### ****1. What do you understand by Natural Language Processing?****

[Natural Language Processing](https://intellipaat.com/blog/what-is-natural-language-processing/" \t "https://intellipaat.com/blog/interview-question/nlp-interview-questions/_blank) is a field of computer science that deals with communication between computer systems and humans. It is a technique used in Artificial Intelligence and Machine Learning. It is used to create automated software that helps understand human spoken languages to extract useful information from the data it gets in the form of audio. Techniques in NLP allow computer systems to process and interpret data in the form of natural languages.

### ****If you want to learn Natural Language Processing then go through the following tutorial:****

### ****2. List any two real-life applications of Natural Language Processing.****

Two real-life applications of Natural Language Processing are as follows:

1. ****Google Translate****: Google Translate is one of the famous applications of Natural Language Processing. It helps convert written or spoken sentences into any language. Also, we can find the correct pronunciation and meaning of a word by using Google Translate. It uses advanced techniques of Natural Language Processing to achieve success in translating sentences into various languages.
2. ****Chatbots****: To provide a better customer support service, companies have started using chatbots for 24/7 service. **[AI Chatbots](https://intellipaat.com/blog/how-to-build-an-artificial-intelligence-chatbot/" \t "https://intellipaat.com/blog/interview-question/nlp-interview-questions/_blank)** help resolve the basic queries of customers. If a chatbot is not able to resolve any query, then it forwards it to the support team, while still engaging the customer. It helps make customers feel that the customer support team is quickly attending them. With the help of chatbots, companies have become capable of building cordial relations with customers. It is only possible with the help of Natural Language Processing.



### ****3. What are stop words?****

Stop words are said to be useless data for a search engine. Words such as articles, prepositions, etc. are considered as stop words. There are stop words such as was, were, is, am, the, a, an, how, why, and many more. In Natural Language Processing, we eliminate the stop words to understand and analyze the meaning of a sentence. The removal of stop words is one of the most important tasks for search engines. Engineers design the algorithms of search engines in such a way that they ignore the use of stop words. This helps show the relevant search result for a query.

### ****4. What is NLTK?****

NLTK is a Python library, which stands for Natural Language Toolkit. We use NLTK to process data in human spoken languages. NLTK allows us to apply techniques such as parsing, tokenization, lemmatization, stemming, and more to understand natural languages. It helps in categorizing text, parsing linguistic structure, analyzing documents, etc.

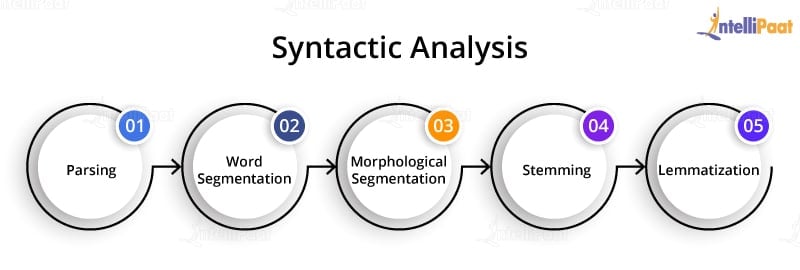
A few of the libraries of the NLTK package that we often use in NLP are:

1. SequentialBackoffTagger
2. DefaultTagger
3. UnigramTagger
4. treebank
5. wordnet
6. FreqDist
7. patterns
8. RegexpTagger
9. backoff\_tagger
10. UnigramTagger, BigramTagger, and TrigramTagger

### ****5. What is Syntactic Analysis?****

****Syntactic analysis**** is a technique of analyzing sentences to extract meaning from it. Using syntactic analysis, a machine can analyze and understand the order of words arranged in a sentence. NLP employs grammar rules of a language that helps in the syntactic analysis of the combination and order of words in documents.

The techniques used for syntactic analysis are as follows:

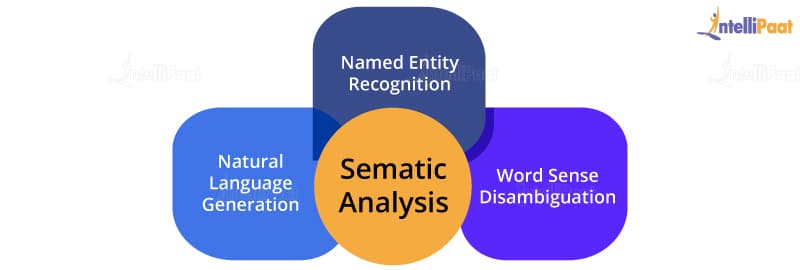


1. ****Parsing****: It helps in deciding the structure of a sentence or text in a document. It helps analyze the words in the text based on the grammar of the language.
2. ****Word segmentation****: The segmentation of words segregates the text into small significant units.
3. ****Morphological segmentation****: The purpose of morphological segmentation is to break words into their base form.
4. ****Stemming****: It is the process of removing the suffix from a word to obtain its root word.
5. ****Lemmatization****: It helps combine words using suffixes, without altering the meaning of the word.

### ****6. What is Semantic Analysis?****

Semantic analysis helps make a machine understand the meaning of a text. It uses various algorithms for the interpretation of words in sentences. It also helps understand the structure of a sentence.

Techniques used for semantic analysis are as given below:

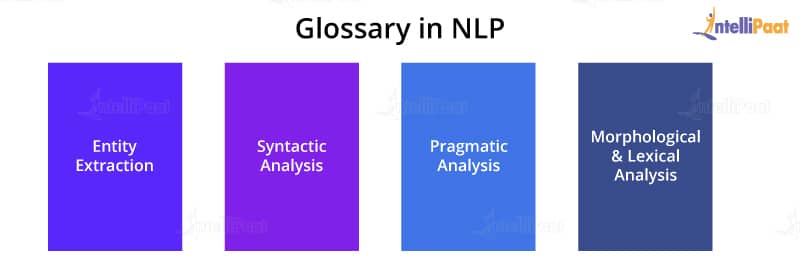


1. ****Named entity recognition:****This is the process of information retrieval that helps identify entities such as the name of a person, organization, place, time, emotion, etc.
2. ****Word sense disambiguation:****It helps identify the sense of a word used in different sentences.
3. ****Natural language generation****: It is a process used by the software to convert the structured data into human spoken languages. By using NLG, organizations can automate content for custom reports.

**If you want to learn Artificial Intelligence then enroll in [Artificial Intelligence Training](https://intellipaat.com/artificial-intelligence-course-india/" \t "https://intellipaat.com/blog/interview-question/nlp-interview-questions/_blank) now!**

### ****7. List the components of Natural Language Processing.****

The major components of NLP are as follows:



* ****Entity extraction:****Entity extraction refers to the retrieval of information such as place, person, organization, etc. by the segmentation of a sentence. It helps in the recognition of an entity in a text.
* ****Syntactic analysis:**** Syntactic analysis helps draw the specific meaning of a text.
* ****Pragmatic analysis:**** To find useful information from a text, we implement pragmatic analysis techniques.
* ****Morphological and lexical analysis:****It helps in explaining the structure of words by analyzing them through parsing.

### ****8. What is Latent Semantic Indexing (LSI)?****

Latent semantic indexing is a mathematical technique used to improve the accuracy of the information retrieval process. The design of LSI algorithms allows machines to detect the hidden (latent) correlation between semantics (words). To enhance information understanding, machines generate various concepts that associate with the words of a sentence.

The technique used for information understanding is called singular value decomposition. It is generally used to handle static and unstructured data. The matrix obtained for singular value decomposition contains rows for words and columns for documents. This method best suits to identify components and group them according to their types.

The main principle behind LSI is that words carry a similar meaning when used in a similar context. Computational LSI models are slow in comparison to other models. However, they are good at contextual awareness that helps improve the analysis and understanding of a text or a document.

### **C**9. What are Regular Expressions?****

A regular expression is used to match and tag words. It consists of a series of characters for matching strings.

Suppose, if A and B are regular expressions, then the following are true for them:

* If {ɛ} is a regular language, then ɛ is a regular expression for it.
* If A and B are regular expressions, then A + B is also a regular expression within the language {A, B}.
* If A and B are regular expressions, then the concatenation of A and B (A.B) is a regular expression.
* If A is a regular expression, then A\* (A occurring multiple times) is also a regular expression.

### ****10. What is Regular Grammar?****

Regular grammar is used to represent a regular language.

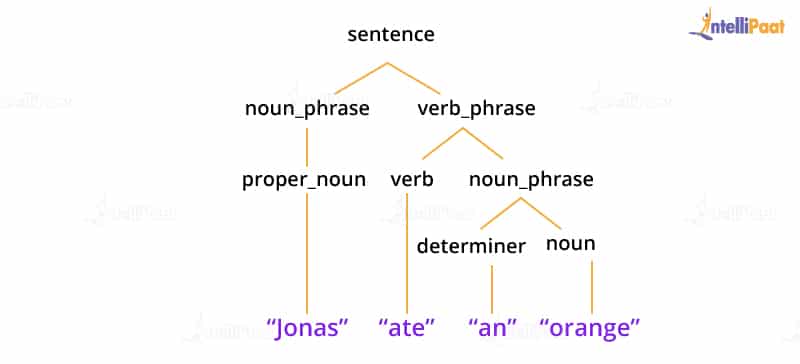
A regular grammar comprises rules in the form of ****A -> a****, ****A -> aB****, and many more. The rules help detect and analyze strings by automated computation.

Regular grammar consists of four tuples:

1. ‘N’ is used to represent the non-terminal set.
2. ‘∑’ represents the set of terminals.
3. ‘P’ stands for the set of productions.
4. ‘S € N’ denotes the start of non-terminal.

### ****11. What is Parsing in the context of NLP?****

Parsing in NLP refers to the understanding of a sentence and its grammatical structure by a machine. Parsing allows the machine to understand the meaning of a word in a sentence and the grouping of words, phrases, nouns, subjects, and objects in a sentence. Parsing helps analyze the text or the document to extract useful insights from it. To understand parsing, refer to the below diagram:



In this, ‘Jonas ate an orange’ is parsed to understand the structure of the sentence.

## ****Intermediate NLP Interview Questions****

### ****12. What is TF-IDF?****

TFIDF or Term Frequency-Inverse Document Frequency indicates the importance of a word in a set. It helps in information retrieval with numerical statistics. For a specific document, TF-IDF shows a frequency that helps identify the keywords in a document. The major use of TF-IDF in NLP is the extraction of useful information from crucial documents by statistical data. It is ideally used to classify and summarize the text in documents and filter out stop words.

****TF**** helps calculate the ratio of the frequency of a term in a document and the total number of terms. Whereas, ****IDF****denotes the importance of the term in a document.

The formula for calculating TF-IDF:

****TF(W) = (Frequency of W in a document)/(The total number of terms in the document)****

****IDF(W) = log\_e(The total number of documents/The number of documents having the term W)****

When ****TF****\*****IDF**** is high, the frequency of the term is less and vice versa.

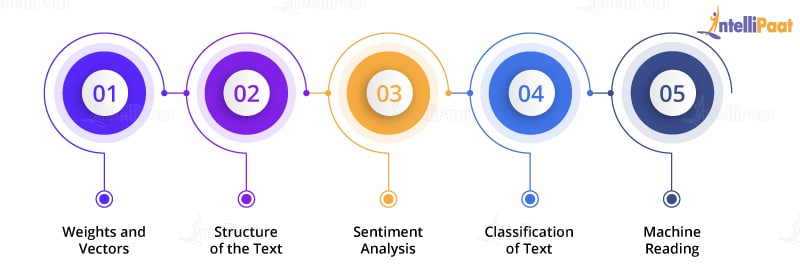
Google uses TF-IDF to decide the index of search results according to the relevancy of pages. The design of the TF-IDF algorithm helps optimize the search results in Google. It helps quality content rank up in search results.

**If you want to know more about ‘What is Natural Language Processing?’ you can go through this [Natural Language Processing Using Python](https://intellipaat.com/nlp-training-course-using-python/" \t "https://intellipaat.com/blog/interview-question/nlp-interview-questions/_blank) course!**

### ****13. Define the terminology in NLP.****

This is one of the most often asked NLP interview questions.

The interpretation of Natural Language Processing depends on various factors, and they are:



****Weights and Vectors****

* Use of TF-IDF for information retrieval
* Length (TF-IDF and doc)
* Google Word Vectors
* Word Vectors

****Structure of the Text****

* POS tagging
* Head of the sentence
* Named Entity Recognition (NER)

****Sentiment Analysis****

* Knowledge of the characteristics of sentiment
* Knowledge about entities and the common dictionary available for sentiment analysis

****Classification of Text****

* Supervised learning algorithm
* Training set
* Validation set
* Test set
* Features of the text
* LDA

****Machine Reading****

* Removal of possible entities
* Joining with other entities
* DBpedia

FRED (lib) Pikes

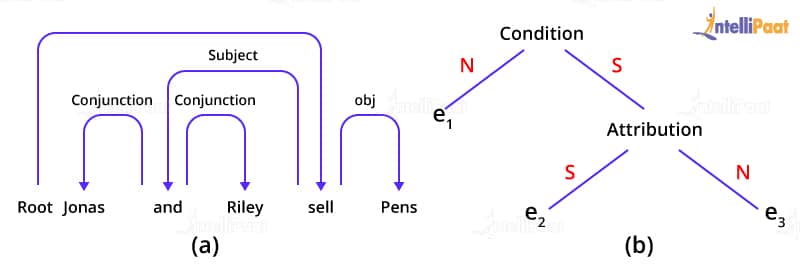
**Courses you may like**

### ****14. Explain Dependency Parsing in NLP.****

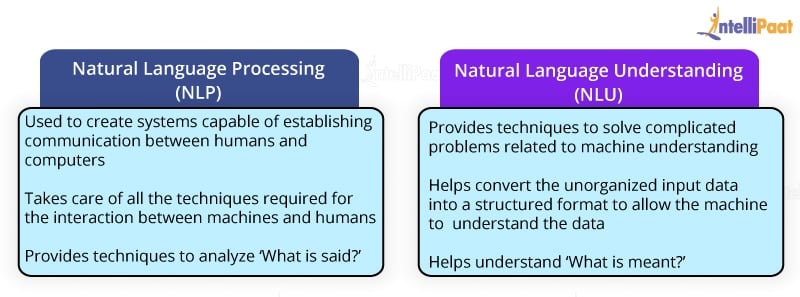
Dependency parsing helps assign a syntactic structure to a sentence. Therefore, it is also called syntactic parsing. Dependency parsing is one of the critical tasks in NLP. It allows the analysis of a sentence using parsing algorithms. Also, by using the parse tree in dependency parsing, we can check the grammar and analyze the semantic structure of a sentence.

For implementing dependency parsing, we use the spacy package. It implements token properties to operate the dependency parse tree.

The below diagram shows the dependency parse tree:

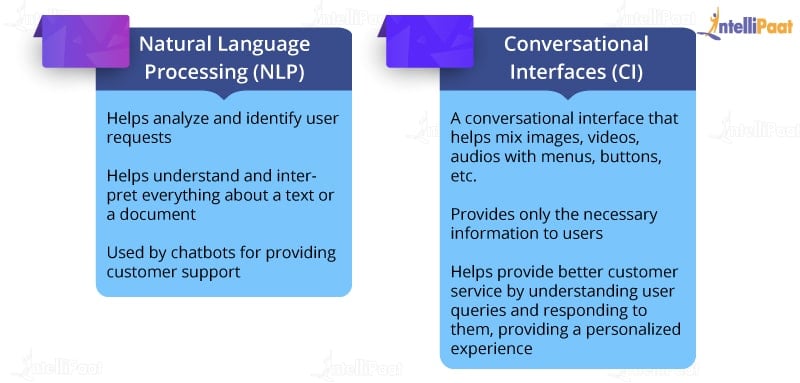


### ****15. What is the difference between NLP and NLU?****

The below table shows the difference between NLP and NLU:

### ****16. What is the difference between NLP and CI?****

The below table shows the difference between NLP and CI:



### ****17. What is Pragmatic Analysis?****

Pragmatic analysis is an important task in NLP for interpreting knowledge that is lying outside a given document. The aim of implementing pragmatic analysis is to focus on exploring a different aspect of the document or text in a language. This requires a comprehensive knowledge of the real world. The pragmatic analysis allows software applications for the critical interpretation of the real-world data to know the actual meaning of sentences and words.

****Example****:

Consider this sentence: ‘Do you know what time it is?’

This sentence can either be asked for knowing the time or for yelling at someone to make them note the time. This depends on the context in which we use the sentence.

### ****18. What is Pragmatic Ambiguity?****

Pragmatic ambiguity refers to the multiple descriptions of a word or a sentence. An ambiguity arises when the meaning of the sentence is not clear. The words of the sentence may have different meanings. Therefore, in practical situations, it becomes a challenging task for a machine to understand the meaning of a sentence. This leads to pragmatic ambiguity.

****Example****:

Check out the below sentence.

‘Are you feeling hungry?’

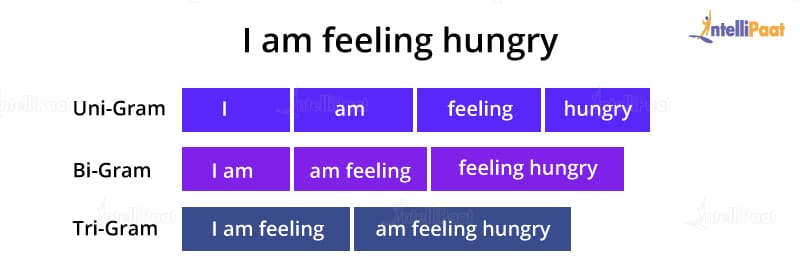
The given sentence could be either a question or a formal way of offering food.

### ****19. What are unigrams, bigrams, trigrams, and n-grams in NLP?****

When we parse a sentence one word at a time, then it is called a unigram. The sentence parsed two words at a time is a bigram.

When the sentence is parsed three words at a time, then it is a trigram. Similarly, n-gram refers to the parsing of n words at a time.

****Example****: To understand unigrams, bigrams, and trigrams, you can refer to the below diagram:



Therefore, parsing allows machines to understand the individual meaning of a word in a sentence. Also, this type of parsing helps predict the next word and correct spelling errors.

**Are you interested in learning Artificial Intelligence from experts? Enroll in our [AI Course in Bangalore now!](https://intellipaat.com/artificial-intelligence-course-toronto/" \t "https://intellipaat.com/blog/interview-question/nlp-interview-questions/_blank)**

### ****20. What are the steps involved in solving an NLP problem?****

Below are the steps involved in solving an NLP problem:

1. Gather the text from the available dataset or by web scraping
2. Apply stemming and lemmatization for text cleaning
3. Apply feature engineering techniques
4. Embed using ****word2vec****
5. Train the built model using neural networks or other Machine Learning techniques
6. Evaluate the model’s performance
7. Make appropriate changes in the model
8. Deploy the model

[](https://intellipaat.com/data-science-architect-masters-program-training/)

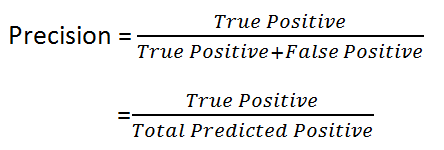
### ****21. What is Feature Extraction in NLP?****

Features or characteristics of a word help in text or document analysis. They also help in sentiment analysis of a text. Feature extraction is one of the techniques that are used by recommendation systems. Reviews such as ‘excellent,’ ‘good,’ or ‘great’ for a movie are positive reviews, recognized by a recommender system. The recommender system also tries to identify the features of the text that help in describing the context of a word or a sentence. Then, it makes a group or category of the words that have some common characteristics. Now, whenever a new word arrives, the system categorizes it as per the labels of such groups.

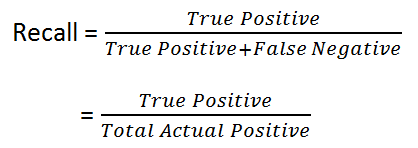
### ****22. What is precision and recall?****

The metrics used to test an NLP model are precision, recall, and F1. Also, we use accuracy for evaluating the model’s performance. The ratio of prediction and the desired output yields the accuracy of the model.

****Precision**** is the ratio of true positive instances and the total number of positively predicted instances.

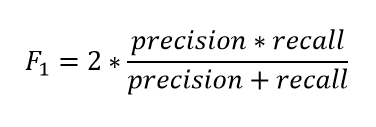


****Recall**** is the ratio of true positive instances and the total actual positive instances.



### ****23. What is F1 score in NLP?****

****F1 score**** evaluates the weighted average of recall and precision. It considers both false negative and false positive instances while evaluating the model. F1 score is more accountable than accuracy for an NLP model when there is an uneven distribution of class. Let us look at the formula for calculating F1 score:



**If you are preparing for an Artificial Intelligence job interview, go through this** **[top Artificial Intelligence Interview Questions and Answers](https://intellipaat.com/blog/interview-question/artificial-intelligence-interview-questions/" \t "https://intellipaat.com/blog/interview-question/nlp-interview-questions/_blank)**!****

## **Advanced NLP Interview Questions**

### ****24. How to tokenize a sentence using the nltk package?****

Tokenization is a process used in NLP to split a sentence into tokens. ****Sentence tokenization**** refers to splitting a text or paragraph into sentences.

