Day 6, 19th Dec 2023

NLP - by mentor Dr Rajat Dhamija, IIT DELHI + Harvard University Alumna

Recap ① spam & ham classifier
② BOW & TF-IDF implementation

Topic : Word Embedding

Limitation of TF-IDF

① Context info is missing
② Word order ✗
③ Sensitive to corpus size
④ biased to rare words

Key points of TF-IDF / Take home message

① useful tool to compare text → No.
② To find important words
↳ Keyword extraction
used in search engine to retrive relevant webpages
③ weightage to words → Advantage
④ context info missing → Disadvantage
⑤ There is a chance of Overfitting
⑥ Relationship b/w words are missing

Word Embedding

I love NLP
⇓
[0.8]
⇓
[0 ; 0.8 ; 0.9 ; 0.26 ; 0.13] 32 ⇝ Deep representation

Advantage

① Semantic info
② Context info
③ Order of word
④ relationship between words

① I was born in France. ..... ② I can speak French fluently

BOW: [0 1 0 1 0 0 1 0 ... 1]

TF-IDF: [0 0.8 ... 0.91 0 0 0.7]

[0.8 ; 0.3 ; ... ; 1] 32 France

[0.7 ; 0.4 ; ⋮] 32

Word embedding : Each word is represented as a vector of high dimension instead of a single number

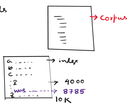
One hot encoding

V = 10,000 unique words

I was born in France. .....

[0 0 0 1 0 0 0] 10K — 8785

[0 0 0 0 0 1 0 0] 10K



Limitation to One hot encoding

① Creates high dimension vectors
if corpus size is big, # unique words / |V| = 10,00,000

[0 ; 0 ; ⋮] 10,00,000

② High dimension sparse matrix [0 0 1 0 0 ; 0 0 0 0 0]

③ Semantic information
④ Order of the word
⑤ Context " " "
} ✗

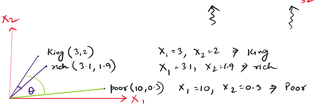
⑥ Time complexity
⑦ memory issue

word2vec → representing a word by a vector
smaller dimension

example

F	boy	Girl	Rich	poor	King	Queen
F_1 gender	+1	-1	0.3	0.02	+0.98	-0.98
F_2 age	0.3	0.9				
F_3 Bungalow			0.8	-0.8	0.83	
	32	32	32	32	32	32

gender

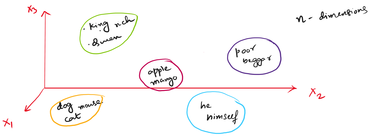


$X_1 = 3,\ X_2 = 2$ → King
$X_1 = 3.1,\ X_2 = 1.9$ → rich
$X_1 = 10,\ X_2 = 0.3$ → Poor

rich [0.8 ; 0.76 ; 0.33 ; ⋮] 32 — [0.81 ; 0.75 ; 0.35 ; ⋮] King

rich - King ≈ 0
King - rich ≈ 0
⇝ Cosine Similarity

rich [ ] 10K — [ ] poor 10K ⇝ rich - poor ≉ 0 ≈ ( )
King - poor ≈ 0 ✗ ✓



n - dimensions

Advantage

① Semantic information
② Context "
③ Sparse matrix ✗
↳ Deep representation
④ Lower dimension wrt One hot encoding
⑤ Relationship between words
⑥ Repeated words

I was born in France. .... Visa process for France..

[⋮ ✓ ⋮] [⋮ ✓ ⋮]

text
↓
□
↓
[✓ | ✓ |] vector
↓
[0 ; 1 ; 0.9 ;] BOW / TF-IDF
↓
[ ] [ ]
↓
One hot encoding
word2vec
Glove
} embedding

Stem
Lemma
stop