**Assignment**

**Text Processing in NLP**

**Name – Sahil Sudhir Bhandari**

Natural Language Processing (NLP) involves the interaction between computers and human language. Text processing is a fundamental component of NLP, focusing on extracting valuable information from textual data. This documentation explores key aspects of text processing in NLP.

1. Tokenization

Tokenization is the process of breaking down text into smaller units, often words or phrases referred to as tokens. In NLP, tokenization is crucial for various tasks, such as text analysis and feature extraction.

Example:

from nltk.tokenize import word\_tokenize

text = "Tokenization is a key step in NLP."

tokens = word\_tokenize(text)

print(tokens)

Output:

['Tokenization', 'is', 'a', 'key', 'step', 'in', 'NLP', '.']

2. Stopword Removal

Stopwords are common words that often do not contribute significant meaning to a sentence. Removing stopwords is essential to focus on meaningful content.

Example:

from nltk.corpus import stopwords

from nltk.tokenize import word\_tokenize

text = "Removing stopwords is crucial for text analysis."

stop\_words = set(stopwords.words('english'))

words = word\_tokenize(text)

filtered\_words = [word for word in words if word.lower() not in stop\_words]

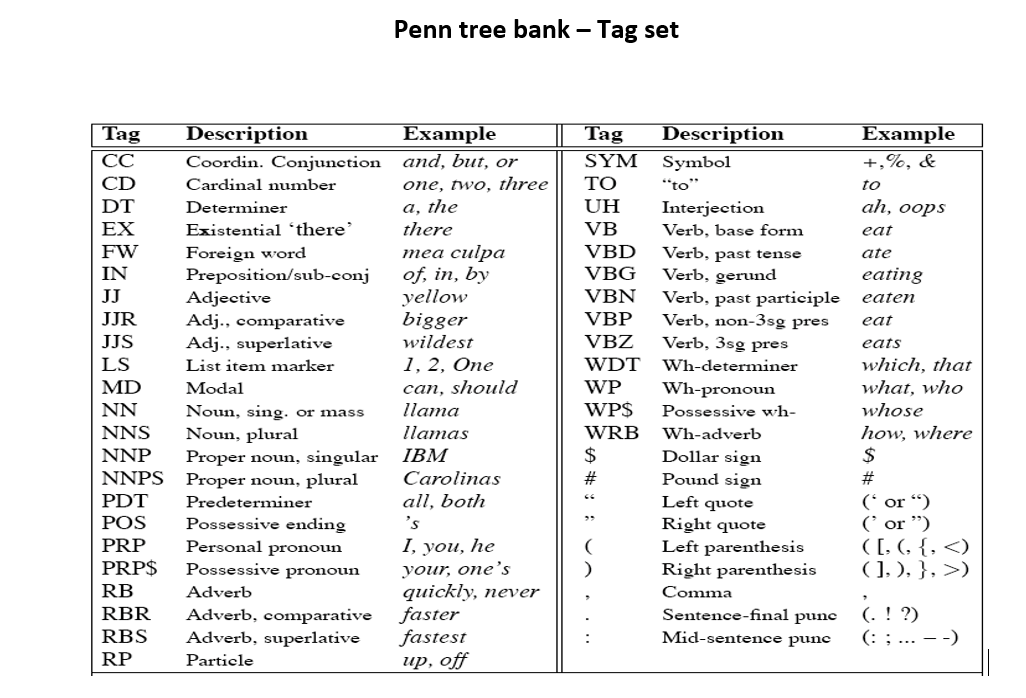
print(filtered\_words)

Output:

['Removing', 'stopwords', 'crucial', 'text', 'analysis', '.']

3. Part-of-Speech Tagging

Part-of-speech tagging involves categorizing words in a text into their respective grammatical categories (e.g., noun, verb, adjective). This information is valuable for understanding the syntactic structure of a sentence.



1."The quick brown fox jumps over the lazy dog."

A) The/**DT** quick /**JJ** brown /**JJ**  fox/**NN** jumps / **VBZ** over/**IN** the/ **DT** lazy/**JJ** dog/**NN** ./**PUNCT**

2. ."He was being opposed by her without any reason.\  
 A plan is being prepared by charles for next project"

1. [('He', 'PRP'), ('was', 'VBD'), ('being', 'VBG'), ('opposed', 'VBN'), ('by', 'IN'), ('her', 'PRP$'), ('without', 'IN'), ('any', 'DT'), ('reason', 'NN'), ('.', '.')] #Sentence 1[('A', 'DT'), ('plan', 'NN'), ('is', 'VBZ'), ('being', 'VBG'), ('prepared', 'VBN'), ('by', 'IN'), ('charles', 'NNS'), ('for', 'IN'), ('next', 'JJ'), ('project', 'NN')] #Sentence 2

3. I am sitting in Royal’s restaurant putting on the

Apollo fish, which is a dish I am very fond of, . . .