For tokenizing, we will import ****sent\_tokenize**** from the ****nltk package****:

from nltk.tokenize import sent\_tokenize<>

We will use the below paragraph for sentence tokenization:  
Para = “Hi Guys. Welcome to Intellipaat. This is a blog on the NLP interview questions and answers.”

sent\_tokenize(Para)

****Output****:

[ 'Hi Guys.' ,

'Welcome to Intellipaat. ',

'This is a blog on the NLP interview questions and answers. ' ]

Tokenizing a word refers to splitting a sentence into words.

Now, to tokenize a word, we will import ****word\_tokenize**** from the nltk package.

from nltk.tokenize import word\_tokenize

Para = “Hi Guys. Welcome to Intellipaat. This is a blog on the NLP interview questions and answers.”

word\_tokenize(Para)

****Output****:

[ 'Hi' , 'Guys' , ' . ' , 'Welcome' , 'to' , 'Intellipaat' , ' . ' , 'This' , 'is' , 'a', 'blog' , 'on' , 'the' , 'NLP' , 'interview' , 'questions' , 'and' , 'answers' , ' . ' ]

### ****25. Explain how we can do parsing.****

Parsing is the method to identify and understand the syntactic structure of a text. It is done by analyzing the individual elements of the text. The machine parses the text one word at a time, then two at a time, further three, and so on.

* When the machine parses the text one word at a time, then it is a ****unigram****.
* When the text is parsed two words at a time, it is a ****bigram****.
* The set of words is a ****trigram**** when the machine parses three words at a time.

Look at the below diagram to understand unigram, bigram, and trigram.



Now, let’s implement parsing with the help of the ****nltk****package.

import nltk

text = ”Top 30 NLP interview questions and answers”

We will now tokenize the text using ****word\_tokenize****.

text\_token= word\_tokenize(text)

Now, we will use the function for extracting unigrams, bigrams, and trigrams.

list(nltk.unigrams(text))

****Output****:

[ "Top 30 NLP interview questions and answer"]

list(nltk.bigrams(text))

****Output****:

["Top 30", "30 NLP", "NLP interview", "interview questions", "questions and", "and answer"]

list(nltk.trigrams(text))

****Output****:

["Top 30 NLP", "NLP interview questions", "questions and answers"]

For extracting ****n-grams****, we can use the function ****nltk.ngrams**** and give the argument n for the number of parsers.

list(nltk.ngrams(text,n))

### ****26. Explain Stemming with the help of an example.****

In Natural Language Processing, stemming is the method to extract the root word by removing suffixes and prefixes from a word.  
For example, we can reduce ‘stemming’ to ‘stem’ by removing ‘m’ and ‘ing.’  
We use various algorithms for implementing stemming, and one of them is PorterStemmer.  
First, we will import ****PorterStemmer**** from the nltk package.

from nltk.stem import PorterStemmer

Creating an object for PorterStemmer

pst=PorterStemmer()

pst.stem(“running”), pst.stem(“cookies”), pst.stem(“flying”)

****Output****:

(‘run’, ‘cooki', ‘fly’ )

### ****27. Explain Lemmatization with the help of an example.****

We use stemming and lemmatization to extract root words. However, stemming may not give the actual word, whereas lemmatization generates a meaningful word.  
In lemmatization, rather than just removing the suffix and the prefix, the process tries to find out the root word with its proper meaning.  
****Example****: ‘Bricks’ becomes ‘brick,’ ‘corpora’ becomes ‘corpus,’ etc.  
Let’s implement lemmatization with the help of some nltk packages.  
First, we will import the required packages.

from nltk.stem import wordnet

from nltk.stem import WordnetLemmatizer

Creating an object for WordnetLemmatizer()

lemma= WordnetLemmatizer()

list = [“Dogs”, “Corpora”, “Studies”]

for n in list:

print(n + “:” + lemma.lemmatize(n))

****Output****:

Dogs: Dog

Corpora: Corpus

Studies: Study

****Interested in learning Artificial Intelligence? Go through this**[Artificial Intelligence Tutorial](https://intellipaat.com/blog/tutorial/artificial-intelligence-tutorial/" \t "https://intellipaat.com/blog/interview-question/nlp-interview-questions/_blank)**!****

### ****28. What is Parts-of-speech Tagging?****

The parts-of-speech (POS) tagging is used to assign tags to words such as nouns, adjectives, verbs, and more. The software uses the POS tagging to first read the text and then differentiate the words by tagging. The software uses algorithms for the parts-of-speech tagging. POS tagging is one of the most essential tools in Natural Language Processing. It helps in making the machine understand the meaning of a sentence.  
We will look at the implementation of the POS tagging using stop words.  
Let’s import the required nltk packages.

import nltk

from nltk.corpus import stopwords

from nltk.tokenize import word\_tokenize, sent\_tokenize

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

txt = "Sourav, Pratyush, and Abhinav are good friends."

Tokenizing using sent\_tokenize

tokenized\_text = sent\_tokenize(txt)

To find punctuation and words in a string, we will use ****word\_tokenizer**** and then remove the stop words.

for n in tokenized\_text:

wordsList = nltk.word\_tokenize(i)

wordsList = [w for w in wordsList if not w instop\_words]

Now, we will use the POS tagger.

tagged\_words = nltk.pos\_tag(wordsList)

print(tagged\_words)

****Output****:

[('Sourav', 'NNP'), ('Pratyush', 'NNP'), ('Abhinav', 'NNP'), ('good', 'JJ'), ('friends', 'NNS')]

### ****29. Explain Named Entity Recognition by implementing it.****

Named Entity Recognition (NER) is an information retrieval process. NER helps classify named entities such as monetary figures, location, things, people, time, and more. It allows the software to analyze and understand the meaning of the text. NER is mostly used in NLP, Artificial Intelligence, and Machine Learning. One of the real-life applications of NER is chatbots used for customer support.  
Let’s implement NER using the ****spacy**** package.  
Importing the spacy package:

import spacy

nlp = spacy.load('en\_core\_web\_sm')

Text = "The head office of Google is in California"

document = nlp(text)for ent in document.ents:

print(ent.text, ent.start\_char, ent.end\_char, ent.label\_)

****Output****:

Office 9 15 Place

Google 19 25 ORG

California 32 41 GPE

### ****30. How to check word similarity using the spacy package?****

To find out the similarity among words, we use word similarity. We evaluate the similarity with the help of a number that lies between 0 and 1. We use the spacy library to implement the technique of word similarity.

import spacy

nlp = spacy.load('en\_core\_web\_md')

print("Enter the words")

input\_words = input()

tokens = nlp(input\_words)

for i in tokens:

print(i.text, i.has\_vector, i.vector\_norm, i.is\_oov)

token\_1, token\_2 = tokens[0], tokens[1]

print("Similarity between words:", token\_1.similarity(token\_2))

****Output****:

hot  True 5.6898586 False

cold True6.5396233 False

Similarity: 0.597265

### ****1. What is Naive Bayes algorithm, When we can use this algorithm in NLP?****

[Naive Bayes algorithm](https://www.mygreatlearning.com/blog/introduction-to-naive-bayes/" \t "https://www.mygreatlearning.com/blog/nlp-interview-questions/_blank) is a collection of classifiers which works on the principles of the Bayes’ theorem. This series of NLP model forms a family of algorithms that can be used for a wide range of classification tasks including sentiment prediction, filtering of spam, classifying documents and more.

Naive Bayes algorithm converges faster and requires less training data. Compared to other discriminative models like logistic regression, Naive Bayes model it takes lesser time to train. This algorithm is perfect for use while working with multiple classes and text classification where the data is dynamic and changes frequently.

### ****2. Explain Dependency Parsing in NLP?****

Dependency Parsing, also known as Syntactic parsing in NLP is a process of assigning syntactic structure to a sentence and identifying its dependency parses. This process is crucial to understand the correlations between the “head” words in the syntactic structure.  
The process of dependency parsing can be a little complex considering how any sentence can have more than one dependency parses. Multiple parse trees are known as ambiguities. Dependency parsing needs to resolve these ambiguities in order to effectively assign a syntactic structure to a sentence.

Dependency parsing can be used in the semantic analysis of a sentence apart from the syntactic structuring.

### ****3. What is text Summarization?****

[Text summarization](https://www.mygreatlearning.com/blog/text-summarization-in-python/" \t "https://www.mygreatlearning.com/blog/nlp-interview-questions/_blank) is the process of shortening a long piece of text with its meaning and effect intact. Text summarization intends to create a summary of any given piece of text and outlines the main points of the document. This technique has improved in recent times and is capable of summarizing volumes of text successfully.

Text summarization has proved to a blessing since machines can summarise large volumes of text in no time which would otherwise be really time-consuming. There are two types of text summarization:

* Extraction-based summarization
* Abstraction-based summarization

### ****4. What is NLTK? How is it different from Spacy?****

NLTK or Natural Language Toolkit is a series of libraries and programs that are used for symbolic and statistical natural language processing. This toolkit contains some of the most powerful libraries that can work on different ML techniques to break down and understand human language. NLTK is used for Lemmatization, Punctuation, Character count, Tokenization, and Stemming. The difference between NLTK and Spacey are as follows:

* While NLTK has a collection of programs to choose from, Spacey contains only the best-suited algorithm for a problem in its toolkit
* NLTK supports a wider range of languages compared to Spacey (Spacey supports only 7 languages)
* While Spacey has an object-oriented library, NLTK has a string processing library
* Spacey can support word vectors while NLTK cannot

### ****5. What is information extraction?****

Information extraction in the context of Natural Language Processing refers to the technique of extracting structured information automatically from unstructured sources to ascribe meaning to it. This can include extracting information regarding attributes of entities, relationship between different entities and more. The various models of information extraction includes:

* Tagger Module
* Relation Extraction Module
* Fact Extraction Module
* Entity Extraction Module
* Sentiment Analysis Module
* Network Graph Module
* Document Classification & Language Modeling Module

### ****6. What is Bag of Words?****

[Bag of Words](https://www.mygreatlearning.com/blog/bag-of-words/" \t "https://www.mygreatlearning.com/blog/nlp-interview-questions/_blank) is a commonly used model that depends on word frequencies or occurrences to train a classifier. This model creates an occurrence matrix for documents or sentences irrespective of its grammatical structure or word order.

### ****7. What is Pragmatic Ambiguity in NLP?****

Pragmatic ambiguity refers to those words which have more than one meaning and their use in any sentence can depend entirely on the context. Pragmatic ambiguity can result in multiple interpretations of the same sentence. More often than not, we come across sentences which have words with multiple meanings, making the sentence open to interpretation. This multiple interpretation causes ambiguity and is known as Pragmatic ambiguity in NLP.

### ****8. What is Masked Language Model?****

Masked language models help learners to understand deep representations in downstream tasks by taking an output from the corrupt input. This model is often used to predict the words to be used in a sentence.

### ****9. What is the difference between NLP and CI(Conversational Interface)?****

The difference between NLP and CI is as follows:

| ****Natural Language Processing (NLP)**** | ****Conversational Interface (CI)**** |
| --- | --- |
| NLP attempts to help machines understand and learn how language concepts work. | CI focuses only on providing users with an interface to interact with. |
| NLP uses AI technology to identify, understand, and interpret the requests of users through language. | CI uses voice, chat, videos, images, and more such conversational aid to create the user interface. |

### ****10. What are the best NLP Tools?****

Some of the best NLP tools from open sources are:

* SpaCy
* TextBlob
* Textacy
* Natural language Toolkit ([NLTK](https://www.mygreatlearning.com/blog/nltk-tutorial-with-python/" \t "https://www.mygreatlearning.com/blog/nlp-interview-questions/_blank))
* Retext
* NLP.js
* Stanford NLP
* CogcompNLP

### ****11. What is POS tagging?****

Parts of speech tagging better known as [POS tagging](https://www.mygreatlearning.com/blog/pos-tagging/" \t "https://www.mygreatlearning.com/blog/nlp-interview-questions/_blank) refer to the process of identifying specific words in a document and grouping them as part of speech, based on its context. POS tagging is also known as grammatical tagging since it involves understanding grammatical structures and identifying the respective component.

POS tagging is a complicated process since the same word can be different parts of speech depending on the context. The same general process used for word mapping is quite ineffective for POS tagging because of the same reason.

### ****12. What is NES?****

Name entity recognition is more commonly known as NER is the process of identifying specific entities in a text document that are more informative and have a unique context. These often denote places, people, organizations, and more. Even though it seems like these entities are proper nouns, the NER process is far from identifying just the nouns. In fact, NER involves entity chunking or extraction wherein entities are segmented to categorize them under different predefined classes. This step further helps in extracting information.

## ****NLP Interview Questions for Experienced****

### ****13. Which of the following techniques can be used for keyword normalization in NLP, the process of converting a keyword into its base form?****

a. Lemmatization  
b. Soundex  
c. Cosine Similarity  
d. N-grams

****Answer:**** a)

Lemmatization helps to get to the base form of a word, e.g. are playing -> play, eating -> eat, etc. Other options are meant for different purposes.

### ****14. Which of the following techniques can be used to compute the distance between two-word vectors in NLP?****

a. Lemmatization  
b. Euclidean distance  
c. Cosine Similarity  
d. N-grams

****Answer:**** b) and c)

Distance between two-word vectors can be computed using Cosine similarity and Euclidean Distance.  Cosine Similarity establishes a cosine angle between the vector of two words. A cosine angle close to each other between two-word vectors indicates the words are similar and vice versa.

E.g. cosine angle between two words “Football” and “Cricket” will be closer to 1 as compared to the angle between the words “Football” and “New Delhi”.

Python code to implement CosineSimlarity function would look like this:

def cosine\_similarity(x,y):

return np.dot(x,y)/( np.sqrt(np.dot(x,x)) \* np.sqrt(np.dot(y,y)) )

q1 = wikipedia.page(‘Strawberry’)

q2 = wikipedia.page(‘Pineapple’)

q3 = wikipedia.page(‘Google’)

q4 = wikipedia.page(‘Microsoft’)

cv = CountVectorizer()

X = np.array(cv.fit\_transform([q1.content, q2.content, q3.content, q4.content]).todense())

print (“Strawberry Pineapple Cosine Distance”, cosine\_similarity(X[0],X[1]))

print (“Strawberry Google Cosine Distance”, cosine\_similarity(X[0],X[2]))

print (“Pineapple Google Cosine Distance”, cosine\_similarity(X[1],X[2]))

print (“Google Microsoft Cosine Distance”, cosine\_similarity(X[2],X[3]))

print (“Pineapple Microsoft Cosine Distance”, cosine\_similarity(X[1],X[3]))

Strawberry Pineapple Cosine Distance 0.8899200413701714

Strawberry Google Cosine Distance 0.7730935582847817

Pineapple Google Cosine Distance 0.789610214147025

Google Microsoft Cosine Distance 0.8110888282851575

Usually Document similarity is measured by how close semantically the content (or words) in the document are to each other. When they are close, the similarity index is close to 1, otherwise near 0.

The **Euclidean distance** between two points is the length of the shortest path connecting them. Usually computed using Pythagoras theorem for a triangle.

### ****15. What are the possible features of a text corpus in NLP?****

a. Count of the word in a document  
b. Vector notation of the word  
c. Part of Speech Tag  
d. Basic Dependency Grammar  
e. All of the above

****Answer:**** e)

All of the above can be used as features of the text corpus.

### ****16. You created a document term matrix on the input data of 20K documents for a Machine learning model. Which of the following can be used to reduce the dimensions of data?****

1. Keyword Normalization
2. Latent Semantic Indexing
3. Latent Dirichlet Allocation

a. only 1  
b. 2, 3  
c. 1, 3  
d. 1, 2, 3

****Answer:**** d)

### ****17. Which of the text parsing techniques can be used for noun phrase detection, verb phrase detection, subject detection, and object detection in NLP.****

a. Part of speech tagging  
b. Skip Gram and N-Gram extraction  
c. Continuous Bag of Words  
d. Dependency Parsing and Constituency Parsing

****Answer:**** d)

### ****18. Dissimilarity between words expressed using cosine similarity will have values significantly higher than 0.5****

a. True  
b. False

****Answer:****a)

### ****19. Which one of the following is keyword Normalization techniques in NLP****

a. Stemming  
b. Part of Speech  
c. Named entity recognition  
d. Lemmatization

Answer: a) and d)

Part of Speech (POS) and Named Entity Recognition(NER) is not keyword Normalization techniques. Named Entity helps you extract Organization, Time, Date, City, etc., type of entities from the given sentence, whereas Part of Speech helps you extract Noun, Verb, Pronoun, adjective, etc., from the given sentence tokens.

### ****20. Which of the below are NLP use cases?****

a. Detecting objects from an image  
b. Facial Recognition  
c. Speech Biometric  
d. Text Summarization

Ans: d)

a) And b) are Computer Vision use cases, and c) is the Speech use case.  
Only d) Text Summarization is an NLP use case.

### ****21. In a corpus of N documents, one randomly chosen document contains a total of T terms and the term “hello” appears K times.****

What is the correct value for the product of TF (term frequency) and IDF (inverse-document-frequency), if the term “hello” appears in approximately one-third of the total documents?  
a. KT \* Log(3)  
b. T \* Log(3) / K  
c. K \* Log(3) / T  
d. Log(3) / KT

****Answer:**** (c)

formula for TF is K/T  
formula for IDF is log(total docs / no of docs containing “data”)  
= log(1 / (⅓))  
= log (3)

Hence, the correct choice is Klog(3)/T

### ****22. In NLP, The algorithm decreases the weight for commonly used words and increases the weight for words that are not used very much in a collection of documents****

a. Term Frequency (TF)  
b. Inverse Document Frequency (IDF)  
c. Word2Vec  
d. Latent Dirichlet Allocation (LDA)

****Answer:**** b)

### ****23. In NLP, The process of removing words like “and”, “is”, “a”, “an”, “the” from a sentence is called as****

a. Stemming  
b. Lemmatization  
c. Stop word  
d. All of the above

****Ans:**** c)

In Lemmatization, all the stop words such as a, an, the, etc.. are removed. One can also define custom stop words for removal.

### ****24. In NLP, The process of converting a sentence or paragraph into tokens is referred to as Stemming****

a. True  
b. False

****Answer:**** b)

The statement describes the process of tokenization and not stemming, hence it is False.

### ****25. In NLP, Tokens are converted into numbers before giving to any Neural Network****

a. True  
b. False

****Answer:**** a)

In NLP, all words are converted into a number before feeding to a Neural Network.

### ****26. Identify the odd one out****

a. nltk  
b. scikit learn  
c. SpaCy  
d. BERT

****Answer:**** d)

All the ones mentioned are NLP libraries except BERT, which is a word embedding.

### ****27. TF-IDF helps you to establish?****

a. most frequently occurring word in document  
b. the most important word in the document

****Answer:**** b)

TF-IDF helps to establish how important a particular word is in the context of the document corpus. TF-IDF takes into account the number of times the word appears in the document and is offset by the number of documents that appear in the corpus.

* TF is the frequency of terms divided by the total number of terms in the document.
* IDF is obtained by dividing the total number of documents by the number of documents containing the term and then taking the logarithm of that quotient.
* Tf.idf is then the multiplication of two values TF and IDF.

Suppose that we have term count tables of a corpus consisting of only two documents, as listed here:

|  |  |  |
| --- | --- | --- |
| **Term** | **Document 1 Frequency** | **Document 2 Frequency** |
| This | 1 | 1 |
| is | 1 | 1 |
| a | 2 |  |
| Sample | 1 |  |
| another |  | 2 |
| example |  | 3 |

The calculation of tf–idf for the term “this” is performed as follows:

for "this"

-----------

tf("this", d1) = 1/5 = 0.2

tf("this", d2) = 1/7 = 0.14

idf("this", D) = log (2/2) =0

hence tf-idf

tfidf("this", d1, D) = 0.2\* 0 = 0

tfidf("this", d2, D) = 0.14\* 0 = 0

for "example"

------------

tf("example", d1) = 0/5 = 0

tf("example", d2) = 3/7 = 0.43

idf("example", D) = log(2/1) = 0.301

tfidf("example", d1, D) = tf("example", d1) \* idf("example", D) = 0 \* 0.301 = 0

tfidf("example", d2, D) = tf("example", d2) \* idf("example", D) = 0.43 \* 0.301 = 0.129

In its raw frequency form, TF is just the frequency of the “this” for each document. In each document, the word “this” appears once; but as document 2 has more words, its relative frequency is smaller.

