Day 2

✱ Preprocessing

① Tokenization → 1) Sentence Tokenizer
2) Word Tokenizer
   - ① Word Tokenizer
   - ② White space
3) Regex Tokenizer

② Normalization - lower case

③ stopwords Removal - ① language specific
② Domain Specific stopwords

④ Remove punctuation

⑤ Contraction mapping

I didn't like the movie → -ve
(not like move)

I like the movie → +ve

didn't = did not, never

don't = do not

⑥ stemming & lemmatization

stem (root word)

lemma (meaningful root word)

Stemming

- Aggressive Stemmer

Word net — Adj, adv

learning → lemmatization → learn

data → learning, swimming, walking

acting, act

act

Num1

record — 1 lakh

70 K

(7) Handling accented characters

à, â, ā

àpple → apple

bat → bat

(8) cleaning → regex

- "html. >> > a b

I like the

c

product

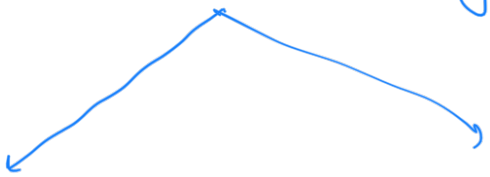
(9) Remove digits

✳ (10) Auto correction → ✳

(*) Feature Engineering

Text → Numerical - Word embedding

Word Embedding



(1) Frequency Based

→ 1) One hot encoding

→ 2) Count vectorizer

→ 3) TF IDF

(2) Prediction Based

→ 1) Word2Vec

→ 2) Doc2Vec

(1) Count vectorizer

Sent (1) ~~He~~ ~~are~~ learning nlp

Sent (2) NLP ~~is~~ ~~a~~ part ~~of~~ data science

Sent (3) Machine learning ~~and~~ deep learning ~~are~~ imp ~~in~~ data science

	learning,	nlp,	data,	science	part	machine	deep	imp	oo
①	1	1	0	0	0	0	0	0	
②	0	1	1	1	1	0	0	0	
③	2	0	1	1	0	1	1	1	
test	0	0	1	1	0	0	0	0	2

test ×

test = data science is interesting field

Drawbacks

(1) Curse Of Dimensionality

(2) It will not give meaning of word

(*) Creating vocabulary

1 lakh → 10 K

max_df = 0.95, 0.92 most frequent

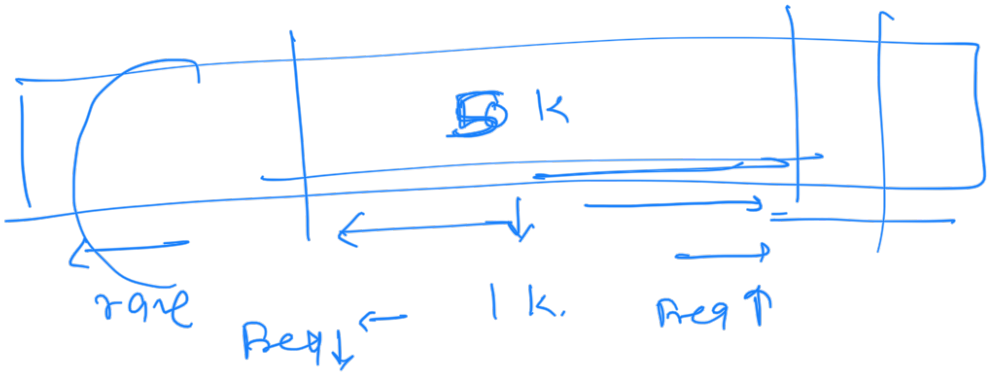
100 doc 95 > ×

1 K

min_df = 1, 5

100 doc → 2 K × rare words

= 1 K



OOV = Out of Vocabulary

CountVectorizer → train → columns → 100 col

→ OOV.

test → C.V. → OOV +3

3

100 words → 100 columns + 1 OOV

TF-IDF

TF = Term frequency

$$= \frac{\text{Count of word in a doc d}}{\text{total number of words in d}}$$

IDF = Inverse Documents Frequency

$$DF = \frac{\text{number of doc. containing term } t}{\text{total no. of doc}}$$

$$IDF = \log\left(\frac{1}{DF}\right)$$

Sent ① = "NLP part data scienc
NLP huge domain"

$$TF(nlp) = \frac{2}{7}$$

Sent ① = " learning nlp"

Sent ② = "NLP part data science"

Sent ③ = "machine learning deep learning
part data science" & nlp

	learning	nlp	part	data	science	machine	deep
①	$\frac{1}{2} \times \log(\frac{3}{\cdot})$	$\frac{1}{2} \times \log(\frac{3}{\cdot})$	0	0	0	0	0

②	0	$\frac{1}{4}\log(\frac{3}{2})$	$\frac{1}{4}\log(\frac{3}{2})$	$\frac{1}{4}\log(\frac{3}{2})$	$\frac{1}{4}\log(\frac{3}{2})$	0	0
③	$\frac{2}{7}\log(\frac{3}{2})$	0	$\frac{1}{7}\log(\frac{3}{2})$	$\frac{1}{7}\log(\frac{3}{2})$	$\frac{1}{7}\log(\frac{3}{2})$	$\frac{1}{7}\log(3)$	$\frac{1}{7}\log(3)$

Freq ↑ → weightage ↓

Drawbacks

① Curse Of dimensionality

100 → movie

↳ not

②