1. I**/PRP**  am**/VBP** sitting**/VBG** in**/IN** Royals**/NNP ’s/POS**

restaurant**/NN** putting**/VBG** on**/ IN** the**/DT** Apollo fish**/NN**

**,** which**/WDT** is**/VBZ** a**/DT** dish**/NN** I**/PRP** am**/VBP**

very**/JJ** fond**/JJ** of**/IN ,/,**

4. Lemmatization

Lemmatization involves reducing words to their base or root form. This is valuable for standardizing words and reducing dimensionality.

from nltk.stem import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()

word = "running"

lemma = lemmatizer.lemmatize(word, pos='v')

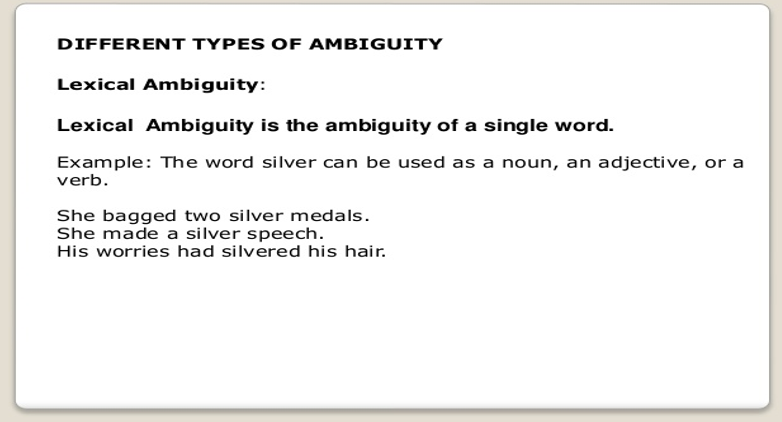
print(lemma)

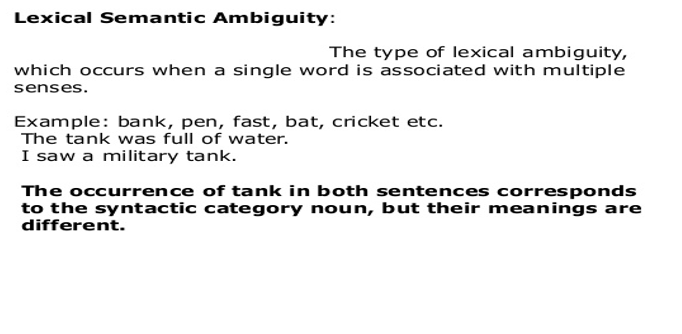
Output:

'run'

5. **Ambiguity**

Ambiguity in computational linguistics is a situation where a word or a sentence may have more than one meaning. That is, a sentence may be interpreted in more than one way. This leads to uncertainty in choosing the right meaning of a sentence especially while processing natural languages by computer.





**Syntactic ambiguity**

It is a type of ambiguity where the doubt is about the syntactic structure of the sentence. That is, there is a possibility that a sentence could be parsed in many syntactical forms (a sentence may be interpreted in more than one way). The doubt is about which one among different syntactical forms is correct.

For example, the sentence “old men and women” is ambiguous. Here, the doubt is that whether the adjective ***old*** is attached with ***both men and women*** or ***men alone***.

Syntactic (structural) ambiguity can be further classified into two ambiguities namely, scope ambiguity and attachment ambiguity.

* [Scope ambiguity](http://www.exploredatabase.com/2020/03/scope-ambiguity-in-natural-language-processing.html)
* [Attachment ambiguity](http://www.exploredatabase.com/2020/03/attachment-ambiguity-in-natural-language-processing.html)

### Scope ambiguity in natural language processing

**Scope ambiguity**

It arises when scope of a part of a sentence is unclear. It usually happens when a sentence have two or more quantifiers (eg. the words like ***every, any, some*** etc.).

For example, the sentence “Every man loves a woman” has two possible meanings as follows;

1.     For every man there is a woman such that he loves her

2.     There is one particular woman who is loved by every man

Based on the scope given to the quantifiers “***every***” and “***a***”, the meaning of the sentences could be decided. For instance, if the scope is given to the quantifier ***every***, then the meaning 1 arises. If the scope is given to the quantifier ***a***, then the meaning 2 arises.