An IDF is constant per corpus, and accounts for the ratio of documents that include the word “this”. In this case, we have a corpus of two documents and all of them include the word “this”. So TF–IDF is zero for the word “this”, which implies that the word is not very informative as it appears in all documents.

The word “example” is more interesting – it occurs three times, but only in the second document. To understand more about NLP, check out these [NLP projects](https://www.mygreatlearning.com/academy/learn-for-free/courses/natural-language-processing-projects" \t "https://www.mygreatlearning.com/blog/nlp-interview-questions/_blank).

### ****28. In NLP, The process of identifying people, an organization from a given sentence, paragraph is called****

a. Stemming  
b. Lemmatization  
c. Stop word removal  
d. Named entity recognition

****Answer:**** d)

### ****29. Which one of the following is not a pre-processing technique in NLP****

a. Stemming and Lemmatization  
b. converting to lowercase  
c. removing punctuations  
d. removal of stop words  
e. Sentiment analysis

****Answer:**** e)

Sentiment Analysis is not a pre-processing technique. It is done after pre-processing and is an NLP use case. All other listed ones are used as part of statement pre-processing.

### ****30. In text mining, converting text into tokens and then converting them into an integer or floating-point vectors can be done using****

a. CountVectorizer  
b.  TF-IDF  
c. Bag of Words  
d. NERs

****Answer:**** a)

CountVectorizer helps do the above, while others are not applicable.

text =["Rahul is an avid writer, he enjoys studying understanding and presenting. He loves to play"]

vectorizer = CountVectorizer()

vectorizer.fit(text)

vector = vectorizer.transform(text)

print(vector.toarray())

****Output****

[[1 1 1 1 2 1 1 1 1 1 1 1 1 1]]

The second section of the interview questions covers advanced NLP techniques such as Word2Vec, GloVe word embeddings, and advanced models such as GPT, Elmo, BERT, XLNET-based*questions, and explanations.*

### ****31. In NLP, Words represented as vectors are called Neural Word Embeddings****

a. True  
b. False

****Answer:**** a)

Word2Vec, GloVe based models build word embedding vectors that are multidimensional.

### ****32. In NLP, Context modeling is supported with which one of the following word embeddings****

1. a. Word2Vec
2. b) GloVe
3. c) BERT
4. d) All of the above

****Answer:**** c)

Only BERT (Bidirectional Encoder Representations from Transformer) supports context modelling where the previous and next sentence context is taken into consideration. In Word2Vec, GloVe only word embeddings are considered and previous and next sentence context is not considered.

### ****33. In NLP, Bidirectional context is supported by which of the following embedding****

a. Word2Vec  
b. BERT  
c. GloVe  
d. All the above

****Answer:**** b)

Only BERT provides a bidirectional context. The BERT model uses the previous and the next sentence to arrive at the context.Word2Vec and GloVe are word embeddings, they do not provide any context.

### ****34. Which one of the following Word embeddings can be custom trained for a specific subject in NLP****

a. Word2Vec  
b. BERT  
c. GloVe  
d. All the above

****Answer:**** b)

BERT allows Transform Learning on the existing pre-trained models and hence can be custom trained for the given specific subject, unlike Word2Vec and GloVe where existing word embeddings can be used, no transfer learning on text is possible.

### ****35. Word embeddings capture multiple dimensions of data and are represented as vectors****

a. True  
b. False

****Answer:**** a)

### ****36.** **In NLP, Word embedding vectors help establish distance between two tokens****

a. True  
b. False

****Answer: a)****

****One can use Cosine similarity to establish the****distance between two vectors represented through Word Embeddings

### ****37. Language Biases are introduced due to historical data used during training of word embeddings, which one amongst the below is not an example of bias****

a. New Delhi is to India, Beijing is to China  
b. Man is to Computer, Woman is to Homemaker

****Answer:**** a)

Statement b) is a bias as it buckets Woman into Homemaker, whereas statement a) is not a biased statement.

### ****38. Which of the following will be a better choice to address NLP use cases such as semantic similarity, reading comprehension, and common sense reasoning****

a. ELMo  
b. Open AI’s GPT  
c. ULMFit

****Answer:****b)

Open AI’s GPT is able to learn complex patterns in data by using the Transformer models Attention mechanism and hence is more suited for complex use cases such as semantic similarity, reading comprehensions, and common sense reasoning.

### ****39. Transformer architecture was first introduced with?****

a. GloVe  
b. BERT  
c. Open AI’s GPT  
d. ULMFit

****Answer:****c)

ULMFit has an LSTM based Language modeling architecture. This got replaced into Transformer architecture with Open AI’s GPT.

### ****40. Which of the following architecture can be trained faster and needs less amount of training data****

a. LSTM-based Language Modelling  
b. Transformer architecture

****Answer:**** b)

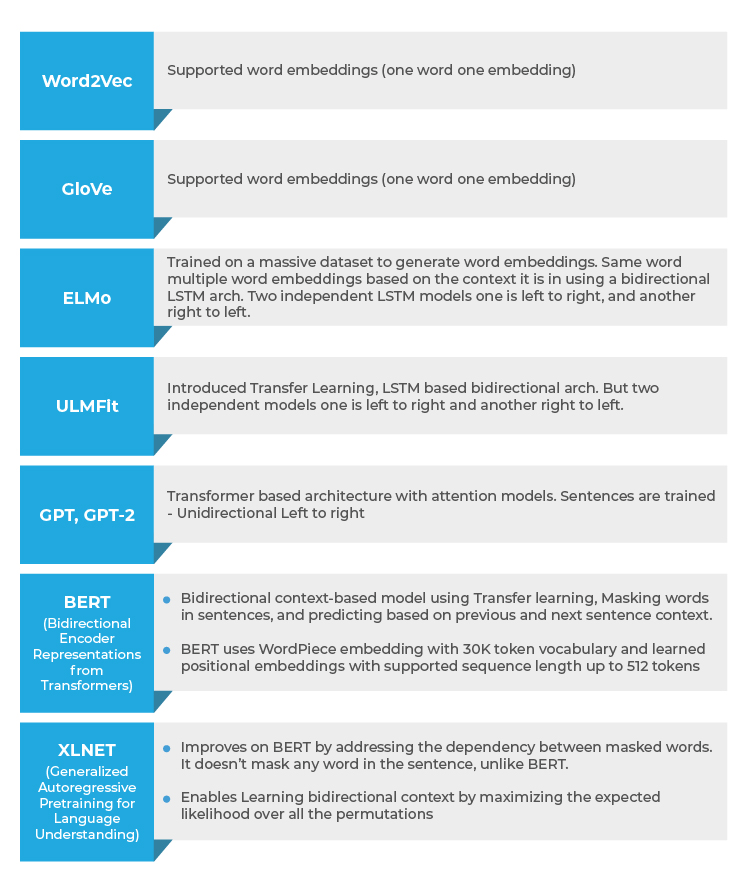
Transformer architectures were supported from GPT onwards and were faster to train and needed less amount of data for training too.

### ****41. Same word can have multiple word embeddings possible with \_\_\_\_\_\_\_\_\_\_\_\_?****

a. GloVe  
b. Word2Vec  
c. ELMo  
d. nltk

****Answer:****c)

EMLo word embeddings support the same word with multiple embeddings, this helps in using the same word in a different context and thus captures the context than just the meaning of the word unlike in GloVe and Word2Vec. Nltk is not a word embedding.



### ****42. For a given token, its input representation is the sum of embedding from the token, segment and position****

****embedding**  
a. ELMo  
b. GPT  
c. BERT  
d. ULMFit  
**Answer:** c)  
BERT uses token, segment and position embedding.**

### ****43. Trains two independent LSTM language model left to right and right to left and shallowly concatenates them.****

**a. GPT  
b. BERT  
c. ULMFit  
d. ELMo  
**Answer:** d)  
ELMo tries to train two independent LSTM language models (left to right and right to left) and concatenates the results to produce word embedding.**

### ****44. Uses unidirectional language model for producing word embedding.****

a. BERT  
b. GPT  
c. ELMo  
d. Word2Vec

****Answer:****b)

GPT is a bidirectional model and word embedding is produced by training on information flow from left to right. ELMo is bidirectional but shallow. Word2Vec provides simple word embedding.

### ****45. In this architecture, the relationship between all words in a sentence is modelled irrespective of their position. Which architecture is this?****

a. OpenAI GPT  
b. ELMo  
c. BERT  
d. ULMFit

****Ans:**** c)

BERT Transformer architecture models the relationship between each word and all other words in the sentence to generate attention scores. These attention scores are later used as weights for a weighted average of all words’ representations which is fed into a fully-connected network to generate a new representation.

### ****46. List 10 use cases to be solved using NLP techniques?****

* Sentiment Analysis
* Language Translation (English to German, Chinese to English, etc..)
* Document Summarization
* Question Answering
* Sentence Completion
* Attribute extraction (Key information extraction from the documents)
* Chatbot interactions
* Topic classification
* Intent extraction
* Grammar or Sentence correction
* Image captioning
* Document Ranking
* Natural Language inference

### ****47. Transformer model pays attention to the most important word in Sentence**.**

a. True  
b. False

****Ans:**** a) Attention mechanisms in the Transformer model are used to model the relationship between all words and also provide weights to the most important word.

### ****48. Which NLP model gives the best accuracy amongst the following?****

a. BERT  
b. XLNET  
c. GPT-2  
d. ELMo

****Ans:**** b) XLNET

XLNET has given best accuracy amongst all the models. It has outperformed BERT on 20 tasks and achieves state of art results on 18 tasks including sentiment analysis, question answering, natural language inference, etc.

### ****49. Permutation Language models is a feature of****

a. BERT  
b. EMMo  
c. GPT  
d. XLNET

****Ans:**** d)

XLNET provides permutation-based language modelling and is a key difference from BERT. In permutation language modeling, tokens are predicted in a random manner and not sequential. The order of prediction is not necessarily left to right and can be right to left. The original order of words is not changed but a prediction can be random. The conceptual difference between BERT and XLNET can be seen from the following diagram.

### ****50. Transformer XL uses relative positional embedding****

a. True  
b. False

****Ans:**** a)

Instead of embedding having to represent the absolute position of a word, Transformer XL uses an embedding to encode the relative distance between the words. This embedding is used to compute the attention score between any 2 words that could be separated by n words before or after.

There, you have it – all the probable questions for your NLP interview. Now go, give it your best shot.

### ****Natural Language Processing FAQs****

### ****1. Why do we need NLP?****

One of the main reasons why NLP is necessary is because it helps computers communicate with humans in natural language. It also scales other language-related tasks. Because of NLP, it is possible for computers to hear speech, interpret this speech, measure it and also determine which parts of the speech are important.

### ****2. What must a natural language program decide?****

A natural language program must decide what to say and when to say something.

### ****3. Where can NLP be useful?****

NLP can be useful in communicating with humans in their own language. It helps improve the efficiency of the machine translation and is useful in emotional analysis too. It can be helpful in [sentiment analysis using python](https://www.mygreatlearning.com/academy/learn-for-free/courses/sentiment-analysis-using-python?gl_blog_id=7798" \t "https://www.mygreatlearning.com/blog/nlp-interview-questions/_blank) too. It also helps in structuring highly unstructured data. It can be helpful in creating chatbots, Text Summarization and virtual assistants.

### ****4. How to prepare for an NLP Interview?****

The best way to prepare for an NLP Interview is to be clear about the basic concepts. Go through blogs that will help you cover all the key aspects and remember the important topics. Learn specifically for the interviews and be confident while answering all the questions.

### ****5. What are the main challenges of NLP?****

Breaking sentences into tokens, Parts of speech tagging, Understanding the context, Linking components of a created vocabulary, and Extracting semantic meaning are currently some of the main challenges of NLP.

### ****6. Which NLP model gives best accuracy?****

Naive Bayes Algorithm has the ****highest accuracy**** when it comes to NLP models. It gives up to 73% correct predictions.

### ****7. What are the major tasks of NLP?****

Translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation are few of the major tasks of NLP. Under unstructured data, there can be a lot of untapped information that can help an organization grow.

### ****8. What are stop words in NLP?****

Common words that occur in sentences that add weight to the sentence are known as stop words. These stop words act as a bridge and ensure that sentences are grammatically correct. In simple terms, words that are filtered out before processing natural language data is known as a stop word and it is a common pre-processing method.

### ****9. What is stemming in NLP?****

The process of obtaining the root word from the given word is known as stemming. All tokens can be cut down to obtain the root word or the stem with the help of efficient and well-generalized rules. It is a rule-based process and is well-known for its simplicity.

### ****10. Why is NLP so hard?****

There are several factors that make the process of Natural Language Processing difficult. There are hundreds of natural languages all over the world, words can be ambiguous in their meaning, each natural language has a different script and syntax, the meaning of words can change depending on the context, and so the process of NLP can be difficult. If you choose to upskill and continue learning, the process will become easier over time.

### ****11. What does a NLP pipeline consist of \*?****

The overall architecture of an ****NLP pipeline consists**** of several layers: a user interface; one or several ****NLP**** models, depending on the use case; a Natural Language Understanding layer to describe the ****meaning of**** words and sentences; a preprocessing layer; microservices for linking the components together and of course.

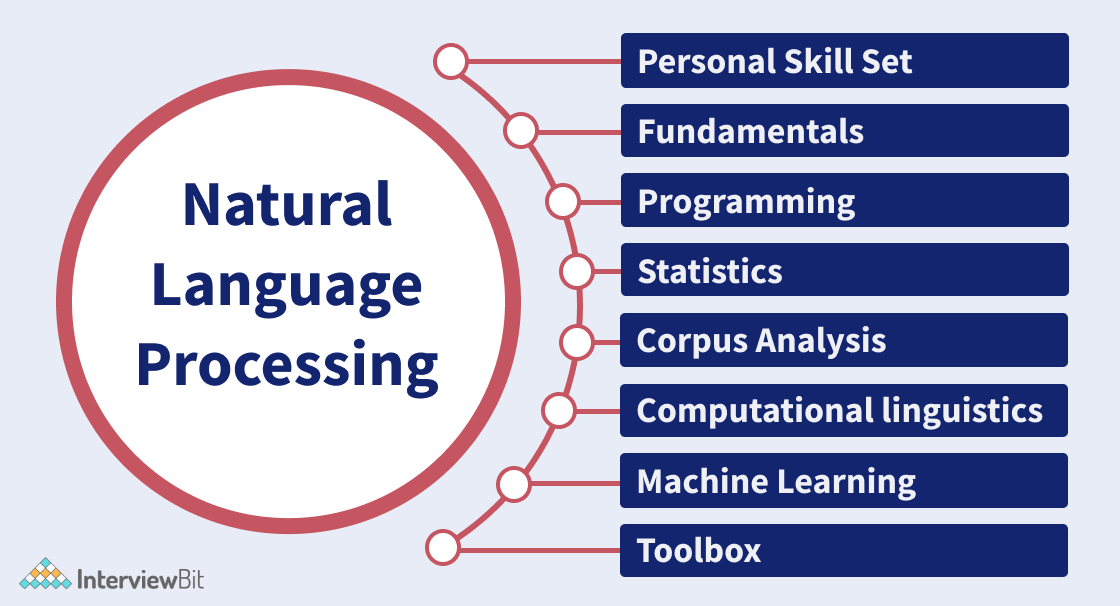
### ****12. How many steps of NLP is there?****

The five phases of NLP involve lexical (structure) analysis, parsing, semantic analysis, discourse integration, and pragmatic analysis.

### **Introduction to Natural Language Processing (NLP)**

Natural language processing (NLP) is a branch of linguistics, computer science, and **[artificial intelligence](https://www.interviewbit.com/artificial-intelligence-interview-questions/" \t "https://www.interviewbit.com/nlp-interview-questions/_blank)** that studies how computers interact with human language, particularly how to design computers to process and analyse massive amounts of natural language data. Google Assistant from Google, Siri speech assistance from Apple, are examples of Natural language processing.

Computational linguistics—rule-based human language modelling—is combined with statistical, **[machine learning](https://www.interviewbit.com/machine-learning-interview-questions/" \t "https://www.interviewbit.com/nlp-interview-questions/_blank)**, and **[deep learning](https://www.interviewbit.com/deep-learning-interview-questions/" \t "https://www.interviewbit.com/nlp-interview-questions/_blank)** models in NLP. These technologies, when used together, allow computers to process human language in the form of text or speech data and 'understand' its full meaning, including the speaker's or writer's intent and sentiment. NLP is used to power computer programmes that translate text from one language to another, respond to spoken commands, and quickly summarise vast amounts of material—even in real-time. Voice-activated GPS systems, digital assistants, speech-to-text dictation software, customer care chatbots, and other consumer conveniences are all examples of NLP in action. However, NLP is increasingly being used in corporate solutions to help businesses streamline operations, boost employee productivity, and streamline important business processes.



**Crack your next tech interview with confidence!**

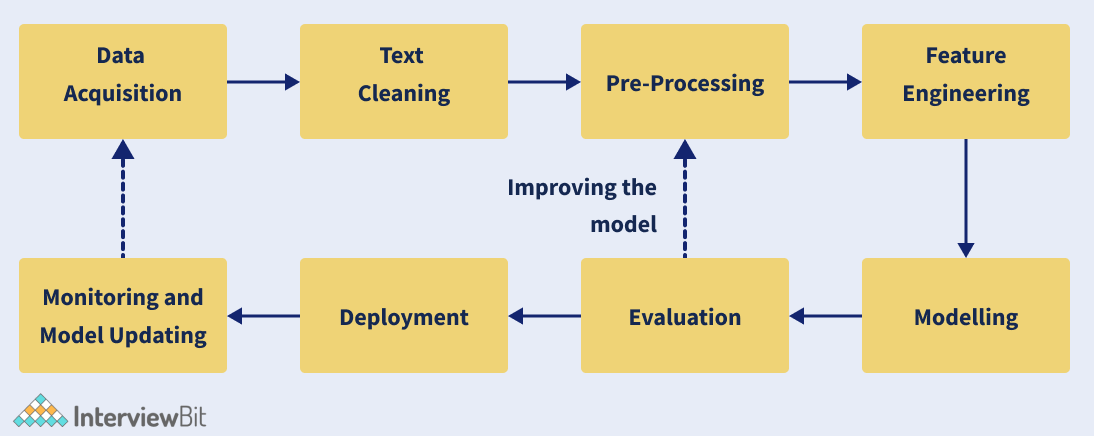
Take a free mock interview, get instant⚡️ feedback and recommendation

**Attempt Now**

## **NLP Interview Questions for Freshers**

### **1. What are the stages in the lifecycle of a natural language processing (NLP) project?**

Following are the stages in the lifecycle of a natural language processing (NLP) project:

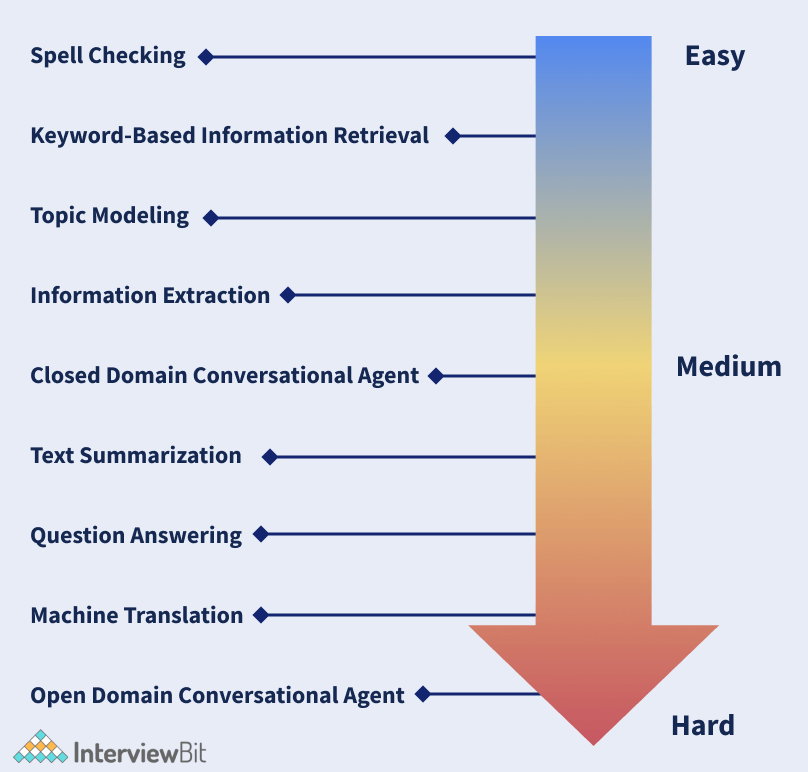


* ****Data Collection:**** The procedure of collecting, measuring, and evaluating correct insights for research using established approved procedures is referred to as data collection.
* ****Data Cleaning:**** The practice of correcting or deleting incorrect, corrupted, improperly formatted, duplicate, or incomplete data from a dataset is known as data cleaning.
* ****Data Pre-Processing:**** The process of converting raw data into a comprehensible format is known as data preparation.
* ****Feature Engineering:**** Feature engineering is the process of extracting features (characteristics, qualities, and attributes) from raw data using domain expertise.
* ****Data Modeling:**** The practice of examining data objects and their relationships with other things is known as data modelling. It's utilised to look into the data requirements for various business activities.
* ****Model Evaluation:**** Model evaluation is an important step in the creation of a model. It aids in the selection of the best model to represent our data and the prediction of how well the chosen model will perform in the future.
* ****Model Deployment:**** The technical task of exposing an ML model to real-world use is known as model deployment.
* ****Monitoring and Updating:**** The activity of measuring and analysing production model performance to ensure acceptable quality as defined by the use case is known as machine learning monitoring. It delivers alerts about performance difficulties and assists in diagnosing and resolving the core cause.

