Date ..../..../.....

Natural Language Processing

THE GOAL :- The ability to process & harness information from a large corpus of text with very little manual intervention

Why is it hard ?
→ idioms, metaphors, sarcasm, double -ves.
→ mixing of visual cues and ambiguous in nature
→ Interpretation depends on real world, common sense and contextual Knowledge.

Probablity and NLP :-
→ we can make informed decision
→ we can get a quantitative description or chances or likelihood associated with various outcomes.
→ predict next word in a sentence
→ probability of a sentence

Vector Space models
Let us consider that the words in a corpus are considered as linearly independent basis vectors.
If a corpus contains |V| words which are linearly independent, then every word represents a continuous vector space R. Each word takes an independent axis which is orthogonal to other words/axis.

Date ..../..../.....

So, If you have 7079 words in a corpus, you will have 7079 axes since each one will be independent of each other

We need to start aligning vectors in a way that if they are very close, they can combine them as one axis.

Creation of Semantically Created Vectors
→ Identify a model that enumerates the relationship between terms & documents
→ Identify a model that tries to put similar items closer to each other in some space or structure.
→ A model that discovers / uncovers the semantic similarity b/w words and documents.
→ Develop a distributed word vectors or dense vectors that captures the linear combination of word vectors in the transformed domain.

Word Embedding
→ Process each word in a vocabulary of words to obtain a respective numeric representation of each word in the vocabulary.
→ Reflect semantic similarities, Syntactic similarities or both, between words they represent.
→ Map each of the plurality of words to a respective vector and output a single merged vector that is a combination of resp vectors

Spiral

Date ..../..../.....

Sequence Learning

Sequence Learning is the study of machine learning algorithms designed for applications that require sequential data or temporal data.

Machine Translation

→ Approach 1 :- have a sentence in english look up French word for it.
→ Approach 2 :- Translation by analogy
→ Approach 3 :- Creating a syntax structure for ex :- english has a structure in terms of forming a sentence. Then there is an equivalent structure in other language to which you want to translate into. When the syntactic structure is similar do some translation.
→ Parallel Corpora is a sample of sentences in the target languge.

Summary :-

1] Gather statistical info about corpus, words
2] Understanding words from context or context words from a word.
3] Learn to encode the contextual info about a word
4] Predict the next word based on context
5] Learn to encode a sentence - understanding the context of a sentence.
6] Predict how likely a new sentence could be valid sentence.
7] Learn to automatically translate.

Spiral

Date ..../..../.....

Preprocessing :-

Preprocessing consists of (a) tokenization (b) normalization (c) substitution.

→ case folding → white space, newline, tabs
→ Stemming → removing contraction
→ Lemmatization → Remove scripts, form variables
→ Tokenization
→ remove misspellings → Punctuations