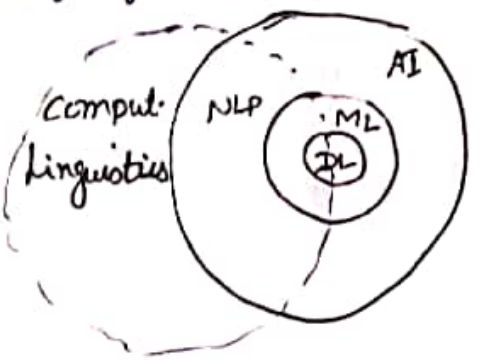
NLP (NATURAL LANGUAGE PROCESSING)

part of AI where computers analyse, understand and derive meaning meaningful info from Human language



Libraries → NTLK (Natural Lang Toolkit)
- ~~Scotty~~
- Spacy

Application
- Automatic Summerisation
- Sentiment Analysis
- Speech Recognition
- Topic Segmentation
- Named Entity Recognition

NLP → NLU (Natural Lang Understanding)
- Lexical Ambiguity
- Syntatical Ambiguity
- Referential Ambiguity

NLP → NLG (—,,—,,— Generation)
- Text Planning
- Sentence planning
- Text Realisation

Lexical Ambiguity

Ambiguity of a single word

She bagged two silver medals.
She made a silver speech.
His wavy hair had silvered.

silver → noun, adj, verb

Syntactical Ambiguity

Two or more possible meanings

Eg: The chicken is ready to eat.

- chicken bird is ready to eat
- chicken dish prepared/cooked?

Referential Ambiguity: Uses have many ref.
John met Tom & Harry. They went to garden.
John went with Tom / Harry or both to garden?

Corpus (collection of text documents) → Docs → Para → Sentences → Tokens

Tokenisation - seperating piece of text into smaller units called tokens

paragraph → sentences (nltk.sent_tokenize(paragraph)) → word (nltk.word_tokenize(paragraph))

word → vector (numerical repres)

Stemming & Lemmatization - used to analyse the meaning of word behind it

Stemming → produce intermediate representation of word
→ reduce word to their word stem

history, historical → histori

finally, final, finalized → fina

going, goes, gone → go

Lemmatization

history → history
historical → history

finally → final
final → final
finalised → final

Converts words to word that is understood by human

Stopwords - stopwords.words('english')

Eg i, me, myself, in, out, your, yours, wouldn't etc.

Bag of Words

Sent1: He is a good boy
Sent 2: She is a good girl
Sent3: Boys & girls are good

stop keywords → Lower Sentence (except country name)

Sent1: good boy
Sent2: good girl
Sent3: Boy girl good

Words	Freq
Good	3
Boy	3
girl	2

Bag of Words converts this table to vector.

	good f_1	boy f_2	girl f_3	f_{n-1}	o/p
S1	1	1	0	—	1
S2	1	0	1	—	1
S3	1	1	1	—	1

Binary Bag of Words (Values 1, or 0)

f_1, f_2, f_3 are independent variables & o/p data will be used to train ML/DL model

Disadv

(i) "Good" & "Boy" have equal representation. Semantics of these words are almost, we aren't able to derive which word is more impt. (good >> boy)

Sol^n → TF IDF (term freq & inverse document freq)
→ Word2vec (huge dataset)

TF - IDF = TF * IDF

$$TF = \frac{\text{No of representation of words in a sentence}}{\text{No of words in sentences}}$$

$$IDF = \log\left(\frac{\text{No of sentences}}{\text{No of sentences containing words}}\right)$$

sent 1 → good boy
sent 2 → good girl
sent 3 → boy girl good

Histogram or Freq Table →

Words	freq
good	3
boy	2
girl	2

↓ tfidf vectors

	TF		
	Sent 1	Sent 2	Sent 3
good	1/2	1/2	1/3
boy	1/2	0	1/3
girl	0	1/2	1/3

IDF	
Words	IDF
good	$\log 3/3 = 0$
boy	$\log(3/2)$
girl	$\log(3/2)$

	f_1	f_2	f_3	o/p
	good	boy	girl	
Sent 1	0	$(\log 3/2)\frac{1}{2}$	0	
Sent 2	0	0	$\frac{1}{2}(\log 3/2)$	
Sent 3	0	$1/2(\log 3/2)$	$1/2(\log 3/2)$	