### **2. What are some of the common NLP tasks?**

Some of the common tasks of NLP include:

* ****Machine Translation:****This helps in translating a given piece of text from one language to another.
* ****Text Summarization:**** Based on a large corpus, this is used to give a short summary that gives an idea of the entire text in the document.
* ****Language Modeling:**** Based on the history of previous words, this helps uncover what the further sentence will look like. A good example of this is the auto-complete sentences feature in Gmail.
* ****Topic Modelling:****This helps uncover the topical structure of a large collection of documents. This indicates what topic a piece of text is actually about.
* ****Question Answering:****This helps prepare answers automatically based on a corpus of text, and on a question that is posed.
* ****Conversational Agent:**** These are basically voice assistants that we commonly see such as Alexa, Siri, Google Assistant, Cortana, etc.
* ****Information Retrieval:**** This helps in fetching relevant documents based on a user’s search query.
* ****Information Extraction:****This is the task of extracting relevant pieces of information from a given text, such as calendar events from emails.
* ****Text Classification:****This is used to create a bucket of categories of a given text, based on its content. This is used in a wide variety of AI-based applications such as sentiment analysis and spam detection.



Common NLP Tasks in order of Difficulty

### **3. What are the different approaches used to solve NLP problems?**

There are multiple approaches to solving NLP problems. These usually come in 3 categories:

* Heuristics
* Machine learning
* Deep Learning

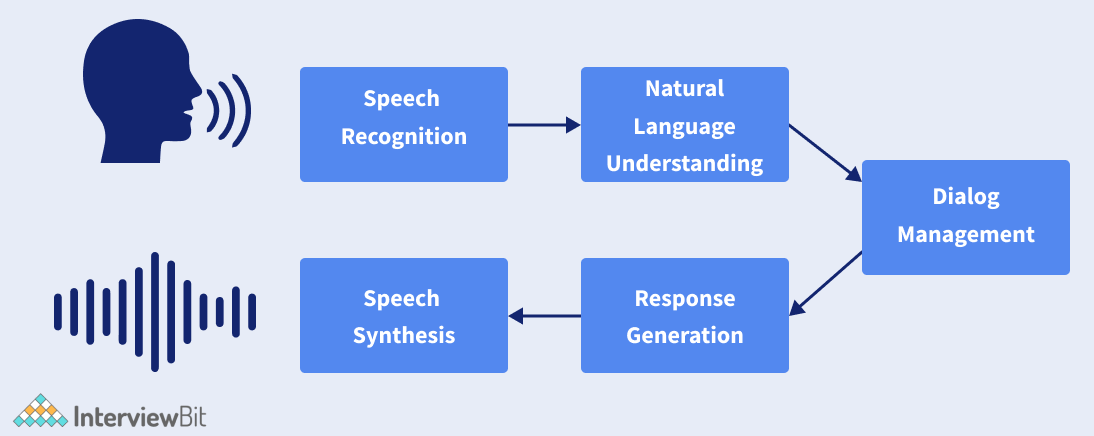
**You can download a PDF version of Nlp Interview Questions.**

**[IMG_259Download PDF](javascript:void(0))**

### **4. How do Conversational Agents work?**

The following NLP components are used in Conversational Agents:

* ****Speech Recognition and Synthesis:**** In the first stage, speech recognition helps convert speech signals to their phonemes, and are then transcribed as words.
* ****Natural Language Understanding (NLU):**** Here, the transcribed text from stage one is further analysed through AI techniques within the natural language understanding system. Certain NLP tasks such as Named Entity Recognition, Text Classification, Language modelling, etc. come into play here.
* ****Dialog Management:**** Once the needed information from text is extracted, we move on to the stage of understanding the user’s intent. The user’s response can then be classified by using a text classification system as a pre-defined intent. This helps the conversational agent in figuring out what is actually being asked.
* ****Generating Response:**** Based on the above stages, the agent generates an appropriate response that is based on a semantic interpretation of the user’s intent.



### **5. What is meant by data augmentation? What are some of the ways in which data augmentation can be done in NLP projects?**

NLP has some methods through which we can take a small dataset and use that in order to create more data. This is called data augmentation. In this, we use language properties to create text that is syntactically similar to the source text data.

Some of the ways in which data augmentation can be done in NLP projects are as follows:

* Replacing entities
* TF-IDF–based word replacement
* Adding noise to data
* Back translation
* Synonym replacement
* Bigram flipping

### **6. How can data be obtained for NLP projects?**

There are multiple ways in which data can be obtained for NLP projects. Some of them are as follows:

* ****Using publicly available datasets:****Datasets for NLP purposes are available on websites like Kaggle as well as Google Datasets.
* ****By using data augmentation:**** These are used to create additional datasets from existing datasets.
* ****Scraping data from the web:**** Using coding in Python or other languages once can scrape data from websites that are usually not readily available in a structured form.

### **7. What do you mean by Text Extraction and Cleanup?**

The process of extracting raw text from the input data by getting rid of all the other non-textual information, such as markup, metadata, etc., and converting the text to the required encoding format is called ****text extraction and cleanup****. Usually, this depends on the format of available data for the required project.

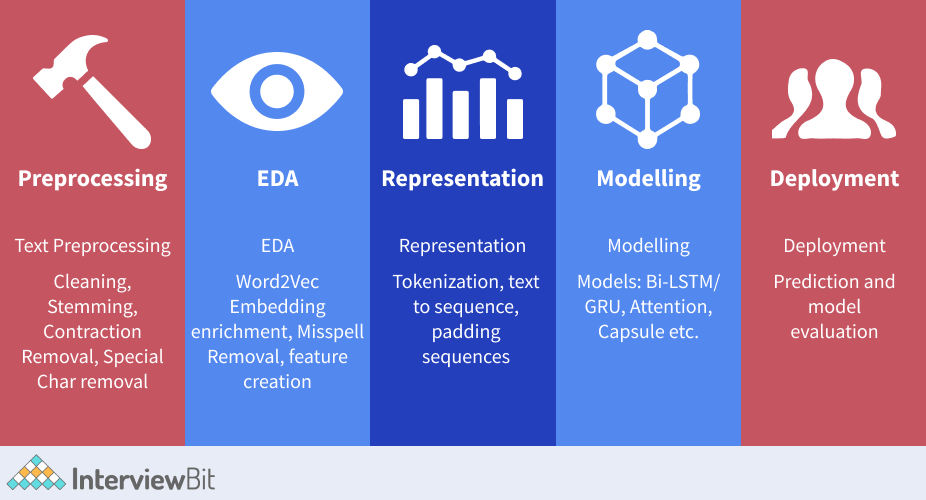
Following are the common ways used for Text Extraction in NLP:

* Named Entity Recognition
* Sentiment Analysis
* Text Summarization
* Aspect Mining
* Topic Modeling

### **8. What are the steps involved in preprocessing data for NLP?**

Here are some common pre-processing steps used in NLP software:

* ****Preliminaries:**** This includes word tokenization and sentence segmentation.
* ****Common Steps:**** Stop word removal, stemming and lemmatization, removing digits/punctuation, lowercasing, etc.
* ****Processing Steps:**** Code mixing, normalization, language detection, transliteration, etc.
* ****Advanced Processing:**** Parts of Speech (POS) tagging, coreference resolution, parsing, etc.



### **9. What do you mean by Stemming in NLP?**

When we remove the suffixes from a word so that the word is reduced to its base form, this process is called stemming. When the word is reduced to its base form, all the different variants of that word can be represented by the same form (e.g., “bird” and “birds” are both reduced to “bird”).

We can do this by using a fixed set of rules. For instance:  if a word ends in “-es,” we can remove the “-es”).

Even though these rules might not really make sense as a linguistically correct base form, stemming is usually carried out to match user queries in search engines to relevant documents. And in text classification, is done to reduce the feature space to train our machine learning (ML) models.

The code snippet given below depicts the way to use a well known NLP algorithm for stemming called Porter Stemmer using NLTK:

from nltk.stem.porter import PorterStemmer

stemmer = PorterStemmer()

word1, word2 = "bikes", "revolution"

print(stemmer.stem(word1), stemmer.stem(word2))

This gives “bike” as the stemmed version for “bikes,” but “revolut” as the stemmed form of “revolution,” even though the latter is not linguistically correct. Even if this might not affect the performance of the search engine, a derivation of the correct linguistic form becomes useful in some other cases. This can be done by another process that is closer to stemming, known as lemmatization.

### **10. What do you mean by Lemmatization in NLP?**

The method of mapping all the various forms of a word to its base word (also called “lemma”) is known as Lemmatization. Although this may appear close to the definition of stemming, these are actually different. For instance, the word “better,” after stemming, remains the same. However, upon lemmatization, this should become “good,”. Lemmatization needs greater linguistic knowledge. Modelling and developing efficient lemmatizers still remains an open problem in NLP research.

The application of a lemmatizer based on WordNet from NLTK is shown in the code snippet below:

from nltk.stem import WordNetLemmatizer

lemmatizer = WordnetLemmatizer()

print(lemmatizer.lemmatize("better", pos="a")) #a is for adjective

## **NLP Interview Questions for Experienced**

### **11. What is the meaning of Text Normalization in NLP?**

Consider a situation in which we’re operating with a set of social media posts to find information events. Social media textual content may be very exceptional from the language we’d see in, say, newspapers. A phrase may be spelt in multiple ways, such as in shortened forms, (for instance, with and without hyphens), names are usually in lowercase, and so on. When we're developing NLP tools to work with such kinds of data, it’s beneficial to attain a canonical representation of textual content that captures these kinds of variations into one representation. This is referred to as text normalization.

Converting all text to lowercase or uppercase, converting digits to text (e.g., 7 to seven), expanding abbreviations, and so on are some frequent text normalisation stages.

### **12. Explain the concept of Feature Engineering.**

After a variety of pre-processing procedures and their applications, we need a way to input the pre-processed text into an NLP algorithm later when we employ ML methods to complete our modelling step. The set of strategies that will achieve this goal is referred to as feature engineering. Feature extraction is another name for it. The purpose of feature engineering is to convert the text's qualities into a numeric vector that NLP algorithms can understand. This stage is called "text representation".

### **13. What is an ensemble method in NLP?**

An ensemble approach is a methodology that derives an output or makes predictions by combining numerous independent similar or distinct models/weak learners. An ensemble can also be created by combining various models such as random forest, SVM, and logistic regression.

Bias, variance, and noise, as we all know, have a negative impact on the mistakes and predictions of any machine learning model. Ensemble approaches are employed to overcome these drawbacks.

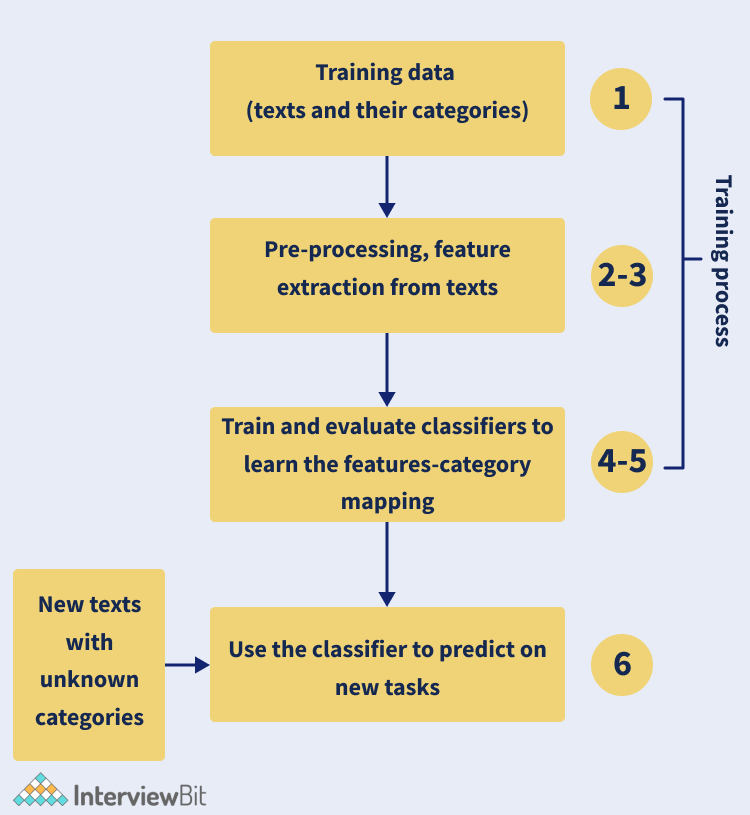
### **14. What do you mean by TF-IDF in Natural language Processing?**

TF-IDF also called ****Term Frequency-Inverse Document Frequency**** helps us get the importance of a particular word relative to other words in the corpus. It's a common scoring metric in information retrieval (IR) and summarization. TF-IDF converts words into vectors and adds semantic information, resulting in weighted unusual words that may be utilised in a variety of NLP applications.

### **15. What are the steps to follow when building a text classification system?**

When creating a text classification system, the following steps are usually followed:

* Gather or develop a labelled dataset that is appropriate for the purpose.
* Decide on an evaluation metric after splitting the dataset into two (training and test) or three parts: training, validation (i.e., development), and test sets (s).
* Convert unprocessed text into feature vectors.
* Utilize the feature vectors and labels from the training set to train a classifier.
* Benchmark the model's performance on the test set using the evaluation metric(s) from Step 2.
* Deploy the model and track its performance to serve a real-world use case.



### **16. Explain how parsing is done in NLP.**

Parsing is the process of identifying and understanding a text's syntactic structure. It is accomplished by examining the text's constituent pieces. The machine parses each word one by one, then two by two, three by three, and so on. It's a unigram when the system parses the text one word at a time. A bigram is a text that is parsed two words at a time. When the machine parses three words at a time, the set of words is called a ****trigram****.

The following points will help us comprehend the importance of parsing in NLP:

* Any syntax errors are reported by the parser.
* It aids in the recovery of often occurring errors so that the remainder of the programme can be processed.
* A parser is used to generate the parse tree.
* The parser is used to construct a symbol table, which is crucial in NLP.
* In addition, a Parser is utilised to generate intermediate representations (IR).

### **17. What do you mean by a Bag of Words (BOW)?**

The ****Bag of Words**** model is a popular one that uses word frequency or occurrences to train a classifier. This methodology generates a matrix of occurrences for documents or phrases, regardless of their grammatical structure or word order.

A bag-of-words is a text representation that describes the frequency with which words appear in a document. It entails two steps:

* A list of terms that are well-known.
* A metric for determining the existence of well-known terms.

Because any information about the sequence or structure of words in the document is deleted, it is referred to as a "bag" of words. The model simply cares about whether or not recognised terms appear in the document, not where they appear.

### **18. What do you mean by Parts of Speech (POS) tagging in NLP?**

A Part-Of-Speech Tagger (POS Tagger) reads the text in a language and assigns parts of speech to each word (and other tokens), such as noun, verb, adjective, and so on.

To label terms in text bodies, PoS taggers employ an algorithm. With tags like "noun-plural" or even more complicated labels, these taggers create more complex categories than those stated as basic PoS.

### **19. What is Latent Semantic Indexing (LSI) in NLP?**

****Latent Semantic Indexing**** (LSI), also known as Latent Semantic Analysis, is a mathematical method for improving the accuracy of information retrieval. It aids in the discovery of hidden(latent) relationships between words (semantics) by generating a set of various concepts associated with the terms of a phrase in order to increase information comprehension. Singular value decomposition is the NLP technique utilised for this aim. It's best for working with small groups of static documents.

### **20. What is the difference between NLP and NLU?**

| **Natural Language Processing (NLP)** | **Natural Language Understanding (NLU)** |
| --- | --- |
| NLP is a system that manages end-to-end conversations between computers and people at the same time. | NLU aids in the solving of Artificial Intelligence's complex problems. |
| Humans and machines are both involved in NLP. | NLU allows machines to interpret unstructured inputs by transforming them into structured text. |
| NLP focuses on interpreting language in its most literal sense, such as what was said. | NLU, on the other hand, concentrates on extracting context and meaning, or what was meant. |
| NLP can parse text-based on grammar, structure, typography, and point of view. | It'll be NLU that helps the machine deduce the meaning behind the language content. |

### **21. What are some metrics on which NLP models are evaluated?**

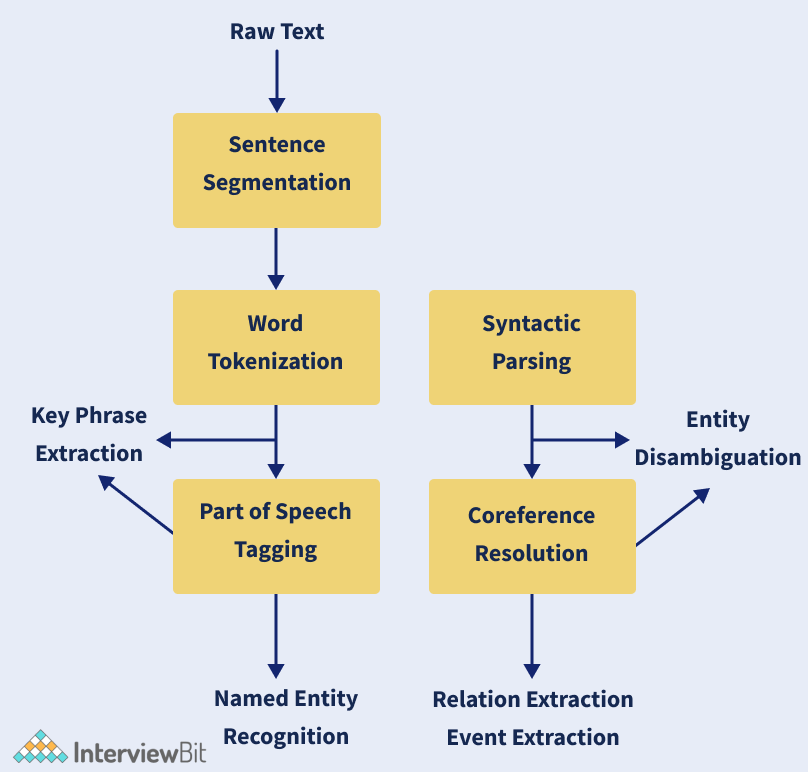
The following are some metrics on which NLP models are evaluated:

* ****Accuracy:****When the output variable is categorical or discrete, accuracy is used. It is the percentage of correct predictions made by the model compared to the total number of predictions made.
* ****Precision:**** Indicates how precise or exact the model's predictions are, i.e., how many positive (the class we care about) examples can the model correctly identify given all of them?
* ****Recall:**** Precision and recall are complementary. It measures how effectively the model can recall the positive class, i.e., how many of the positive predictions it generates are correct.
* ****F1 score:****This metric combines precision and recall into a single metric that also represents the trade-off between accuracy and recall, i.e., completeness and exactness.  
  (2 Precision Recall) / (Precision + Recall) is the formula for F1.
* ****AUC:****As the prediction threshold is changed, the AUC captures the number of correct positive predictions versus the number of incorrect positive predictions.

### **22. Explain the pipeline for Information extraction (IE) in NLP.**

In comparison to text classification, the typical pipeline for IE necessitates more fine-grained NLP processing. For example, we'd need to know the part-of-speech tags of words to identify named entities (people, organisations, etc.). We would require coreference resolution to connect various references to the same entity (e.g., Albert Einstein, Einstein, the scientist, he, etc.). It's worth noting that none of these stages are required for creating a text classification system. As a result, IE is a more NLP-intensive operation than text categorization. Not all steps in the pipeline are required for all IE jobs, as shown in the diagram, and the figure shows which IE tasks necessitate which degrees of analysis.