**Attachment ambiguity**

It arises from uncertainty of attaching a phrase or clause to a part of sentence. It usually happens when a sentence has more than two prepositional phrases.

***Example 1***

In the sentence “***the boy saw the girl with the telescope***”, the uncertainty is about relating the prepositional phrase “***with the telescope***” to “***the boy***” or to “***the girl***”. This could end up with the following meaning based on the attachment;

1.     The boy saw the girl carrying a telescope

2.     The boy saw the girl through the telescope

The first meaning arises it we attach the prepositional phrase with “***the girl***” whereas the second one arises if we attach the prepositional phrase with “***the boy***”.

***Example 2***

Consider the following sentence;

“***Guna ate an ice cream with fruits from Chennai***”

In this sentence, we have two prepositional phrases “***with fruits***” and “***from Chennai***”. Here the possible meanings are as follows;

1.     Guna who is from Chennai ate an ice cream filled with fruits.

2.     Guna ate an ice cream filled with fruits and the ice cream is brought from Chennai.

3.     Guna who is from Chennai ate the ice cream with the help of fruits.

4.     Guna with the help of fruits ate the ice cream which is brought from Chennai

Here we got four possibilities due to two prepositional phrases. Each one arises from how we attach the prepositional phrases “***with fruits***” and “***from Chennai***” to either “***Guna***” or the “***ice cream***”.

### 3. Semantic ambiguity in natural language processing

**Semantic ambiguity**

Even after the syntax and the meanings of individual words are resolved, still there are more than one way of reading a sentence. For example, the sentence “the dog has been domesticated for more than 1000 years” could be semantically interpreted as follows;

1.     A particular dog has been domesticated or

2.     The dog species has been domesticated.

Semantic ambiguity is an uncertainty that occurs when a word, phrase or sentence has more than one interpretation.

Another example

“Ram kissed his son, and so did John”

The possible interpretations would be,

1.     Ram kissed his son and John too kissed Ram’s son

2.     Ram kissed his son and John kissed his son (actual meaning)

### 4. Anaphoric ambiguity in natural language processing

**Anaphoric ambiguity**

Anaphora in linguistics is about referring backwards (or an entity in another context) in a text.

“Suresh kicked the ball. It went out of the stadium”

In this sentence, the pronoun “***it***” is an anaphor. This anaphor refers to the entity “***the ball***” in the previous sentence. The entity is described as the ***antecedent*** of anaphor “***it***”. This example is simple and does not show any ambiguity. Let us see one more example sentence;

In this sentence, how do we relate the pronoun “***it***” with the previous sentence? We have three **antecedents** namely “***London***”, “***snow***” and “***yesterday***”. We can relate the anaphor to either “***London***”, or “***snow***”, or “***yesterday***”. We would be able to get the correct meaning if we relate the anaphor to the **antecedent** “***snow***”.

Some words (anaphors) in a sentence have little or no meaning of their own but instead refer to other words in the same or other sentences. ***Anaphoric ambiguity*** refers to such a situation where an anaphor have more than one possible reference in the same or other sentence.

**Some facts about anaphoric ambiguity**

* Anaphors are mostly pronouns, or noun phrases in some cases.

**Example:** “Darshan plays keyboard. He loves music”. In this sentence, the anaphor “***He***” is a pronoun.

“A puppy drank the milk. The cute little dog was satisfied”. In this sentence, the anaphor “***The cute little dog***” is a noun pharse.

* Anaphoric references may not explicitly present in the previous sentence. Instead it may refer a part of an entity (antecedent) in the previous sentence.

**Example:** “I went to the hospital, and they told me to go home and rest”. Here, the anaphor “***they***” refers not to the “***hospital***” directly, instead to the “***hospital staff***”.

* Anaphors may not be in the immediately previous sentence. They may present in the sentences before the previous one or may present in the same sentence.

**Example:** “The horse ran up the hill. It was very steep. It soon got tired”. Here, the anaphor “***it***” of the third sentence refers the “***horse***” in the first sentence.

6. N-grams

Definition: N-grams are contiguous sequences of n items (words or characters) from a given sample of text or speech.

Significance: Analyzing n-grams helps capture contextual information in a sequence of words. Bigrams (n=2) and trigrams (n=3) are commonly used to analyze relationships between adjacent words.