

27/4/2024

Summary of NLP (as of 27/4)

1. Search Engine

World's Application application

Nutch → crawler → web → extract

BigData → Unstructured →

2. Preprocessing

1. lower case

2. punctuation

3. stop words

4. Numbers

5. Tags (HTML, XML)

6. Contraction I've, shouldn't

7. lemmatization I have shouldn't

8. Stemming slow - acce. - large

↳ fast - less accurate, large

9. Tokenization

↳ accented words â, ä,

Text to Numbers

Count Vectorizer	Bag of Words	TF-IDF	Word Embedding
<u>disadvantage</u> No Frequency within document NO Order NO Context NO Frequency within the corpus	<u>advantage</u> 1. Frequency within the document <u>disadvantage</u> NO Frequency within corpus	<u>advantage</u> 1. Frequency within corpus is also identity <u>disadvantage</u> 1. No context 2. No order	<u>advantage</u> ^{word sense, gloss, context} 1. Context is preserved <u>disadvantage</u> 1. No order preserved 2. Same word can have different meaning based on context

1. Hi <u>good</u> morning,	Hi	How	good	<u>morning</u>	your	beautiful
2. How <u>is</u> your morning	1	1	1	1		
3. <u>Morning</u> is beautiful				1		1

I went to HDFC Bank.
 I went to Rena Bank.

I don't have any interest buying loan
 The Interest rate is very high to get loan

Unsupervised → Text Clustering - k-means - bucket the similar document

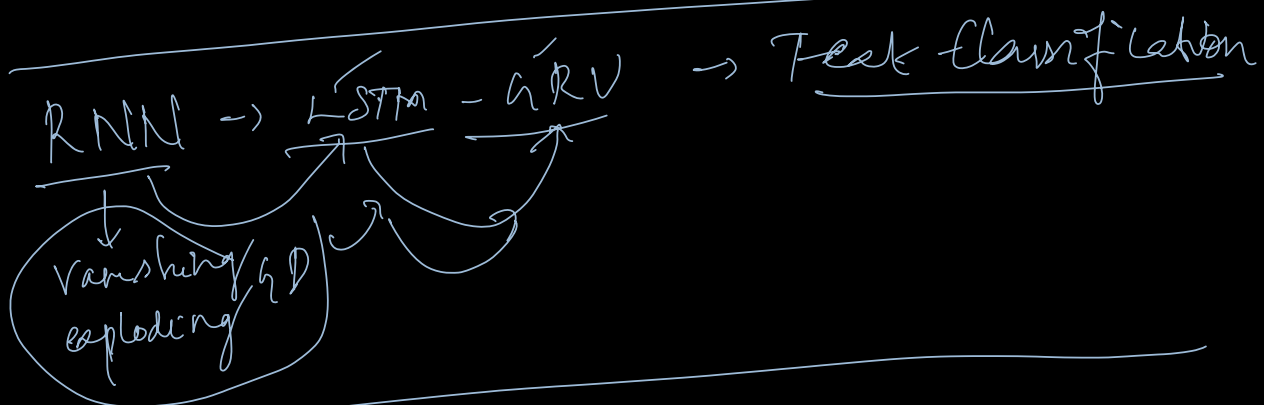
Cosine Similarity : Recommendation engine
Content

Sentiment Analysis - lexicon model -
Unsupervised textblob, Afinn, Vader

Text Classification
Bag of Words
embedding

logistic
machine
learning

Deep learning

RNN \rightarrow LSTM - GRU \rightarrow Text Classification


Seq2Seq Model \rightarrow encoder - decoder
Machine Translation

Image Captioning