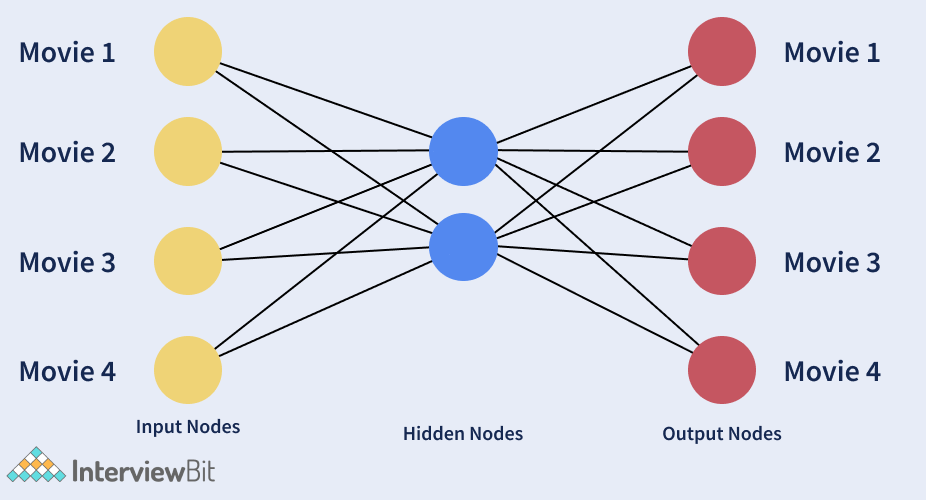
Other than named entity recognition, all other IE tasks require deeper NLP pre-processing followed by models developed for those specific tasks. Key phrase extraction is the task that requires the least amount of NLP processing (some algorithms also do POS tagging before extracting key phrases), whereas all other IE tasks require deeper NLP pre-processing followed by models developed for those specific tasks. Standard evaluation sets are often used to assess IE tasks in terms of precision, recall, and F1 scores. Because of the various levels of NLP pre-processing required, the accuracy of these processing steps has an impact on IE jobs. All of these factors should be considered when collecting relevant training data and, if necessary, training our own models for IE.



### **23. What do you mean by Autoencoders?**

A network that is used for learning a vector representation of the input in a compressed form, is called an autoencoder. It is a type of unsupervised learning since labels aren’t needed for the process. This is mainly used to learn the mapping function from the input. In order to make the mapping useful, the input is reconstructed from the vector representation. After training is complete, the vector representation that we get helps encode the input text as a dense vector. Autoencoders are generally used to make feature representations.

In the figure below, the hidden layer depicts a compressed representation of the source data that captures its essence. The input representation is reconstructed by the output layer called the decoder.



### **24. What do you mean by Masked language modelling?**

Masked language modelling is an NLP technique for extracting the output from a contaminated input. Learners can use this approach to master deep representations in downstream tasks. Using this NLP technique, you may predict a word based on the other words in the sentence.

The following is the process for Masked language modelling:

* Our text is tokenized. We start with text tokenization, just as we would with transformers.
* Make a tensor of labels. We're using a labels tensor to calculate loss against — and optimise towards — as we train our model.
* Tokens in input ids are masked. We can mask a random selection of tokens now that we've produced a duplicate of input ids for labels.
* Make a loss calculation. We use our model to process the input ids and labels tensors and determine the loss between them.

### **25. What is the meaning of Pragmatic Analysis in NLP?**

Pragmatic Analysis is concerned with outside word knowledge, which refers to information that is not contained in the documents and/or questions. The many parts of the language that require real-world knowledge are derived from a pragmatics analysis that focuses on what was described and reinterpreted by what it truly meant.

### **26. What is the meaning of N-gram in NLP?**

Text N-grams are commonly used in text mining and natural language processing. They're essentially a collection of co-occurring words within a specific frame, and when computing the n-grams, you usually advance one word (although you can move X words forward in more advanced scenarios).

### **27. What do you mean by perplexity in NLP?**

It's a statistic for evaluating the effectiveness of language models. It is described mathematically as a function of the likelihood that the language model describes a test sample. The perplexity of a test sample X = x1, x2, x3,....,xn is given by,

PP(X)=P(x1,x2,…,xN)-1N

The total number of word tokens is N.

The more perplexing the situation, the less information the language model conveys.

### **Conclusion**

One of the most important reasons for NLP is that it allows computers to converse with people in natural language. Other language-related activities are also scaled. Computers can now hear, analyse, quantify, and identify which parts of speech are significant thanks to Natural Language Processing (NLP). NLP has a wide range of applications, including chatbots, sentiment analysis, and market intelligence. Since its introduction, NLP has grown in popularity. Today, devices like Amazon's Alexa are extensively used all over the world. And, for businesses, business intelligence and consumer monitoring are quickly gaining traction and will soon rule the industry.

### 1) What is the full form of NLP? / What is Natural Language Processing?

NLP stands for ****"Natural Language Processing".**** NLP is a field of computer science that deals with communication between computer systems and humans. This technique uses Artificial Intelligence and Machine Learning to create automated software that helps understand the human spoken languages and extract useful information from the data gathered from the audio.

The techniques used in NLP allow computer systems to process and interpret data in the form of natural languages. It designs algorithms that can extract meaning from large datasets in audio or text format by applying machine learning algorithms. In other words, we can say that NLP is software that uses artificial intelligence and machine learning algorithms to understand natural languages or the way human beings read and write in a language and extracts required information from such data.

### 2) What are some real-life applications / real-world examples of Natural Language Processing (NLP)?

Some real-life applications of NLP or Natural Language Processing are as follows:

****Spelling/Grammar Checking Apps:**** Spelling and grammar checking applications are real-life examples of Natural Language Processing. These apps are mainly used in mobile applications and websites that facilitate users to correct grammar mistakes in the entered text rely on NLP algorithms. They also recommend the best possible substitutes that the users might type. This is possible because of specific NLP models being used in the backend.

****Google Translate:**** Google Translate is the most famous application of Natural Language Processing. Using this, you can convert your written or spoken sentences into any language. You can also get the correct pronunciation and meaning of a word by using Google Translate. The Google Translate application uses some advanced techniques of Natural Language Processing to provide translation of sentences into various languages.

****Chatbots apps:**** Chatbots applications provide a better customer support service. Many websites and companies use this to offer customer support through these virtual bots that chat with the user and resolve their problems. Many companies use chatbots for 24/7 service to resolve the basic queries of customers. Generally, it filters the basic issues that do not require an interaction with the companies' customer executives. It makes the customers feel that the customer support team quickly attends them. If a chatbot cannot resolve any user's query, it forwards it to the support team while still engaging the customer. Chatbots also make companies capable of building cordial relations with customers. These all are only possible because of Natural Language Processing.

### 3) What are the most used NLP (Natural Language Processing) Terminologies?

Following is the list of most used NLP (Natural Language Processing) Terminologies:

* ****Preprocessing:**** This is a method used to remove unwanted text or noise from the given text and make it "clean." It is the first step of any NLP task. s
* ****Documents:**** Documents are the body of text and are collectively used to form a corpus.
* ****Corpus, or Corpora (Plural):**** It is a collection of text of similar type, for example, movie reviews, social media posts, etc.
* ****Vocabulary:**** It is a group of terms used in a text or speech.
* ****Out of Vocabulary:**** It specifies the terms not included in the vocabulary. We put the terms created during the model's training in this category.
* ****Tokenization:**** It is used in NLP to break down large sets of text into small parts for easy readability and understanding. Here, the small parts are referred to as 'text' and provide a piece of meaningful information.
* ****N-grams:**** It specifies the continuous sequence (similar to the power set in number theory) of n-tokens of a given text.
* ****Parts of Speech (POS):**** It specifies the word's functions, such as a noun, verb, etc.
* ****Parts of Speech Tagging:**** It is the process of tagging words in the sentences into different parts of speech.

### 4) What are the most used NLP (Natural Language Processing) Terminologies?

Following is the list of most used NLP (Natural Language Processing) Terminologies:

* ****Embeddings (Word):**** This process is used to embed each token as a vector and then pass it into a machine learning model. We can apply embeddings also on phrases and characters, apart from words.
* ****Stop Words:**** These are used to remove the unwanted text from further text processing, for example, a, to, can, etc.
* ****Transformers:**** Transformers are deep learning architectures that can parallelize computations. They are used to learn long-term dependencies.
* ****Normalization:**** This is a process of mapping similar terms to a canonical form, i.e., a single entity.
* ****Lemmatization:**** Lemmatization is a type of normalization used to group similar terms to their base form according to their parts of speech. For example, talking and talking can be mapped to a single term, talk.
* ****Stemming:**** Stemming is also a type of normalization similar to lemmatization. But, it is different in the term that it segregates the words without the parts of speech tags. It is faster than lemmatization and also be more precise in some cases.

### 5) What are some of the major components of Natural Language Processing?

Following is a list of some of the major components of Natural Language Processing:

****Entity extraction:**** It is used for segmenting a sentence to identify and extract entities, such as a person (real or fictional), organization, geographies, events, etc. 85

****Pragmatic analysis:**** Pragmatic analysis extracts information from the input text. It is part of the process of data extraction.

****Syntactic analysis:**** Syntactic analysis is used for the proper ordering of words.

### 6) What do you understand by Dependency Parsing in NLP or Natural Language Processing?

In Natural Language Processing, Dependency Parsing is a process of assigning syntactic structure to a sentence and identifying its dependency parses. This is an important process to understand the correlations between the "head" words in the syntactic structure. That's why it is also known as syntactic parsing.

The process of dependency parsing becomes a little complex if there are more sentences that have more than one dependency parses. Multiple parse trees are known as ambiguities. The main task of dependency parsing is to resolve these ambiguities to assign a syntactic structure to a sentence effectively. It is also used in semantic analysis apart from syntactic structuring.

### 7) What are some most common areas of usage of Natural Language Processing?

Following is a list of some most common areas of usage of Natural Language Processing:

* Semantic Analysis
* Text classification
* Automatic summarization
* Questioning Answering

Some real-life examples of Natural Language Processing are chatbots, IOS Siri, Google Assistant, Amazon echo, Spelling, grammar checking apps, and Google translate.

### 8) What do you understand by NLTK in Natural Language Processing?

In Natural Language Processing, NLTK stands for Natural Language Toolkit. It is a Python library used to process data in human spoken languages. NLTK facilitates developers to apply parsing, tokenization, lemmatization, stemming techniques, and more to understand natural languages. It is also used for categorizing text, parsing linguistic structure, analyzing documents, etc.

Following is the list of some libraries of the NLTK package that are often used in NLP:

* DefaultTagger
* UnigramTagger
* RegexpTagger
* backoff\_tagger
* SequentialBackoffTagger
* UnigramTagger
* BigramTagger
* TrigramTagger
* treebank
* wordnet
* FreqDist
* Patterns etc.

### 9) What is the use of TF-IDF? Why is it used in Natural language Processing?

In Natural language Processing, tf-idf, TF-IDF, or TFIDF stands for Term Frequency-Inverse Document Frequency. It is a numerical statistic used to specify how important a word is to a document in a collection or the collection of a set.

### 10) What is the difference between formal and natural languages?

The main difference between a formal language and a natural language is that a formal language is a collection of strings. Each string contains symbols from a finite set called alphabets. On the other hand, a natural language is a language that humans use to speak. This is completely different from a formal language as it contains fragments of words and pause words like uh, um, etc.

### 11) What are the tools used for training NLP models?

The most common tools used for training NLP models are NLTK, spaCY, PyTorch-NLP, openNLP, etc.

### 12) What do you understand by information extraction? What are the various models of information extraction?

In Natural Language Processing, information extraction is a technique of automatically extracting structured information from unstructured sources to get useful information. It extracts information such as attributes of entities, the relationship between different entities, and more.

Following is a list of various models of information extraction in Natural Language Processing:

* Fact Extraction Module
* Entity Extraction Module
* Sentiment Analysis Module
* Tagger Module
* Relation Extraction Module
* Network Graph Module
* Document Classification and Language Modeling Module

### 13) What are the stop words in Natural Language Processing?

In Natural Language Processing, stop words are regarded as useless data for a search engine. It includes the words like articles, prepositions, was, were, is, am, the, a, an, how, why, and many more. The algorithm used in Natural Language Processing eliminates the stop words to understand and analyze the meaning of the sentences. Eliminating the stop words is one of the most important tasks for search engines to process data.

Software developers design the algorithms of search engines so that they ignore the use of stop words and only show the relevant search result for a query.

### 14) What is Bag of Words in Natural Language Processing?

Bag of Words is a commonly used model in Natural Language Processing that depends on word frequencies or occurrences to train a classifier. This model creates an occurrence matrix for documents or sentences without depending on their grammatical structure or word order.

### 15) What do you understand by semantic analysis? What are the techniques used for semantic analysis?

Semantic analysis is a process that makes a machine understand the meaning of a text. It uses several algorithms to interpret the words in sentences. It is also used to understand the structure of a sentence.

Following are the techniques used for semantic analysis:

****Named entity recognition:**** This technique is used to specify the process of information retrieval that helps identify the entities like the name of a person, organization, place, time, emotion, etc.

****Natural language generation:**** This technique specifies a process used by the software to convert the structured data into human spoken languages. By using natural language generation, organizations can automate content for custom reports.

****Word sense disambiguation:**** It technique is used to identify the sense of a word used in different sentences.

### 16) What is pragmatic ambiguity in NLP?

Pragmatic ambiguity is used to specify words with more than one meaning, and they can be used in any sentence depending on the context. In pragmatic ambiguity, words have multiple interpretations.

Pragmatic ambiguity occurs when the meaning of the words is not specific. For example, if a word gives different meanings. Due to pragmatic ambiguity, a sentence can have multiple interpretations. Sometimes, we come across sentences that have words with multiple meanings, making the sentence open to interpretation.

### 17) What is Latent Semantic Indexing (LSI)? What is the use of this technique?

LSI or Latent Semantic Indexing is a mathematical technique used in Natural Language Processing. This technique is used to improve the accuracy of the information retrieval process. The LSI algorithm is designed to allow machines to detect the latent correlation between semantics.

The machines generate various concepts to enhance information understanding. The technique used for information understanding is called singular value decomposition. It is mainly used to handle static and unstructured data. This is one of the best-suited models to identify components and group them according to their types.

Latent Semantic Indexing or LSI is based on a principle that specifies that words carry a similar meaning when used in a similar context. The computational LSI models are slow compared to other models, but they can improve a text or document's analysis and understanding.

### 18) What do you understand by MLM in Natural Language Processing?

In Natural Language Processing, MLM is a term that stands for Masked Language Model. It helps learners understand deep representations in downstream tasks by taking the output from the corrupt input.

This model is mainly used to predict the words used in a sentence.

### 19) What are the most commonly used models to reduce data dimensionality in NLP?

The most commonly used models to reduce the dimensionality of data in NLP are TF-IDF, Word2vec/Glove, LSI, Topic Modelling, Elmo Embeddings, etc.

### 20) What is Lemmatization in Natural Language Processing?

Lemmatization is a process of doing things properly using a vocabulary and morphological analysis of words. It is mainly used to remove the inflectional endings only and return the base or dictionary form of a word, known as the lemma. It is just like cutting down your beard or shaving to get the original shape of your face.

****For example:**** girl's = girl, bikes= bike, leaders= leader etc.

So, the main task of Lemmatization is to identify and return the root or original words of the sentence to explore various additional information.

### 21) What is Stemming in Natural Language Processing?

Stemming is a process of extracting the base form of a word by removing the affixes from them. It is just like cutting down the branches of a tree to its stems.

****For example:**** After stemming, the words go, goes, and going would be 'go'.

Search engines use stemming for indexing the words. It facilitates them to store only the stems rather than storing all forms of a word. By using stemming, the search engines reduce the size of the index and increase the retrieval accuracy.

### 22) What is the difference between Stemming and Lemmatization in NLP?

Stemming and Lemmatization are both the text normalization techniques used in Natural language Processing. Both are used to prepare text, words, and documents for further processing. They seem very similar techniques, but there are quite differences between them. Let's see the main differences between them:

|  |  |
| --- | --- |
| **Stemming** | **Lemmatization** |
| Stemming is the process of extracting the base form of a word by removing the affixes from them. It produces the morphological variants of a root/base word. Stemming programs are commonly known as stemming algorithms or stemmers. | Lemmatization is a more advanced process and looks beyond word reduction, just like stemming. It considers a full vocabulary of a language and applies a morphological analysis to words. For example, the lemma of 'went' is 'go', and the lemma of 'mice' is 'mouse'. |
| Stemming is not as much informative as Lemmatization. It is a somewhat crude method for cataloging related words. It essentially cuts letters from the end until the stem is reached. | Lemmatization is much more informative than simple Stemming; that is why ****Spacy**** has opted to only have Lemmatization available instead of Stemming. |
| Stemming is not as efficient as Lemmatization. This method works fairly well in most cases, but unfortunately, English has many exceptions requiring a more sophisticated process. | Lemmatization is more efficient than Stemming as it works well in exceptional words. |
| Following are some examples of Stemming: run: run runner: runner running: run ran: ran runs: run easily: easili fairly: fair etc. | Following are some examples of Lemmatization: run: run runner: run running: run ran: run runs: run goes: go go: go went: go saw: see mice: mouse |

### 23) Which NLP techniques use a lexical knowledge base to obtain the correct base form of the words?

The NLP techniques that use a lexical knowledge base to obtain the correct base form of the words are Lemmatization and stemming.

### 24) What is tokenization in Natural Language Processing?

In Natural Language Processing, tokenization is a method of dividing the text into various tokens. These tokens are the form of the words, just like a word forms into a sentence. In NLP, the program computers process large amounts of natural language data. These large amounts of natural language data have to be cut into shorter forms. So, tokenization is an important step in NLP that cuts the text into minimal units for further processing.

### 25) What are some open-source libraries used in NLP?

Some popular open-source libraries used in NLP are NLTK (Natural Language ToolKit), SciKit Learn, Textblob, CoreNLP, spaCY, Gensim, etc.

### 26) What are the key differences between NLP and NLU?

Following is the list of key differences between NLP and NLU:

|  |  |
| --- | --- |
| **NLP** | **NLU** |
| NLP is a short form of Natural Language Processing. | NLU is a short form of Natural Language Understanding. |
| NLP or Language Processing is used to create a system that can make and establish communication between humans and computers. | NLU or Natural Language Understanding provides techniques that can solve complicated problems related to machine understanding. |
| It includes all the techniques required for the interaction between computers and humans. | It converts the uncategorized input data into a structured format and allows the computers to understand the data. |
| It includes the techniques focused on analyzing "what is said?" | It includes the techniques to understand "what is meant?" |

### 27) What are the key differences between NLP (Natural Language Processing) and CI (Conversational Interface)?

Following is the list of key differences between NLP (Natural Language Processing) and CI (Conversational Interface):

|  |  |
| --- | --- |
| **Natural Language Processing (NLP)** | **Conversational Interface (CI)** |
| The full form of NLP is Natural Language Processing. | The full form of CI is Conversational Interface. |
| The main focus of NLP is to make computers understand and learn how the normal human being languages' concepts work. | The main and only focus of CI is to provide users with an interface to interact with. |
| Natural Language Processing uses AI technology to identify, understand, and interpret users' requests through languages. | CI or Conversational Interface uses voice, chat, videos, images, and other conversational aid to create the user interface for communication. |

### 28) What do you understand by Pragmatic Analysis?

Pragmatic analysis is an important task used in Natural Language Processing for interpreting knowledge lying outside a given document. It is mainly implemented to focus on exploring a different aspect of the document or text in a language. It requires a comprehensive knowledge of the real world to make software applications capable of critical interpretation of the real-world data to know the actual meaning of sentences and words.

****For example, see the following sentence:****

'Do you know what time it is?'

This sentence can be used to ask for knowing the time or for yelling at someone to make them note the time. It completely depends on the context in which this sentence is used.

### 29) What are the best open sources of NLP Tools available in the market?

Some of the best open sources NLP tools available in the market are:

* SpaCy
* TextBlob
* Textacy
* Natural language Toolkit (NLTK)
* Retext
* NLP.js
* Stanford NLP
* CogcompNLP etc.

### 30) How can you differentiate Artificial Intelligence, Machine Learning, and Natural Language Processing?

Following are the key differences between Artificial Intelligence, Machine Learning, and, Natural Language Processing:

|  |  |  |
| --- | --- | --- |
| **Artificial Intelligence** | **Machine Learning** | **Natural Language Processing** |
| Artificial Intelligence is a technique used to create smarter machines and computers. | Machine Learning is a term used for systems that learn from experience. | Natural Language Processing or NLP is the set of systems that can understand the languages used by humans and process these languages to make them understood by computers. |
| Artificial Intelligence requires human intervention. Without human intervention, it is not possible to create intelligent machines. | Machine Learning doesn't require human intervention. It purely involves the working of computers and machines. | Natural Language Processing uses both computer and human languages to work properly. |
| Artificial Intelligence is a broader concept than Machine Learning. It includes a lot of working fields. | Machine Learning is a narrow concept and is a subset of Artificial Intelligence. | Natural Language Processing uses the concept of both Artificial Intelligence and Machine Learning to make the tools that can process human language and make it understandable by machines. |

### 31) What do you understand by POS tagging?

The full form of POS tagging is Parts of speech tagging. It is most commonly known as POS tagging. According to their context, it specifies a process of identifying specific words in a document and groups them as part of speech.

