computational information
computational information engine
information engine uses.

} N=3

:
:
2

List of items from the 3 documents as shown below

	D1	D2	D3	Count-in 3 docs
research	1	1	1	3
focus	1	0	0	1
process	1	0	0	1
models	1	1	0	2
machine	1	0	1	2
learning	1	0	1	2
statistical	1	1	0	2
develop	0	1	0	1
Information	0	1	1	2
generate	0	1	0	1
casual	0	1	0	1
complex	0	0	1	1
build	0	0	1	1
interaction	0	0	1	1
combine	0	0	1	1
tination				

b) TF :- It is the term frequency which measures how frequently a term occurs in a document since every document is different in length. It is possible that a term would appear many more times in a long document

$$TF(t) = \frac{\text{No of times term t appears in a document}}{\text{total no of terms in the document}}$$

IDF :- which measures how important a term is, while computing TF, terms are considered equally important.

$$IDF(t) = \log_e \left(\text{total no of documents} / \text{no of documents with term t} \right)$$

TF-IDF :- TF-IDF stands for term frequency inverse document frequency and the tf-IDF weight is a weight often used in information retrieval and text mining

Now we are Calculating TF-IDF valy for each term D:

for research

$TF = \frac{1}{12}$ $IDF = \log_e(3/3)$

$TF\text{-}IDF = \frac{1}{12} \times \log_e(3/3) = 0.0146$

for computation:

$TF = \frac{1}{12}$ $IDF = \log_e(3/1) = 0.477$

$TF\text{-}IDF = \frac{1}{12} \times 0.477 = 0.039$

for disease :

$TF = \frac{1}{12}$ $IDF = \log_e(3/1) = 3$

$TF\text{-}IDF = \frac{1}{12} \times 0.477 = 0.39$

for machine :

$TF = \frac{1}{12}$ $IDF = \log_e - 3/2 = 0.176$

$TF\text{-}IDF = \frac{1}{12} \times 0.176$

$= 0.0146$

for learning :

$TF = \frac{1}{12}$ $IDF = \log - 3/2 = 0.176$

$TFIDF = \frac{1}{12} \times 0.176 = 0.0146$