**Assignment\_2**

1. What are Corpora?

**Ans: The corpora are the translations of each other. For example, a novel and its translation or a translation memory of a CAT tool could be used to build a parallel corpus. Both languages need to be aligned, i.e. corresponding segments, usually sentences or paragraphs, need to be matched.**

1. What are Tokens?

**Ans: Tokens are the building blocks of Natural Language. Tokenization is a way of separating a piece of text into smaller units called tokens. Here, tokens can be either words, characters, or subwords. Hence, tokenization can be broadly classified into 3 types – word, character, and subword (n-gram characters) tokenization**

1. What are Unigrams, Bigrams, Trigrams?

Ans:

1. How to generate n-grams from text?

**Ans: def generate\_N\_grams(text,ngram=1):**

**words=[word for word in text.split(" ") if word not in set(stopwords.words('english'))]**

**print("Sentence after removing stopwords:",words)**

**temp=zip(\*[words[i:] for i in range(0,ngram)])**

**ans=[' '.join(ngram) for ngram in temp]**

**return ans**

1. Explain Lemmatization

**Ans: Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma**

1. Explain Stemming

**Ans: Stemming is a rule-based approach because it slices the inflected words from prefix or suffix as per the need using a set of commonly underused prefix and suffix, like “-ing”, “-ed”, “-es”, “-pre”, etc. It results in a word that is actually not a word**.

1. Explain Part-of-speech (POS) tagging

**Ans:  It is a process of converting a sentence to forms – list of words, list of tuples (where each tuple is having a form (word, tag)). The tag in case of is a part-of-speech tag, and signifies whether the word is a noun, adjective, verb, and so on.**

1. Explain Chunking or shallow parsing

**Ans: Shallow parsing (also chunking or light parsing) is an analysis of a sentence which first identifies constituent parts of sentences (nouns, verbs, adjectives, etc.) and then links them to higher order units that have discrete grammatical meanings (noun groups or phrases, verb groups, etc.)**

1. Explain Noun Phrase (NP) chunking

**Ans: Chunking is defined as the process of natural language processing used to identify parts of speech and short phrases present in a given sentence.**

1. Explain Named Entity Recognition

**Ans: Named Entity Recognition is one of the key entity detection methods in NLP. 2. Named entity recognition is** **a natural language processing technique that can automatically scan entire articles and pull out some fundamental entities in a text and classify them into predefined categories.**