POS tagging is also known as grammatical tagging because it involves understanding grammatical structures and identifying the respective component. It is a very complicated process because the same word can be different parts of speech depending on the situation and the structure of the sentence.

### 32) What is NES in Natural Language Processing? Why is it used?

NES is an acronym that stands for Name Entity Recognition. It is used in Natural Language Processing and is most commonly known as NER. It is the process of identifying specific entities in a text document that is more informative and have a unique context. It includes places, people, organizations, and more. After identification, it extracts these entities and categorizes them under different predefined classes. Later, this step helps in extracting information.

### 33) What is parsing in Natural Language Processing? What are the different types of parsing used in NLP?

Parsing is a technique or a method of analyzing the sentences automatically according to their syntactic structure.

Following is a list of different types of parsing used in Natural Language Processing:

****Dependency parsing / Syntactic parsing:**** Dependency parsing is also known as syntactic parsing. It recognizes a dependency parse of a sentence and assigns a syntax structure to the sentence. It mainly focuses on the relationship between different words.

****Semantic parsing:**** Semantic parsing is a method of converting the natural language into machine language that a computer can understand and process.

****Constituency parsing:**** Constituency parsing is a specific parsing method where a division of sentences is divided into sub-parts or constituencies. It is mainly used to extract a constituency-based parse tree from the constituencies of the sentences.

****Shallow parsing / Light parsing:**** Shallow parsing is also known as light parsing and chunking. It identifies constituents of sentences and then links them to different groups of grammatical meanings.

### 34) What is language modeling in NLP?

In Natural Language Processing, language modeling creates a probability distribution of a sequence of words. It provides probability to all the words present in that sequence.

### 35) What is topic modeling in NLP?

In NLP, topic modeling is finding abstract topics in a document or set of documents to find hidden semantic structures.

### 36) What is the key difference between dependency parsing and shallow parsing?

The key difference between dependency parsing and shallow parsing is that dependency parsing is the process of finding a relation between all the different words. On the other hand, shallow parsing is the parsing of a selected limited part of the information.

### 37) What do you understand by Pragmatic Ambiguity in NLP?

In Natural Language Processing, pragmatic ambiguity specifies multiple descriptions of a word or a sentence. It occurs when the words of the sentence may have different meanings, and the correct meaning of the sentence is not clear. In this case, it becomes very difficult for a machine to understand a sentence's meaning, which causes pragmatic ambiguity.

****For example, see the following sentence:****

****"Are you feeling hungry?"****

The above sentence could be either a generally asked question or a formal way of offering food.

### 38) What are the steps used to solve an NLP problem?

Following is a list of steps used to solve an NLP problem:

* In the first step, get the text from the available dataset.
* Now, apply stemming and lemmatization to clean the text.
* Now, apply feature engineering techniques to the received text.
* Embed using word2vec.
* Now, train the built model using neural networks or other Machine Learning techniques.
* Now it turns to evaluate the model's performance.
* Make the appropriate changes in the model.
* Now, the model is complete. Deploy the model.

### 39) What is noise removal in NLP? Why is it used?

Noise removal is one of the NLP techniques. As the name specifies, it is used to remove unnecessary pieces of text from the sentences.

### **1. What are the natural language processing (NLP) project's lifecycle stages?**

The stages of a natural language processing (NLP) project's lifespan are as follows:

* ****Data collection:**** Refers to the process of gathering, measuring, and assessing accurate insights for study following defined, recognised techniques.
* ****Data Cleaning:**** The process of removing or repairing inaccurate, corrupted, incorrectly formatted, duplicate, or incomplete data from a dataset is called data cleaning.
* ****Data Pre-Processing:**** Data preparation is transforming raw data into a usable format. Feature Engineering extracts the characteristics, traits, and attributes from raw data.
* ****Data modelling:****Studying data items and their connections with other objects. It is used to investigate data requirements for a variety of commercial operations.
* ****Model Evaluation:**** A crucial phase in building a model is model evaluation. It aids in selecting the optimal model to describe our data and predict how well that model will function in the future.
* ****Model Deployment:**** Process of making an ML model available for real-world application.
* ****Monitoring and Updating:**** Evaluate and analyse production model performance to achieve acceptable quality defined by the use case. It sends out notifications when there are problems with performance and aids in detecting and treating the root cause.

### **2. What are some examples of typical NLP tasks?**

NLP is used to do various tasks, including

* ****Machine Translation:**** This assists in translating a text from one language to another.
* ****Text Summarization:**** This is used to generate a concise summary of the complete text in the document based on a significant corpus.
* ****Language modelling:**** This determines how the following phrase will seem based on the history of preceding words. The auto-complete sentences tool in Gmail is an excellent illustration of this.
* ****Topic modelling:**** A technique for determining the subject organisation of many documents. It identifies the actual topic of a piece of writing.
* ****Query Answering:**** This aids in automatically preparing replies based on a corpus of text and an asked question.
* ****Conversational Agent:****These are voice assistants like Alexa, Siri, Google Assistant, Cortana, and others that we see all the time.
* ****Information Retrieval:**** This aids in retrieving relevant documents in response to a user's search query.
* ****Information Extraction:**** Extraction of useful information from a text, such as email calendar events, is known as information extraction.
* ****Text Classification:**** This is used to categorise a given text into a set of categories based on its content. It is also utilised in many AI-based applications, including sentiment analysis and spam identification.

### **3. How do Conversational Agents work?**

Conversational Agents use the following NLP components:

* ****Voice Recognition and Synthesis:**** Speech recognition aids in converting speech impulses to phonemes, which are subsequently transcribed as words.
* ****Natural Language Understanding (NLU):**** The transcribed text from stage one is analysed using AI algorithms within the natural language understanding system. Named Entity Recognition, Text Classification, Language Modeling, and other NLP tasks are relevant here.
* ****Management of Conversations:**** After extracting the necessary data from text, we proceed to the stage of determining the user's purpose. The user's response can then be categorised as a pre-defined intent using a text classification system. This assists the conversational agent in determining what is being asked.
* ****Generating Answer:**** The agent creates an appropriate response based on the initial phases' semantic understanding of the user's purpose.

|  |
| --- |
| If you would like to become a Python certified professional, then visit Mindmajix - A Global online training platform:  “**[Python Training](https://mindmajix.com/python-training" \o "Python Training" \t "https://mindmajix.com/_blank)**”  Course.  This course will help you to achieve excellence in this domain. |

### **4. What does data augmentation imply? What are some examples of data augmentation techniques used in NLP projects?**

NLP offers various ways for taking a small dataset and combining it with other data to build larger datasets. Data augmentation is the term for this. Language attributes are used to generate text that is syntactically comparable to the original text data.

The following are some examples of how data augmentation may be used in NLP projects:

* Entities are being replaced.
* Word substitution based on the TF-IDF
* Back data translation with noise
* Synonym replacement
* Bigram flipping

### **5. What procedures should you take while creating a text categorisation system?**

The following stages are commonly performed while establishing a text categorisation system:

1. Gather or create a labelled dataset suitable for the task.

2. After separating the dataset into two (training and test) or three sections (training, validation (i.e., development), and test sets), choose an assessment measure (s).

3. Unprocessed text is converted into feature vectors.

4. To train a classifier, use the feature vectors and labels from the training set.

5. Using the evaluation metric(s) from Step 2, compare the model's performance on the test set.

6. To service a real-world use case, deploy the model and track its performance to service a real-world use case.

### **6. Describe how parsing is done in NLP?**

The process of recognising and comprehending a text's syntactic structure is parsing. It's done by dissecting the text's fundamental parts. Each word is parsed one at a time, then two at a time, three at a time. When the machine parses the text one word at a time, it's called a unigram. A bigram is a text in which two words are analysed at a time. A trigram is a group of three words that the computer parses simultaneously.

The following points will help us understand why parsing is vital in NLP:

* The parser will report any syntax mistakes.
* It assists in the recovery of frequently recurring mistakes, allowing the rest of the programme to be processed.
* The parse tree is created using a parser.
* The parser creates a symbol table, vital in natural language processing.
* In addition, to construct intermediate representations, a Parser is used (IR)

### **7. What exactly is a "Bag of Words" (BOW)?**

The Bag of Words model, which employs word frequency or occurrences to train a classifier, is a popular one. This approach creates a matrix of occurrences for texts or phrases, regardless of grammatical structure or word order.

A bag-of-words is a text representation that indicates how often words appear in a document. It consists of two steps:

A glossary of well-known terms—a metric for determining if they exist.

The document is a "bag" of words since all information about the order or organisation of words is removed. The model is just concerned with whether or not recognised terms occur in the document, not with the place of those phrases.

### **8. What is Regular Grammar?**

Regular grammar represents a common language.

A -> a, A -> aB, and many additional rules exist in regular grammar. The rules allow for the identification and analysis of strings to be automated.

In regular grammar, there are four tuples:

* ‘N’ represents the non-terminal set.
* ‘∑’ represents the set of terminals.
* ‘P’ stands for the set of productions.
* ' € N’ denotes the start of non-terminal.

### **9. What is Latent Semantic Indexing (LSI) in Natural recovering using?**

The mathematical approach of Latent Semantic Indexing (LSI), also known as Latent Semantic Analysis, is used to improve the accuracy of information retrieval. It facilitates the finding of hidden (latent) associations between words (semantics) by developing a collection of various ideas connected with a phrase's terms to improve information understanding. The NLP approach used for this is singular value decomposition. It works well with small sets of static content.

### **10. What are some of the measures used to assess NLP models?**

The following are some of the measures used to evaluate NLP models:

* Accuracy is utilised when the output variable is categorical or discrete. It is the model's proportion of right predictions relative to the total number of forecasts.
* The precision parameter indicates how accurate or exact the model's predictions are, i.e., how many positive (the class we're interested in) instances can the model properly identify.
* F1 score: This measure combines precision and recalls into a single metric that indicates the accuracy-recall trade-off, i.e., completeness and exactness.
* The formula for F1 is (2 Precision Recall) / (Precision + Recall).
* The AUC measures the number of correct positive predictions vs the number of wrong positive predictions as the prediction threshold is modified.

### **11. What is Pragmatic Analysis, exactly?**

In NLP, pragmatic analysis is a crucial job for understanding knowledge that exists outside of a given document. Using pragmatic analysis aims to concentrate on a specific component of a document or text in a language. This necessitates a thorough understanding of the real world. The pragmatic analysis helps software programmes to know the true meaning of phrases and words through critical interpretation of real-world data.

### **12. How can data for NLP projects be obtained?**

There are several methods for obtaining data for NLP projects. The following are a few:

****Using publicly accessible datasets:**** Datasets for NLP may be found on sites such as Kaggle and Google Datasets.

****Using data augmentation:****This technique produces new datasets from current ones.

Scraping data from the web: Using Python or other programming languages, one may scrape data from websites that aren't generally available in an organised format.

### **13. What does Text Extraction and Cleanup imply?**

Text extraction and cleaning is the process of extracting raw text from input data while removing all non-textual information such as markup, metadata, and other non-textual information and converting the text to the needed encoding type. This is usually determined by the format of the available data for the project.

The following are some of the most frequent methods for text extraction in NLP.

* Sentiment Analysis for Named Entity Recognition
* Summarization of Text
* Topic Modeling using Aspect Mining

### **14. What actions are required in resolving an NLP issue?**

The steps for addressing an NLP problem are as follows:

* Obtain the text via scraping the web or using the provided dataset.
* For text cleaning, use stemming and lemmatisation.
* Feature engineering strategies should be used.
* Use word2vec to embed
* Neural networks, or other Machine Learning techniques can train the created model.
* Examine the model's results.
* Make the necessary adjustments to the model.
* Set up the model.

### **15. What are Regular Expressions?**

To match and tag words, a regular expression is employed. It is made up of a set of characters that are used to match strings.

If A and B are regular expressions, then they must satisfy the following conditions:

* It is a regular language, then for it is a regular expression.
* A + B is a regular expression within the language A, B if A and B are regular expressions.
* The concatenation of A and B (A.B) is a regular expression if A and B are regular expressions.
* A\* (A occurring multiple times) is a regular expression if A is a regular expression.

### **16. What is the difference between Natural Language Processing (NLP) and Natural Language Understanding (NLU)?**

****Natural Language Processing (NLP)****

* NLP is a system that handles simultaneous end-to-end talks between computers and people.
* In NLP, both humans and robots are engaged.
* NLP is concerned with understanding language in its purest form, such as stated.
* Grammar, structure, typography, and point of view may all be used to parse text using NLP.

****Natural Language Understanding(NLU)****

* NLU assists in resolving Artificial Intelligence's most complex challenges.
* NLU transforms unstructured inputs into structured text, allowing machines to comprehend them
* NLU, on the other hand, focuses on obtaining context and meaning or determining what was intended.
* NLU will assist the machine in deducing the meaning of the linguistic material.

### **17. What is a Masked Language Model, and how does it work?**

By generating an output from the defective input, masked language models assist learners in comprehending deep representations in downstream tasks. This approach is frequently used to anticipate the words in a phrase.

### **18. What is POS tagging?**

POS tagging, or parts of speech tagging, is the basis for identifying individual words in a document and classifying them as part of speech based on their context. Because it entails analysing grammatical structures and selecting the appropriate component, POS tagging is also known as grammatical tagging.

Because the same word might be several parts of speech depending on the context, POS tagging is a complicated procedure. Because of the same reason, the same general approach used for word mapping is unsuccessful for POS tagging.

### **19. What exactly is NES?**

The practice of recognising certain entities in a text document that are more informative and have a distinct context is known as named entity recognition (NER). These are frequently referred to as places, individuals, organisations, and others. Even though these things appear to be proper nouns, the NER approach does not recognise them. In reality, NER entails entity chunking or extraction, which includes segmenting entities into many specified classes. This stage also aids in the extraction of data.

### **20. What exactly is NLTK? What distinguishes it from Spacy?**

Natural Language Toolkit (NLTK) is a set of libraries and applications for processing symbolic and statistical natural language. This toolkit includes some of the most sophisticated libraries for breaking down and understanding human language using machine learning approaches. Lemmatization, Punctuation, Character Count, Tokenization, and Stemming are all done with NLTK. The following are the differences between NLTK and Spacey:

* While NLTK provides various programmes to pick from, Spacey's toolkit only contains the best-suited algorithm for a given scenario.
* In comparison to Spacey, NLTK supports many languages (Spacey supports only seven languages)
* NLTK provides a string processing library, but Spacey has an object-oriented library. Spacey can handle word vectors, whereas NLTK cannot.

### **21. What is the definition of information extraction?**

In the context of Natural Language Processing, information extraction refers to the process of mechanically extracting structured information from unstructured sources to assign meaning to it. This might involve retrieving entity properties, relationships between entities, and more. The following are some examples of information extraction models:

* Module for Taggers
* Module for Extracting Relationships
* Module for Fact Extraction
* Module for Extracting Entities
* Module for Sentiment Analysis
* Module for Network Graphs
* Module for Document Classification and Language Modeling

### **22. What are the most effective NLP tools?**

Some of the most excellent open-source NLP tools are

* SpaCy
* TextBlob
* Textacy
* Natural language Toolkit (NLTK)
* Retext
* NLP.js
* Stanford NLP
* CogcompNLP

### **23. List 10 use cases to be solved using NLP techniques?**

* Sentiment Analysis
* Language Translation (English to German, Chinese to English, etc..)
* Document Summarization
* Question Answering
* Sentence Completion
* Attribute extraction (Key information extraction from the documents)
* Chatbot interactions
* Topic classification
* Intent extraction
* Grammar or Sentence correction
* Image captioning
* Document Ranking
* Natural Language inference

### **24. In NLP, what are stop words?**

Stop words are common words that appear in sentences and provide weight to the phrase. These stop words serve as a link between phrases, ensuring grammatically accurate. Stop words are taken out before natural language data is processed, and they are a frequent pre-processing strategy.

### **25. In NLP, what is stemming?**

Stemming is the process of extracting the root word from a given term. With efficient and well-generalized principles, all tokens may be broken down to retrieve the root word or stem. It's a rule-based system that's well-known for its ease of use.

### **26. Give two instances of real-world NLP uses?**

****1. Spelling/Grammatical Checking Apps:****NLP algorithms are used in mobile applications and websites that help users fix grammar problems in the submitted text. These days, they may even suggest the next few words that the user might input, thanks to the employment of particular NLP models on the backend.

****2. ChatBots:****Many websites now provide customer service via virtual bots that talk with users and help them solve problems. It functions as a filter for concerns that do not require engagement with the customer service representatives of the firms.

### **27. Define Dependency Parsing?**

Dependency parsing is a technique for understanding the grammatical structure by highlighting the relationships between its components. It investigates how the words of a phrase are related linguistically. Dependencies are the names given to these connections.

### **28. What is the difference between false positives and false negatives?**

A false negative occurs when a machine learning system incorrectly forecasts a negative outcome as positive.

A false positive is defined as a machine learning system that incorrectly forecasts a positive outcome as a negative.

### **29. List a few ways for tagging parts of speech?**

Rule-based tagging, HMM tagging, transformation-based tagging, and memory-based tagging are all examples of tagging techniques.

### **30. List a few examples of how the n-gram model is used in the real world?**

1. Communication Enhancement

2. Tagging of parts of speech

3. Generation of natural language

4. The similarity of Words

5. Identification of Authorship

6. Sentiment Analysis

7. Text Input Predictive

### **31. In NLP, what is the bigram model?**

A bigram model is an NLP model that uses the conditional probability of the preceding word to predict the likelihood of a word in a phrase. It is critical to know all of the last words to calculate the conditional probability of the previous word.

### **32. What are your impressions of the Masked Language Model?**

The Masked Language Model is a model that takes a phrase as input and attempts to finish it by accurately predicting a few concealed (masked) words.

### **33. List the types of sorts available in linguistic ambiguity?**

****1. Lexical Ambiguity:****This sort of ambiguity occurs when a phrase has homonyms and polysemy.

****2. Syntactic Ambiguity:**** Syntactic ambiguity occurs when the grammar of a statement allows for many interpretations.

****3. Semantic Ambiguity:**** When a statement comprises ambiguous words or phrases with unclear meanings, this ambiguity occurs.

### **34. Explain the N-gram model in NLP in a few words?**

The N-gram model is an NLP model that predicts the likelihood of a word in a phrase based on the conditional probability of n-1 preceding terms. The essential idea behind this method is that rather than utilising all of the previous words to predict the future word, we simply utilise a handful of them.

### **35. What is the bigram model's Markov assumption?**

For the bigram model, the Markov assumption assumes that the probability of a word in a phrase depends solely on the preceding word in that sentence rather than all last words.

### **36. What exactly do you mean by word embedding?**

Word embedding is the method of expressing textual data using a real-number vector in natural language processing. This technique allows words with similar meanings to be represented simultaneously.

### **37. What is an embedding matrix, and how does it work?**

A word embedding matrix is a matrix that contains all of the words in a text's embedding vectors.

### **38. List a few popular word embedding techniques?**

A few word embedding approaches are listed below.

* Word2Vec Glove
* Embedding Layer

### **39. What are the first few steps you'll take before applying a natural language processing (NLP) machine-learning algorithm on a corpus?**

1. Eliminating white spaces

2. Eliminating Punctuation

3. Lowercase to Uppercase Conversion

4. Tokenisation

5. Getting Rid of Stopwords

6.Lemmatization

### **40. What is the difference between an hapax and an hapax legomenon?**

Hapaxes are unusual words that only appear once in a sample text or corpus. Each one is referred to as an hapax or hapax legomenon ('read-only once' in Greek). It's also known as a singleton.

****1. What do you know about NLP?****

 NLP stands for Natural Language Processing. It deals with making a machine understand the way human beings read and write in a language. This task is achieved by designing algorithms that can extract meaning from large datasets in audio or text format by applying machine learning algorithms.

****2. Give examples of any two real-world applications of NLP.****

****1. Spelling/Grammar Checking Apps:**** The mobile applications and websites that offer users correct grammar mistakes in the entered text rely on NLP algorithms. These days, they can also recommend the following few words that the user might type, which is also because of specific NLP models being used in the backend.

****2.**** ****ChatBots:**** Many websites now offer customer support through these virtual bots that chat with the user and resolve their problems. It acts as a filter to the issues that do not require an interaction with the companies’ customer executives.

****3. What is tokenization in NLP?****

Tokenization is the process of splitting running text into words and sentences.

