M3L3

Step1: Text → data

nltk → text processing

conv text into numbers into numerical/ vector

embedding

Pre-processing step

1. First we remove numbers from text.
2. Handling capitalization & punctuation

The = the → count for representation

, comes 10 times

cat (can't justify importance)

3. Stemming and lemmatizing

playing played raw form: play

Both have the same intent.

bring all the words to raw form.

———— × ———— × ———— × ———— × ————

Troubling clean text

lemmatize = False stemmer= True troubl

= True = False trouble

won't completely bring to raw form

Model will understand

1) Bag of words:

Broke paragraph into words.

Give a number to all number

50 distinct words (50 sized vector)

Ex: Trouble → 26 position

one-hot encoding

represent count of words.

higher term freq → higher imp

the (common word) → No sense to take this into consider.

Term frequency → this word in the doc ↗ more imp

TF - IDF

data → 10 diff doc what you all call as a dataset

a word occurs 7 times, 3 times

→ The number of times a term occurs in a document is called term frequency.

$T.F \times \frac{1}{\text{doc freq}} \rightarrow$ lower down importance

word that occurs in 3 documents

Review - dataset

↳ is it a positive or negative

we use BoW & TF-~~$~~IDF Text into numbers

featurization

62% acc 70% acc

cross-validation;

Spam dataset; 92% acc with BoW

98% acc with TF-IDF

till what extent we need to bring to a raw form

lemmatization brings to a raw form, better choice

↳ raw form.

what emotions are ~~being~~ reviewed

Spam detection : whether it is SPAM or HAM on Kaggle.

good 256 size 10, lakh

on how

n gram → 7 word Sentence

1 gram → 1 word at a time

2 gram → 2 words at a time.

T1