27/4/2024

Summary of NLP (as of 27/4)

1. Search Engine
   World's Application application

Nutch → crawler → web → extract

BigData → Unstructured →

2. Preproccessing {

   1. lower case
   2. punctuation
   3. stop words
   4. Numbers
   5. Tags (HTML, XML)
   6. Contraction I 're, shouldn't
      I have Shouldnot
   7. lemmatatization → slow – acce. – large
   8. Stemming
      ↳ fast – less accurate, large
   9. Tokenzation
   10. accents words â, ä,

Text to Numbers

Count Vectorizer	Bag of Words	TF-IDF	word embedding word2vec glove, Fasttext
disadvantage	advantage	advantage	advantage
No Frequency within document NO Order NO Context NO Frequent within the Corpus	1. Frequency within the document	1. Frequency with corpus is also identify	1. Context is preserved
	disadvantage No Frequency with Corpus	disadvantage 1. No context 2. No order	disadvantage 1. No order preserved 2. Same word can have different meaning based on Context

1. Hi good morning,
2. How is your morning
3. Morning is beautiful

Hi	How	good	morning	your	beautiful
1		1	1 ✓		
	1		1 ✓		
			1 ✓		1

I went to HDFC Bank.
I went to River Bank

I dont have any intrest buying loan
The Intrest 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 -> LSTM - GRU -> Text Classification

vanishing/exploding GD

Seq2Seq Model -> encoder - Decoder

Machine Translation

Image Captioning