**New Projects**

**[PyTorch Project to Build a GAN Model on MNIST Dataset](https://www.projectpro.io/project-use-case/gan-mnist-pytorch?utm_source=439&utm_medium=fold2)**[View Project](https://www.projectpro.io/project-use-case/gan-mnist-pytorch?utm_source=439&utm_medium=fold2)

**[CycleGAN Implementation for Image-To-Image Translation](https://www.projectpro.io/project-use-case/cyclegan-implementation-pytorch-for-image-to-image-translation?utm_source=439&utm_medium=fold2)**[View Project](https://www.projectpro.io/project-use-case/cyclegan-implementation-pytorch-for-image-to-image-translation?utm_source=439&utm_medium=fold2)

**[Build a Speech-Text Transcriptor with Nvidia Quartznet Model](https://www.projectpro.io/project-use-case/speech-to-text-transcriptor-wth-nvidia-quartznet-nemo?utm_source=439&utm_medium=fold2)**[View Project](https://www.projectpro.io/project-use-case/speech-to-text-transcriptor-wth-nvidia-quartznet-nemo?utm_source=439&utm_medium=fold2)

**[CycleGAN Implementation for Image-To-Image Translation](https://www.projectpro.io/project-use-case/cyclegan-implementation-pytorch-for-image-to-image-translation?utm_source=439&utm_medium=fold2)**[View Project](https://www.projectpro.io/project-use-case/cyclegan-implementation-pytorch-for-image-to-image-translation?utm_source=439&utm_medium=fold2)

**[CycleGAN Implementation for Image-To-Image Translation](https://www.projectpro.io/project-use-case/cyclegan-implementation-pytorch-for-image-to-image-translation?utm_source=439&utm_medium=fold2)**[View Project](https://www.projectpro.io/project-use-case/cyclegan-implementation-pytorch-for-image-to-image-translation?utm_source=439&utm_medium=fold2)

**[Talend Real-Time Project for ETL Process Automation](https://www.projectpro.io/project-use-case/talend-real-time-project-example-for-etl-process-automation?utm_source=439&utm_medium=fold2)**[View Project](https://www.projectpro.io/project-use-case/talend-real-time-project-example-for-etl-process-automation?utm_source=439&utm_medium=fold2)

**[Build a Credit Default Risk Prediction Model with LightGBM](https://www.projectpro.io/project-use-case/credit-risk-prediction-modelling-in-python-with-lightgbm?utm_source=439&utm_medium=fold2)**[View Project](https://www.projectpro.io/project-use-case/credit-risk-prediction-modelling-in-python-with-lightgbm?utm_source=439&utm_medium=fold2)

**[Build an End-to-End AWS SageMaker Classification Model](https://www.projectpro.io/project-use-case/aws-sagemaker-classification-model-example?utm_source=439&utm_medium=fold2)**[View Project](https://www.projectpro.io/project-use-case/aws-sagemaker-classification-model-example?utm_source=439&utm_medium=fold2)

**[Build an AI Chatbot from Scratch using Keras Sequential Model](https://www.projectpro.io/project-use-case/build-a-chatbot-from-scratch-using-keras-sequential-model?utm_source=439&utm_medium=fold2)**[View Project](https://www.projectpro.io/project-use-case/build-a-chatbot-from-scratch-using-keras-sequential-model?utm_source=439&utm_medium=fold2)

**[PyTorch Project to Build a GAN Model on MNIST Dataset](https://www.projectpro.io/project-use-case/gan-mnist-pytorch?utm_source=439&utm_medium=fold2)**[View Project](https://www.projectpro.io/project-use-case/gan-mnist-pytorch?utm_source=439&utm_medium=fold2)

**[PyTorch Project to Build a GAN Model on MNIST Dataset](https://www.projectpro.io/project-use-case/gan-mnist-pytorch?utm_source=439&utm_medium=fold2)**[View Project](https://www.projectpro.io/project-use-case/gan-mnist-pytorch?utm_source=439&utm_medium=fold2)

**[CycleGAN Implementation for Image-To-Image Translation](https://www.projectpro.io/project-use-case/cyclegan-implementation-pytorch-for-image-to-image-translation?utm_source=439&utm_medium=fold2)**[View Project](https://www.projectpro.io/project-use-case/cyclegan-implementation-pytorch-for-image-to-image-translation?utm_source=439&utm_medium=fold2)

**[Build a Speech-Text Transcriptor with Nvidia Quartznet Model](https://www.projectpro.io/project-use-case/speech-to-text-transcriptor-wth-nvidia-quartznet-nemo?utm_source=439&utm_medium=fold2)**[View Project](https://www.projectpro.io/project-use-case/speech-to-text-transcriptor-wth-nvidia-quartznet-nemo?utm_source=439&utm_medium=fold2)

**[CycleGAN Implementation for Image-To-Image Translation](https://www.projectpro.io/project-use-case/cyclegan-implementation-pytorch-for-image-to-image-translation?utm_source=439&utm_medium=fold2)**[View Project](https://www.projectpro.io/project-use-case/cyclegan-implementation-pytorch-for-image-to-image-translation?utm_source=439&utm_medium=fold2)

**[CycleGAN Implementation for Image-To-Image Translation](https://www.projectpro.io/project-use-case/cyclegan-implementation-pytorch-for-image-to-image-translation?utm_source=439&utm_medium=fold2)**[View Project](https://www.projectpro.io/project-use-case/cyclegan-implementation-pytorch-for-image-to-image-translation?utm_source=439&utm_medium=fold2)

**[Talend Real-Time Project for ETL Process Automation](https://www.projectpro.io/project-use-case/talend-real-time-project-example-for-etl-process-automation?utm_source=439&utm_medium=fold2)**[View Project](https://www.projectpro.io/project-use-case/talend-real-time-project-example-for-etl-process-automation?utm_source=439&utm_medium=fold2)

**[Build a Credit Default Risk Prediction Model with LightGBM](https://www.projectpro.io/project-use-case/credit-risk-prediction-modelling-in-python-with-lightgbm?utm_source=439&utm_medium=fold2)**[View Project](https://www.projectpro.io/project-use-case/credit-risk-prediction-modelling-in-python-with-lightgbm?utm_source=439&utm_medium=fold2)

**[Build an End-to-End AWS SageMaker Classification Model](https://www.projectpro.io/project-use-case/aws-sagemaker-classification-model-example?utm_source=439&utm_medium=fold2)**[View Project](https://www.projectpro.io/project-use-case/aws-sagemaker-classification-model-example?utm_source=439&utm_medium=fold2)

**[Build an AI Chatbot from Scratch using Keras Sequential Model](https://www.projectpro.io/project-use-case/build-a-chatbot-from-scratch-using-keras-sequential-model?utm_source=439&utm_medium=fold2)**[View Project](https://www.projectpro.io/project-use-case/build-a-chatbot-from-scratch-using-keras-sequential-model?utm_source=439&utm_medium=fold2)

**[PyTorch Project to Build a GAN Model on MNIST Dataset](https://www.projectpro.io/project-use-case/gan-mnist-pytorch?utm_source=439&utm_medium=fold2)**[View Project](https://www.projectpro.io/project-use-case/gan-mnist-pytorch?utm_source=439&utm_medium=fold2)

[View all New Projects](https://www.projectpro.io/projects?utm_source=439&utm_medium=fold2)

****4. What is the difference between a formal language and a natural language?****

|  |  |
| --- | --- |
| Formal Language | Natural Language |
| A formal language is a collection of strings, where each string contains symbols from a finite set called alphabets. | A natural language is a language that humans utilize to speak. It is usually a lot different from a formal language. These typically contain fragments of words and pause words like uh, um, etc. |

****5. What is the difference between stemming and lemmatization?****

Both stemming and lemmatization are keyword normalization techniques aiming to minimize the morphological variation in the words they encounter in a sentence. But, they are different from each other in the following way.

|  |  |
| --- | --- |
| Stemming | Lemmatization |
| This technique involves removing the affixes added to a word and leaving us with the rest of the word. | Lemmatization is the process of converting a word into its lemma from its inflected form. |
| Example: ‘Caring’→ ’Car’ | Example: ‘Caring’→ ’Care’ |

****6. What is NLU?****

 NLU stands for Natural Language Understanding. It is a subdomain of NLP that concerns making a machine learn the skills of reading comprehension. A few applications of NLU include Machine translation (MT), Newsgathering, and Text categorization. It often goes by the name Natural Language Interpretation (NLI) as well.

****Get Closer To Your Dream of Becoming a Data Scientist with 70+ Solved**[End-to-End ML Projects](https://www.projectpro.io/projects?utm_source=Blg439&utm_medium=ConvertLink&utm_campaign=TXTCTA1" \o "Machine Learning Projects" \t "https://www.projectpro.io/article/nlp-interview-questions-and-answers/_blank)**

****7. List the differences between NLP and NLU.****

|  |  |
| --- | --- |
| Natural Language Processing | Natural Language Understanding |
| NLP is a branch of AI that deals with designing programs for machines that will allow them to process the language that humans use. The idea is to make machines imitate the way humans utilize language for communication. | In NLU, the aim is to improve a computer’s ability to understand and analyze human language. This aim is achieved by transforming unstructured data into a machine-readable format. |

****8. What do you know about Latent Semantic Indexing (LSI)?****

LSI is a technique that analyzes a set of documents to find the statistical coexistence of words that appear together. It gives an insight into the topics of those documents.

LSI is also known as Latent Semantic Analysis.

****9. List a few methods for**[extracting features](https://www.projectpro.io/recipes/load-features-from-dictionary-in-python" \o "extracting features" \t "https://www.projectpro.io/article/nlp-interview-questions-and-answers/_blank)**from a corpus for NLP.****

1. Bag-of-Words

2. Word Embedding

****10. What are stop words?****

Stop words are the words in a document that are considered redundant by NLP engineers and are thus removed from the document before processing it. Few examples are ‘is’, ‘the’, ‘are, ‘am’.

****11. What do you know about Dependency Parsing?****

Dependency parsing is a technique that highlights the dependencies among the words of a sentence to understand its grammatical structure. It examines how the words of a sentence are linguistically linked to each other. These links are called dependencies.

****12. What is Text Summarization? Name its two types.****

Text Summarization is a method of converting a long-form text into a summary. The summary thus generated is expected to have critical ideas of the lengthy text. Two main types of Text Summarization are:

1. Extraction-based Summarization
2. Abstraction-based Summarization

****Get FREE Access to**[Machine Learning Example Codes](https://www.projectpro.io/recipes?utm_source=Blg439&utm_medium=RcpLink&utm_campaign=TXTCTA2" \o "Data Science and Machine Learning Python Example Codes" \t "https://www.projectpro.io/article/nlp-interview-questions-and-answers/_blank)**for Data Cleaning, Data Munging, and Data Visualization****

****13. What are false positives and false negatives?****

If a machine learning algorithm falsely predicts a negative outcome as positive, then the result is labeled as a false negative.

And, if a machine learning algorithm falsely predicts a positive outcome as negative, then the result is labeled as a false positive.

****14. List a few methods for part-of-speech tagging.****

Rule-based tagging, HMM-tagging, transformation-based tagging, and memory-based tagging.

****15. What is a corpus?****

‘Corpus’ is a Latin word that means ‘body.’ Thus, a body of the written or spoken text is called a corpus.

****Recommended Reading:**[10 NLP Techniques Every Data Scientist Should Know](https://www.projectpro.io/article/10-nlp-techniques-every-data-scientist-should-know/415" \o "10 NLP Techniques Every Data Scientist Should Know" \t "https://www.projectpro.io/article/nlp-interview-questions-and-answers/_blank)**

### ****NLP Algorithm Interview Questions with Answers****

Most recruiters usually try to understand how well you know the models that are used widely in NLP. Take a look at these interview questions in NLP with answers that will help you upgrade your NLP algorithm skills.

****1. List a few real-world applications of the n-gram model.****

1. Augmentive Communication

2. Part-of-speech Tagging

3. Natural language generation

4. Word Similarity

5. Authorship Identification

6. Sentiment Extraction

7. Predictive Text Input

****2. What does TF\*IDF stand for? Explain its significance.****

TF\*IDF stands for Term-Frequency/Inverse-Document Frequency. It is an information-retrieval measure that encapsulates the semantic significance of a word in a particular document N, by degrading words that tend to appear in a variety of different documents in some huge background corpus with D documents.

Let nw denote the frequency of a word w in the document N, m represents the total number of documents in the corpus that contain w. Then, TF\*IDF is defined as

TF\*IDF(w)=nw×lognm

****3. What is perplexity in NLP?****

It is a metric that is used to test the performance of language models. Mathematically, it is defined as a function of the probability that the language model represents a test sample. For a test sample X = x1, x2, x3,....,xn , the perplexity is given by,

PP(X)=P(x1,x2,…,xN)-1N

where N is the total number of word tokens.

Higher the perplexity, lesser is the information conveyed by the language model.

****4. Which algorithm in NLP supports bidirectional context?****

 BERT

****5. What is the Naive Bayes algorithm?****

Naive Bayes is a [classification machine learning algorithm](https://www.projectpro.io/article/7-types-of-classification-algorithms-in-machine-learning/435" \o "classification machine learning algorithm" \t "https://www.projectpro.io/article/nlp-interview-questions-and-answers/_blank) that utilizes Baye’s Theorem for labeling a class to the input set of features. A vital element of this algorithm is that it assumes that all the feature values are independent.

**Explore Categories**

**[Data Science Projects in Python](https://www.projectpro.io/projects/data-science-projects/data-science-projects-in-python?utm_source=439&utm_medium=fold3)** **[Deep Learning Projects](https://www.projectpro.io/projects/data-science-projects/deep-learning-projects?utm_source=439&utm_medium=fold3)** **[Neural Network Projects](https://www.projectpro.io/projects/data-science-projects/neural-network-projects?utm_source=439&utm_medium=fold3)** **[Tensorflow Projects](https://www.projectpro.io/projects/data-science-projects/tensorflow-projects?utm_source=439&utm_medium=fold3)** **[H2O R Projects](https://www.projectpro.io/projects/data-science-projects/h2o-r-projects?utm_source=439&utm_medium=fold3)** **[IoT Projects](https://www.projectpro.io/projects/data-science-projects/iot-projects?utm_source=439&utm_medium=fold3)** **[Keras Deep Learning Projects](https://www.projectpro.io/projects/data-science-projects/keras-deep-learning-projects?utm_source=439&utm_medium=fold3)** **[NLP Projects](https://www.projectpro.io/projects/data-science-projects/nlp-projects?utm_source=439&utm_medium=fold3)** **[Pytorch](https://www.projectpro.io/projects/data-science-projects/pytorch?utm_source=439&utm_medium=fold3)** **[Data Science Projects in Banking and Finance](https://www.projectpro.io/projects/data-science-projects/data-science-in-finance?utm_source=439&utm_medium=fold3)** **[Data Science Projects in Retail & Ecommerce](https://www.projectpro.io/projects/data-science-projects/machine-learning-in-retail-and-ecommerce?utm_source=439&utm_medium=fold3)** **[Data Science Projects in Entertainment & Media](https://www.projectpro.io/projects/data-science-projects/machine-learning-in-media-and-entertainment?utm_source=439&utm_medium=fold3)** **[Data Science Projects in Telecommunications](https://www.projectpro.io/projects/data-science-projects/machine-learning-in-telecommunications-industry?utm_source=439&utm_medium=fold3)**

****6. What is Part-of-Speech tagging?****

 Part-of-speech tagging is the task of assigning a part-of-speech label to each word in a sentence. A variety of part-of-speech algorithms are available that contain tagsets having several tags between 40 and 200.

****7. What is the bigram model in NLP?****

 A bigram model is a model used in NLP for predicting the probability of a word in a sentence using the conditional probability of the previous word. For calculating the conditional probability of the previous word, it is crucial that all the previous words are known.

****8. What is the significance of the Naive Bayes algorithm in NLP?****

The Naive Bayes algorithm is widely used in NLP for various applications. For example: to determine the sense of a word, to predict the tag of a given text, etc.

****9. What do you know about the Masked Language Model?****

The Masked Language Model is a model that takes a sentence with a few hidden (masked) words as input and tries to complete the sentence by correctly guessing those hidden words.

****10. What is the Bag-of-words model in NLP?****

Bag-of-words refers to an unorganized set of words. The Bag-of-words model is NLP is a model that assigns a vector to a sentence in a corpus. It first creates a dictionary of words and then produces a vector by assigning a binary variable to each word of the sentence depending on whether it exists in the bag of words or not.

****11. Briefly describe the N-gram model in NLP.****

 N-gram model is a model in NLP that predicts the probability of a word in a given sentence using the conditional probability of n-1 previous words in the sentence. The basic intuition behind this algorithm is that instead of using all the previous words to predict the next word, we use only a few previous words.

****12. What is the Markov assumption for the bigram model?****

 The Markov assumption assumes for the bigram model that the probability of a word in a sentence depends only on the previous word in that sentence and not on all the previous words.

****13. What do you understand by word embedding?****

 In NLP, word embedding is the process of representing textual data through a real-numbered vector. This method allows words having similar meanings to have a similar representation.

****14. What is an embedding matrix?****

 A word embedding matrix is a matrix that contains embedding vectors of all the words in a given text.

****15. List a few popular methods used for word embedding.****

Following are a few methods of word embedding.

1. Embedding Layer
2. Word2Vec
3. Glove

### ****NLP Interview Coding Questions with Answers****

In the NLP interview questions round, the interviewer will be interested in your coding skills as well. Thus, you mustn’t miss the NLP interview questions below before going for your interview.

****1. How will you use Python’s concordance command in**[NLTK](https://www.projectpro.io/recipes/download-nltk-library" \o "NLTK" \t "https://www.projectpro.io/article/nlp-interview-questions-and-answers/_blank)**for a text that does not belong to the package?****

The concordance() function can easily be accessed for a text that belongs to the NLTK package using the following code:

>>>from nltk.book import \*

>>>text1.concordance("monstrous")

However, for a text that does not belong to the NLTK package, one has to use the following code to access that function.

>>>import nltk.corpus

>>>from nltk.text import Text

>>>NLTKtext = Text(nltk.corpus.gutenberg.words('Your\_file\_name\_here.txt'))

>>>NLTKtext.concordance('word')

Here, we have created a Text object to access the concordance() function. The function displays the occurrence of the chosen word and the context around it.

****2. Write the code to count the number of distinct tokens in a text?****

len(set(text))

****3. What are the first few steps that you will take before applying an NLP machine-learning algorithm to a given corpus?****

Ans: 1. Removing white spaces

2. Removing Punctuations

3. Converting Uppercase to Lowercase

4. [Tokenization](https://www.projectpro.io/recipes/explain-difference-between-word-tokenizer" \o "Tokenization" \t "https://www.projectpro.io/article/nlp-interview-questions-and-answers/_blank)

5. [Removing Stopwords](https://www.projectpro.io/recipes/add-custom-stopwords-and-then-remove-them-from-text" \o "Removing Stopwords" \t "https://www.projectpro.io/article/nlp-interview-questions-and-answers/_blank)

6. Lemmatization

****4. For correcting spelling errors in a corpus, which one is a better choice: a giant dictionary or a smaller dictionary, and why?****

Initially, a smaller dictionary is a better choice because most NLP researchers feared that a giant dictionary would contain rare words that may be similar to misspelled words. However, later it was found (Damerau and Mays (1989)) that in practice, a more extensive dictionary is better at marking rare words as errors.

****5. Do you always recommend removing punctuation marks from the corpus you’re dealing with? Why/Why not?****

No, it is not always a good idea to remove punctuation marks from the corpus as they are necessary for certain NLP applications that require the marks to be counted along with words.

For example: Part-of-speech tagging, parsing, speech synthesis.

****6. List a few libraries that you use for NLP in Python.****

NLTK, Scikit-learn,GenSim, SpaCy, CoreNLP, TextBlob.

****7, Suggest a few machine learning/deep learning models that are used in NLP.****

Support Vector Machines, Neural Networks, Decision Tree, Bayesian Networks.

****8. Which library contains the Word2Vec model in Python?****

GenSim

### ****Advanced NLP Interview Questions with Answers****

It is not always the case in an NLP interview that you’ll be asked common questions. Sometimes, to test whether you are genuinely interested in the field of NLP, an interviewer may ask you slightly advanced questions. And, we don't want those advanced questions to refrain you from achieving your dream job. So, go through the following NLP interview questions and answers that will give you an edge over other applicants.

****1. What are homographs, homophones, and homonyms?****

|  |  |  |
| --- | --- | --- |
| Homographs | Homophones | Homonyms |
| “Home”=same “graph”=write | “Home”=same “phone”=sound | “Homo”=same, “onym” = name |
| These are the words that have the same spelling but may or may not have the same pronunciations. | These are the words that sound similar but have different spelling and different meanings. | These are the words that have the same spelling and pronunciation but different meanings. |
| To live a life, airing a show live | Eye, I | River Bank, Bank Account |

****2. Is converting all text in uppercase to lowercase always a good idea? Explain with the help of an example.****

No, for words like The, the, THE, it is a good idea as they all will have the same meaning. However, for a word like brown which can be used as a surname for someone by the name Robert Brown, it won’t be a good idea as the word ‘brown’ has different meanings for both the cases. We, therefore, would want to treat them differently. Hence, it is better to change uppercase letters at the beginning of a sentence to lowercase, convert headings and titles to which are all in capitals to lowercase, and leave the remaining text unchanged.