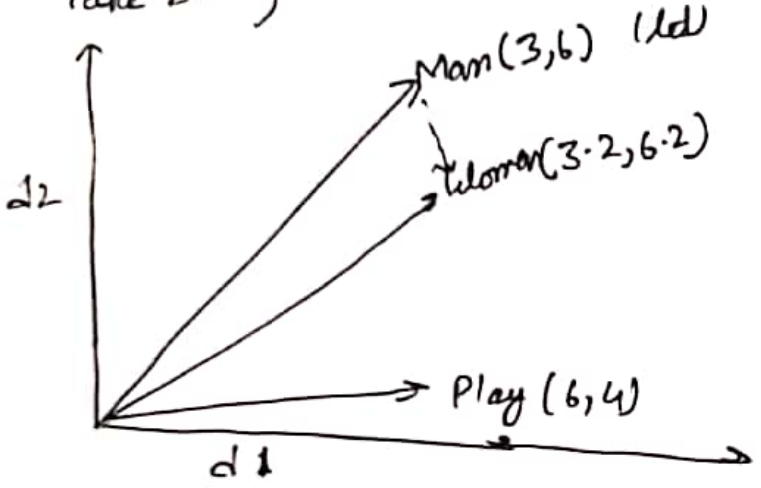
definite imp given to good, boy, girl, trying to bring in semantic meaning

Word2 Vec

- Both BoW & Tf-Idf semantic info is not stored properly (order of word)
  - Tf-Idf gives impt to uncommon words
  - Chance of overfitting

- Each word is represented as a vector 32 or 32+ dimension & not a single word.
- Hence semantic info & relation b/w diff^nt word is preserved

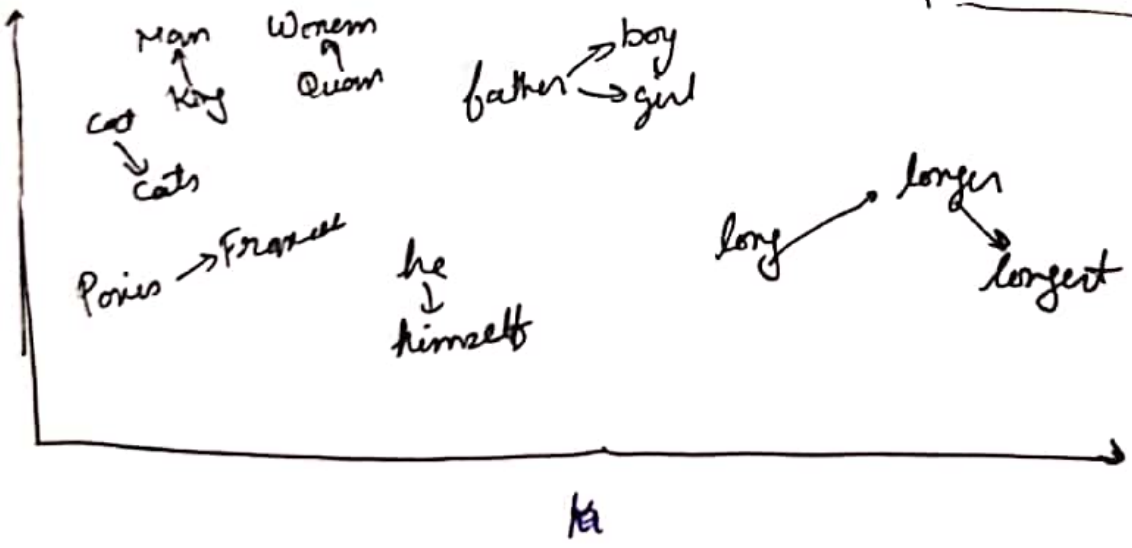
Take 2-D eg



- Men & Women are similar words, related with semantics
  ∴ distance b/w them is small

opr can be done

King − Man + Women = Queen



tokenize → histogram → take most freq words

↓

matrix of all unique words