2024-06-12

12 June 2024 10:05

-> NLP where **Embeddings**

-> Prereq : Processing the text data.

[______] ~ Non-numeric Data

-> Math model ~ Numeric Data.

-> Tokenization : [CORPUS ~ 1000 lines]

[_/_/_/_/_/_/_] ~ **Parameters**

-> Preprocessing : **STOPWORDS**

[-> The propensity of action has consequences.]

Semantic ~ propensity action consequences

Connecting sentence grammatically.

-> TF-IDF does not have correlation index of words

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 1.2 \\ 22.3 & 0 & 0 & 4.5 & \\ \vdots & - & - & - & \end{bmatrix}$$

(5572, 8800)

~ Sparsity (Sparse)

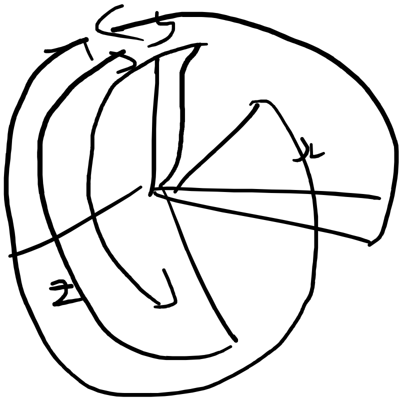
[Very big & has multiple zeroes]

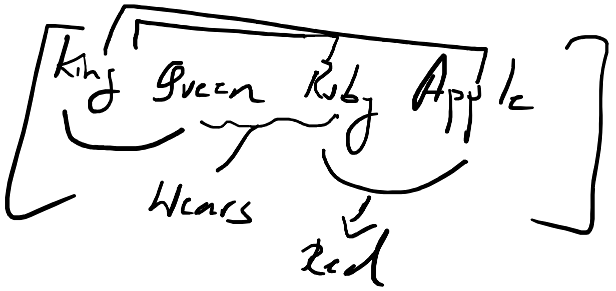
16 to 32 dimensions

Density of words

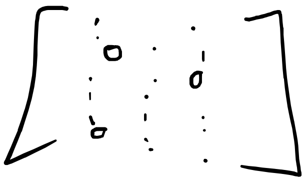
-> Word Embeddings (Deep Learning)

Density of words

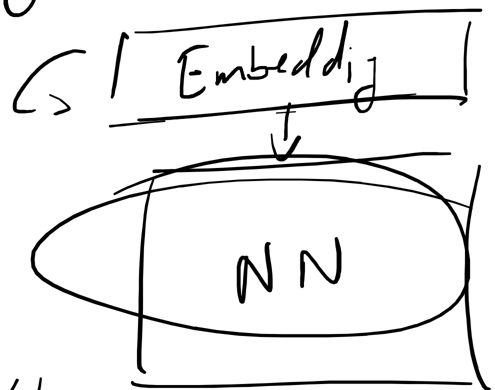




Probabilities



-> Train Embedding



~ Transfer learning

~ Training overhead (System)

Word2Vec

[GloVe]

500 words (tokens) vocab & 32 dimension

6 billion tokens

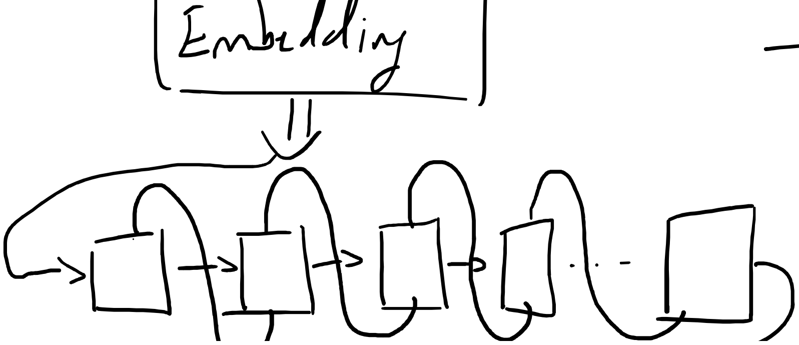
500

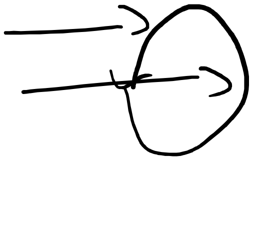
~ [6 0 0 0 0 0 0 0 0 0]

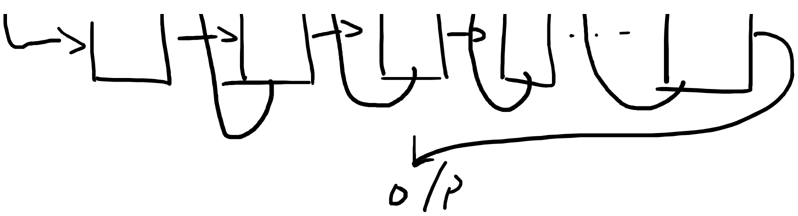
100 / 200 / 300 dimensions

-> LSTM

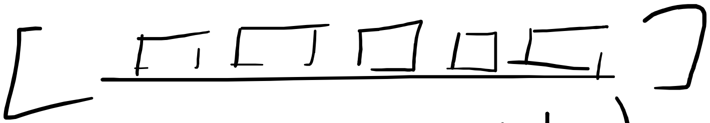
Bidirectional LSTM





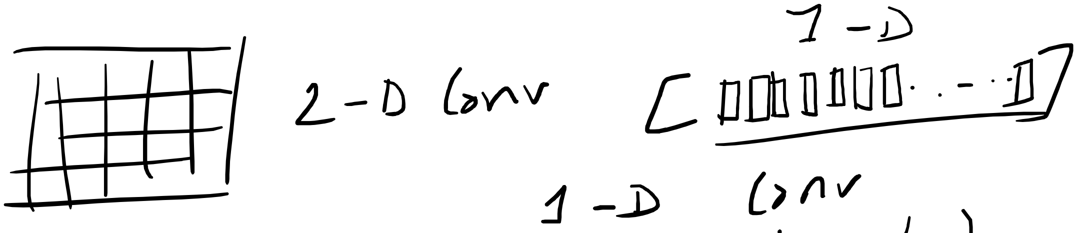


→ Creating Embedding on words

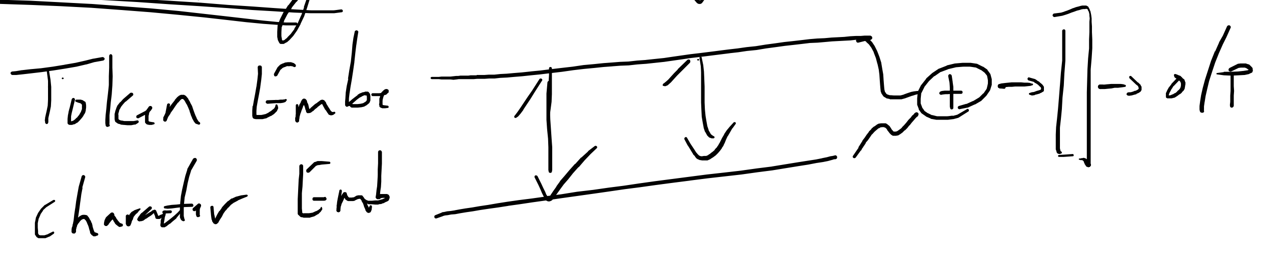


uncomplete (incomplete) ✓ Character Level Semantics

→ Character Level Embedding _ 1-D Convolutional Layer



→ Parallel Processing (Siamese style N/w)


Token Embe
Character Emb
+
o/p