****3. What is a hapax/hapax legomenon?****

The rare words that only occur once in a sample text or corpus are called hapaxes. Each one of them is called an hapax or hapax legomenon (greek for ‘read-only once’). It is also called a singleton.

****4. Is tokenizing a sentence based on white-space ‘ ‘ character sufficient? If not, give an example where it may not work.****

Tokenizing a sentence using the white space character is not always sufficient.

Consider the example,

“ One of our users said, ‘I love Dezyre’s content’. ”

Tokenizing purely based on white space would result in the following words:

‘I        said,   content’.

****5. What is a collocation?****

A collocation is a group of two or more words that possess a relationship and provide a classic alternative of saying something. For example, ‘strong breeze’, ‘the rich and powerful’, ‘weapons of mass destruction.

****6. List a few types of linguistic ambiguities.****

****1. Lexical Ambiguity:**** This type of ambiguity is observed because of homonyms and polysemy in a sentence.

****2. Syntactic Ambiguity:**** A syntactic ambiguity is observed when based on the sentence’s syntax, more than one meaning is possible.

****3. Semantic Ambiguity:**** This ambiguity occurs when a sentence contains ambiguous words or phrases that have ambiguous meanings.

****7. What is a Turing Test? Explain with respect to NLP-based systems.****

Alan Turing developed a test, called Turing Test, that could differentiate between humans and machines. A computer machine is considered intelligent if it can pass this test through its use of language. Alan believed that if a machine could use language the way humans do, it was sufficient for the machine to prove its intelligence.

****8. What do you understand by regular expressions in NLP?****

Regular expressions in natural language processing are algebraic notations representing a set of strings. They are mainly used to find or replace strings in a text and can also be used to define a language in a formal way.

****9. Differentiate between orthographic rules and morphological rules with respect to singular and plural forms of English words.****

|  |  |
| --- | --- |
| Orthographiical Rules | Morphological Rules |
| These are the rules that contain information for extracting the plural form of English words that end in ‘y’. Such words are transformed into their plural form by converting ‘y’ into ‘i’ and adding the letters ‘es’ as suffixes. | These rules contain information for words like fish; there are null plural forms. And words like goose have their plural generated by a change of the vowel. |

****10. Define the term parsing concerning NLP.****

Parsing refers to the task of generating a linguistic structure for a given input. For example, parsing the word ‘helping’ will result in ****verb****-pass + ****gerund****-ing.

## ****The Most Common Use-cases of NLP are:****

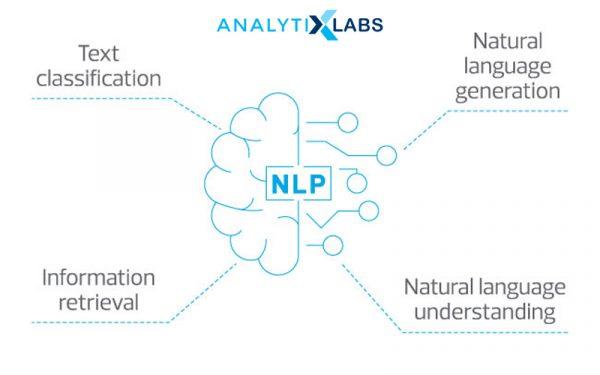
1. Sentiment Analysis
2. Text Summarization
3. Content Categorization
4. Speech-to-Text conversion and vice-versa
5. Machine Translation

### ****Let Us Look At The Most Common NLP Terms****

* **Vocabulary:** The group of terms used in a text or speech.
* **Corpus, or Corpora(Plural):** It is a collection of text of similar type, for example, movie review, social media posts, etc.
* **Documents:** They are the body of text and collectively form a corpus.
* **Out of Vocabulary:** Terms that are not included in the vocabulary that we created during our model’s training are included in this category.
* **Preprocessing:** It is a method that attempts to remove unwanted text, or noise, from the given text and to make it “clean.” It is the first step of any NLP task. You may also like to read about [Data Preprocessing in Machine Learning](https://www.analytixlabs.co.in/blog/data-preprocessing-in-machine-learning/" \t "https://www.analytixlabs.co.in/blog/nlp-interview-questions/_blank).
* **Tokenization:** **Tokenization in NLP**breaks down the large sets of text into small parts for easy readability and understanding. Each small part is referred to as ‘text’ and provides a piece of meaningful information.
* **Embeddings (Word):** It is the process of embedding each token as a vector before passing it into a machine learning model. Embeddings can also be done on phrases and characters as well, apart from words.
* **N-grams**: It is a continuous sequence (similar to the power set in number theory) of n-tokens of a given text.
* **Transformers:** They are deep learning architectures that can have the ability to parallelize computations. Transformers are used to learn long term dependencies.
* **Parts of Speech (POS):** They are the word’s functions, like a noun, verb, etc.
* **Parts of Speech Tagging:** It is the process of tagging words in the sentences into different parts of speech.
* **Stop Words:** It is the removal of unwanted text from further processing of text, for example, a, to, can, etc.
* **Normalization:** It is the process of mapping similar terms to a canonical form, i.e., a single entity.
* **Lemmatization:** Lemmatization in NLP is a type of normalization used to group similar terms to their base form based on the parts of speech. For example, talking and talking can be mapped to a single term, walk.
* **Stemming:** **Stemming in NLP** is also a type of normalization and is similar to lemmatization, but the difference here is that it segregates the words without the parts of speech tags. It is faster than lemmatization and can also be more accurate in some cases.

## ****NLP Interview Questions** With Answers**

In this section, let us see the extensive set of NLP interview questions. **Following the question links, you can also find answers to these NLP interview questions.**



### ****1. What is the NLG (Natural Language Generation)?****

Natural Language Generation is a part of AI and generates natural language texts from structured data to produce an output. It can be seen as NLP’s reverse process, where NLP is used to understand and interpret the natural language to form data, and NLU is used to generate outputs in natural language from structured data.

### ****2. What is the order of steps in natural language understanding?****

The order of steps that are to be followed in Natural Language Understanding is as follows:

1. Signal Processing
2. Syntactic Analysis
3. Semantic Analysis
4. Pragmatics

### ****3. What is signal processing in NLP?****

 Signal processing is a method that enables software to analyze, modify, and synthesize signals. In NLP, these can be sound or text signals.

### ****4. What is pragmatic analysis in NLP?****

The pragmatic analysis is the process of information extraction from the given text. It is a set of linguistic and logical tools that enable us to churn out the meaning of the given structure of a text.

### ****5. What is syntactic analysis in NLP?****

  The syntactic analysis, also referred to as parsing and syntax analysis, is the phase in which we try to process the given text’s structure. This process tries to draw meaning from the text by comparing it to formal grammar rules or syntax.

### ****6. What is semantic analysis in NLP?****

The semantic analysis is the process of understanding the meaning of the text in the way humans perceive and communicate. It focuses on larger parts of data for processing, as compared to other analysis techniques.

### ****7. What is sentiment analysis in NLP?****

The sentiment analysis, also known as opinion mining and emotion AI, is a process of detecting the polarity of the opinion in the text or can be a part of it. It is majorly used to identify, extract, and quantify user or customer reviews’ polarity, survey responses, or social media opinions.

### ****8. What is discourse analysis in NLP?****

Discourse is a structured group of the sentence. Discourse analysis can be termed as an approach to analyzing the discourse, i.e., text or language. It involves text interpretations and interactions.

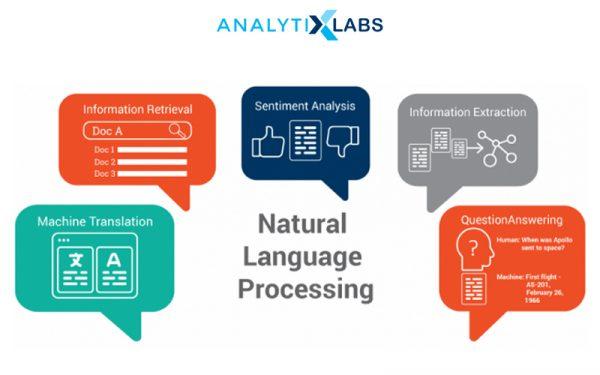
### ****9. What is pragmatic ambiguity in NLP?****

Pragmatic ambiguity can be referred to as a condition where words have multiple interpretations. This condition arises when the meaning of words is not specific; i.e., it can give different meanings.

### ****10. What are the major applications of NLP?****

 The major applications of NLP are:

1. Machine Translation
2. Speech Recognition
3. Sentiment Analysis
4. Text Classification



### ****11. List any real-world application of NLP?****

 The most used real-world application of NLP is speech recognition. Examples of speech recognition applications are Amazon Alexa, Google Assistant, Siri, HP Cortana.

### ****12. What are the common NLP techniques?****

 The common NLP techniques for text extraction are:

1. Named Entity Recognition
2. Sentiment Analysis
3. Text Summarization
4. Aspect Mining
5. Text Modelling

### ****What are the components of NLP?****

  The components of NLP are:

1. Lexical Analysis
2. Syntactic Analysis
3. Semantic Analysis
4. Discourse Integration
5. Pragmatic Analysis

### ****14. What are the tools used for training NLP models?****

The tools used to train NLP models are NLTK, spaCY, PyTorch-NLP, openNLP.

### ****15. Which NLP technique uses a lexical knowledge base to obtain the correct base form of the words?****

The **NLP techniques**that use lexical knowledge to obtain the correct base form are lemmatization and stemming.

### ****16. List the models to reduce the dimensionality of data in NLP.****

 The commonly used models are TF-IDF, Word2vec/Glove, LSI, Topic Modelling, Elmo Embeddings.

### ****17. List some open-source libraries for NLP.****

            The popular libraries are NLTK (Natural Language ToolKit), SciKit Learn, Textblob, CoreNLP, spaCY, Gensim.

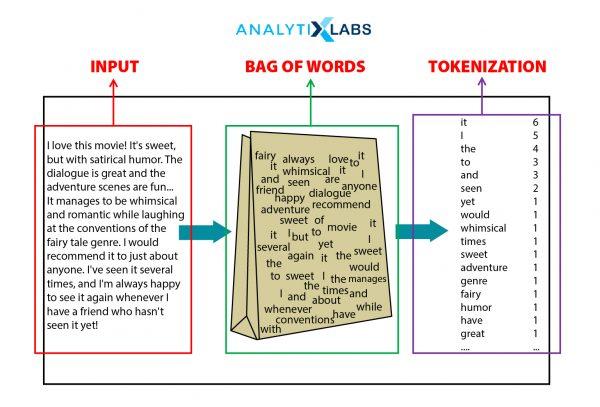


### ****18. Explain the masked language model.****

Masked modeling is an example of autoencoding language modeling. Here the output is predicted from corrupted input. By this model, we can predict the word from other words present in the sentences.

### ****19. What is the bag of words model?****

 The Bagofwords model is used for information retrieval. Here the text is represented as a multiset, i.e., a bag of words. We don’t consider grammar and word order, but we surely maintain the multiplicity.



### ****20. What is CBOW in NLP?****

CBOW or continuous bag of words is a model that tries to predict the target word from the available source context words, i.e., the surrounding words. Here the context words are taken into account as multiple words for a given target word.

### ****21. What is TF-IDF and what are its uses?****

 TF-IDF is an abbreviation for the term frequency-inverse documentary frequency. It is used to provide a numeric value to a word to show how important it is in the document or a corpus.

### ****22. What are POS and tagging?****

 Parts Of Speech (POS) are the functions of the word, like a noun, verb, etc., and tagging is labeling the words present in the sentences into different parts of speech.

### ****23. What is n-gram in NLP?****

   N-grams are the continuous sequence (similar to the power set in number theory) of n-tokens of a given text.

### ****24. What is skip-gram?****

 Skip gram is an unsupervised learning technique used to find the most related words to a target word. It is a reverse process of the continuous bag of words model.

### ****25. What is the corpus in NLP?****

Corpus or corpora (plural), is a collection of the text of a similar type, for example, movie reviews, social media posts, etc.

### ****26. What are the features of the text corpus in NLP?****

  The features of text corpus are:

1. Word count
2. Vector notation
3. Part of speech tag
4. Boolean feature
5. Dependency grammar

### ****27. What is normalization in NLP?****

Normalization is the process of mapping similar terms to a canonical form, i.e., a single entity.

### ****28. What is keyword normalization?****

  Keyword normalization is an **NLP technique**where we apply normalization on a word to condense it to its most basic form.

### ****29. What is lemmatization in NLP?****

 Lemmatization is a type of normalization used to group similar terms to their base form-based on the parts of speech. For example, talking and talking can be mapped to a single term, walk.

### ****30. What is stemming in NLP?****

Stemming in NLP is also a type of normalization and is similar to lemmatization, but the difference here is that it segregates the words without the parts of speech tags. It is faster than lemmatization and can also be more accurate in some cases.

### ****31. What is ambiguity in NLP?****

 Ambiguity can be referred to as a condition when a word can have multiple interpretations and results in being misunderstood. Natural languages are ambiguous and can make it difficult to process**NLP techniques** on them, resulting in the wrong output.

### ****32. What is tokenization in NLP?****

Tokenization is the process of breaking down large sets of text into small parts for easy readability and understanding. Each small part is referred to as ‘text’ and provides a piece of meaningful information.

### ****33. What are stop words in NLP?****

 Stop words are the unwanted text that is present in the input. It is the process of removal of unwanted text from further processing of text, for example, a, to, can, etc.

### ****34. How to find word similarity in NLP?****

 Word similarity in NLP is done by calculating the word vectors of the words in the vector space and then calculating the similarity on a scale of 0 to 1.

### ****35. How to find sentence similarity in NLP?****

 Sentence similarity is done in NLP by finding the cosine similarity between the two sentences. It can be done by finding the cosine angle between the vectors of two sentences in the inner product space.

### ****36. How to find document similarity in NLP?****

 Document similarity is done in NLP by converting the documents into the TF-IDF vectors form and finding their cosine similarity.

### ****37. What are transformers?****

Transformers are deep learning architectures that can parallelize computations. They are used to learn long-term dependencies.

### ****38. What are punctuations in NLP, and how can we remove them?****

Punctuations are the punctuations in the corpus or the input text. We can remove them by using the tokenizer function of NLTK. We can use nltk.RegexpTokenizer() to remove all punctuations.

### ****39. What is latent semantic indexing (LSI)?****

Latent Semantic Indexing,, also referred to as latent semantic analysis, is an **NLP technique**used to remove stop words from processing the text into the text’s main content. It is used to find relationships between different words.



### ****40. What is named entity recognition (NER)?****

Named Entity Recognition is a part of information retrieval, a method to locate and classify the entities present in the unstructured data provided and convert them into predefined categories.

### ****41. What is NLTK in NLP?****

 NTLK, an abbreviation of Natural Language Toolkit, is one of NLP’s most popular libraries. It was written in Python and contained libraries for tokenization, classification, tagging, stemming, parsing, and semantic reasoning.

### ****42. What is spaCY?****

spaCY is an open-source library for natural language processing on an advanced level. It is mostly used for production-level usage and uses convolutional neural network models.

### ****43. What is openNLP?****

openNLP is a java based library used for Natural Language Processing, and it supports most of the NLP tasks such as tokenization, language detection, etc.

### ****44. What is the difference between NLTK and openNLP?****

There is a small difference between NTLK and openNLP, i.e., NLTK is written in python, and openNLP is based on java. One other difference is that NTLK has an option of downloading corpora by an in-built method.

### ****45. What is parsing?****

Parsing is the method of analyzing the sentence automatically based on the syntactic structure.

### ****46. What is dependency parsing?****

Dependency parsing, also called syntactic parsing, recognizes a dependency parse of a sentence and assigns a syntax structure to a sentence. It focuses on the relationship between different words.

## ****47. What is semantic parsing?****

Semantic parsing is a method of conversion of natural language into machine-understandable form.

### ****48. What is constituency parsing?****

 Constituency parsing is a method of division of sentences into sub-parts or constituencies. It aims to extract a constituency-based parse tree from the constituencies of the sentences.

### ****49. What is shallow parsing?****

Shallow parsing, also known as light parsing and chunking, identifies constituents of sentences and then links them to different groups of grammatical meanings.

### ****50. What are the differences between dependency parsing and shallow parsing?****

The difference between shallow parsing and dependency parsing is that shallow parsing is the parsing of limited parts of the information. In contrast, dependency parsing is the process of finding relations between all the different words.

### ****51. What is language modeling?****

Language modeling is the process of creating a probability distribution of a sequence of words. It is used to provide probability to all the words present in the sequence.

### ****52. What is topic modeling?****

 Topic modeling is a method of finding abstract topics in a document or set of documents to find hidden semantic structures.

### ****53. What is text summarization in NLP?****

**Text summarization in NLP**is the process of conversion of large pieces of text to short text. It is intended to summarize the given text, keeping the main contents and overall meaning intact.

### ****54. What is the difference between a regular expression and regular grammar?****

The difference between regular and regular grammar is that regular grammar is used to generate regular language, and regular expression is used to represent regular language.

### ****55. What is perplexity in NLP?****

Perplexity is the condition when the system encounters something unaccountable or which is not meaningful.

### ****56. What is the Naive Bayes algorithm, and where is it used in NLP?****

Naive Bayes algorithm is used to predict tags of text by calculating the probability for each tag for the text and then providing the one with the highest probability.

### ****57. What is the PageRank algorithm?****

Google uses the PageRank algorithm. It is the algorithm to rank web pages in the search engine results.

### ****58. What is noise removal?****

 Noise removal is one of the **NLP techniques**i.e., used to remove pieces of text from the corpus that is not necessary as they can hinder our text analysis.

### ****59. What is word embedding?****

 Word embedding is the process of mapping words from the vocabulary to vectors of real numbers.



### ****60. What are the word embedding libraries?****

 The libraries that provide word embedding features are spaCY and genism.

### ****61. What is word2vec?****

Word2vec is a collection of models that are used to generate word embeddings. These models are trained to reconstruct the linguistic context of the words in the corpus.

### ****61. What is doc2vec?****

 Doc2vec is one of the unsupervised algorithms used to generate vectors of sentences or documents irrespective of their length.

### ****63. What is a document-term matrix?****

The document-term matrix, also called the term-document matrix, is the matrix that describes the frequency of terms occurring in a document.

### ****64. What is wordnet?****

 Wordnet can be described as a database created to store words from different languages connected by their semantic relationships.

### ****65. What is GloVe in NLP?****

   The gloVebased on their pronunciation.

### ****66. What is a flexible string matching?****

 Flexible string matching or fuzzy string matching is a method to find strings that are likely to match a specific pattern. It is also called approximate string matching as it uses an approximation to find patterns between strings.

### ****67. What is cosine similarity?****

 Cosine similarity is the measure of cosine difference between two non-zero vectors in the inner product space. It is used to find the similarity between documents irrespective of their size.

### ****68. What is information extraction?****

Information extraction is the process of extracting useful data in a structured way from a given unstructured set of data.

### ****69. What is object standardization, and when is it used?****

Object standardization is the process of extracting useful information from abbreviations and other acronyms that can not be meaningful in lexical dictionaries.

### ****71. What is text generation, and when is it done?****

   Text-generation is the process of generating natural language texts automatically in response to the communication. It uses artificial intelligence and computational linguistic knowledge to perform this task.

### ****71. How can we estimate the entropy of the English language?****

N-grams can estimate the entropy of the English language. The entropy of a letter is calculated by knowing the entropy of all the previous letters.

### ****72. What is Latent Dirichlet Allocation?****

 Latent Dirichlet Allocation is a topic modeling method where each topic represents a set of words, and every document is made of various words.

### ****73. What are the conditional random fields?****

Conditional Random Fields (CRFs ) are a collection of statistical modeling methods. It is used for pattern recognition and structure prediction.

### ****74. What are the hidden Markov random fields?****

Hidden Markov Random fields are a derivation of the Hidden Markov Model. It is a process generated by a Markov chain, whose state sequence can only be observed by a sequence of observations.

### ****75. What is a coreference resolution?****

Coreference resolution is the process of collecting all the expressions that are referring to the same entity in a text. It is used in information extraction, document summarization, and question answering.

### ****76. What is PAC learning?****

Probably Approximately Correct learning is a mathematical analysis framework. It is used for the analysis of generalization error of the learning algorithms.

### ****77. What is sequence learning?****

Sequence learning is a method of learning where both input and output are sequences.

### ****78. What is an ensemble method?****

 The ensemble method uses multiple learning algorithms to get enhanced and more accurate performance compared to the performance of an algorithm alone.

So these were the most frequently asked **NLP interview questions**, prepare them well, and increase your chances of getting selected.