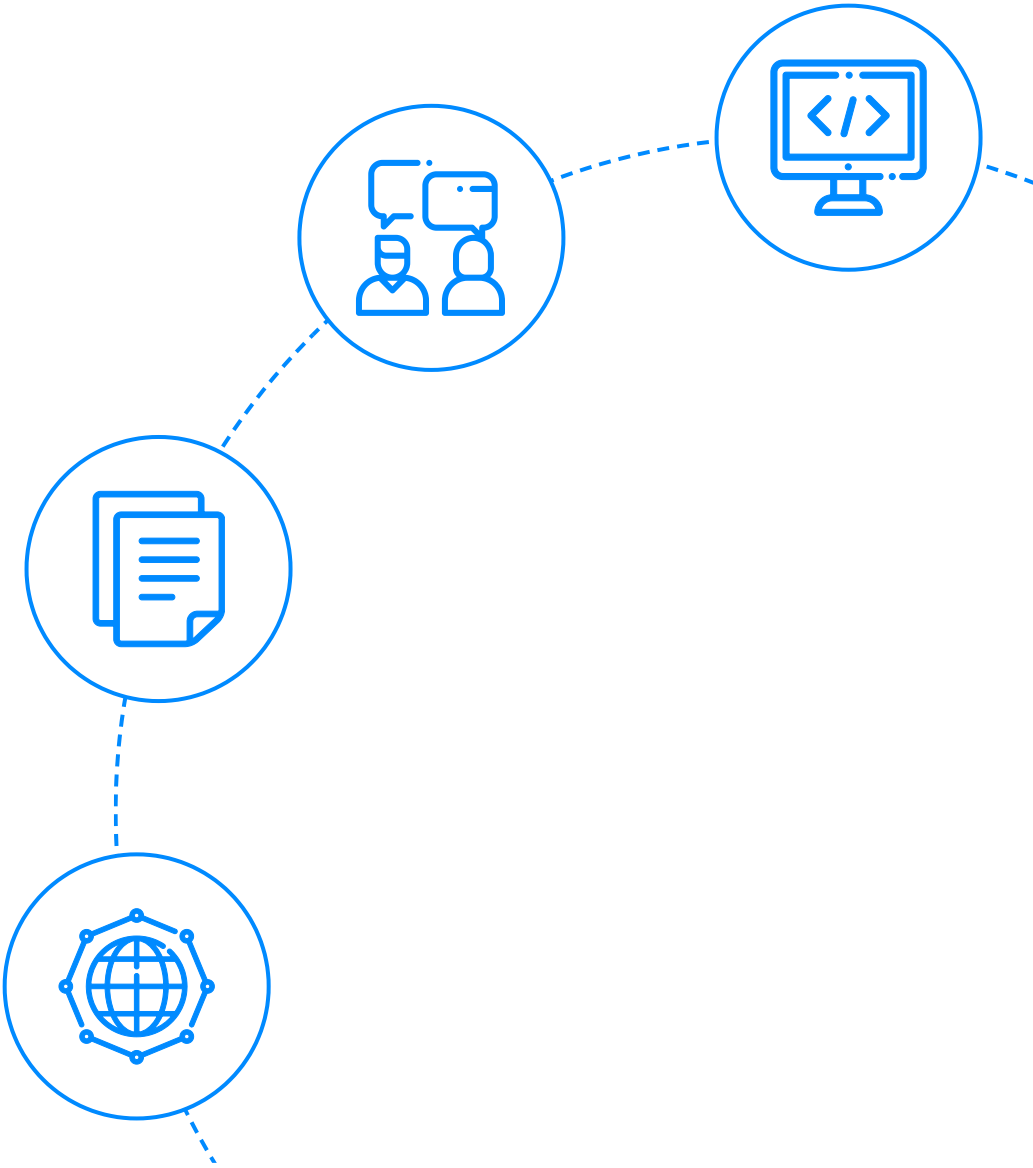
InterviewBit

NLP Interview Questions



To view the live version of the page, click here.

© Copyright by Interviewbit

Contents

NLP Interview Questions for Freshers

NLP Interview Questions for Experienced

© Copyright by Interviewbit

NLP Interview Questions for Experienced (.....Continued)

19. What is Latent Semantic Indexing (LSI) in NLP?
20. What is the difference between NLP and NLU?
21. What are some metrics on which NLP models are evaluated?
22. Explain the pipeline for Information extraction (IE) in NLP.
23. What do you mean by Autoencoders?
24. What do you mean by Masked language modelling?
25. What is the meaning of Pragmatic Analysis in NLP?
26. What is the meaning of N-gram in NLP?
27. What do you mean by perplexity in NLP?

© Copyright by Interviewbit

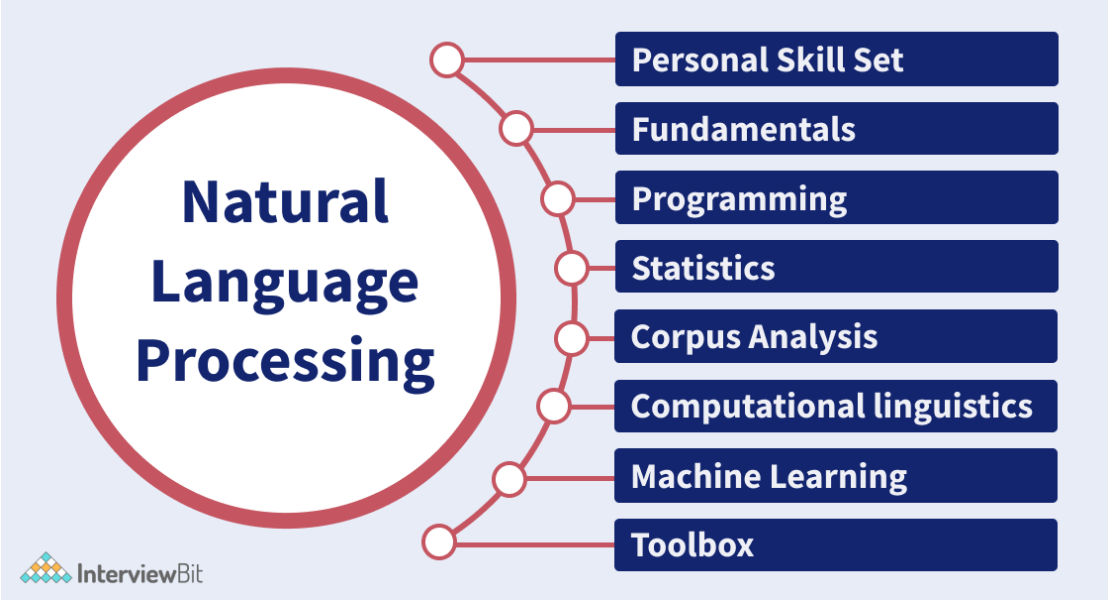
Let's get Started

Introduction to Natural Language Processing (NLP)

Natural language processing (NLP) is a branch of linguistics, computer science, and **artificial intelligence** 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**, and **deep learning** 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.

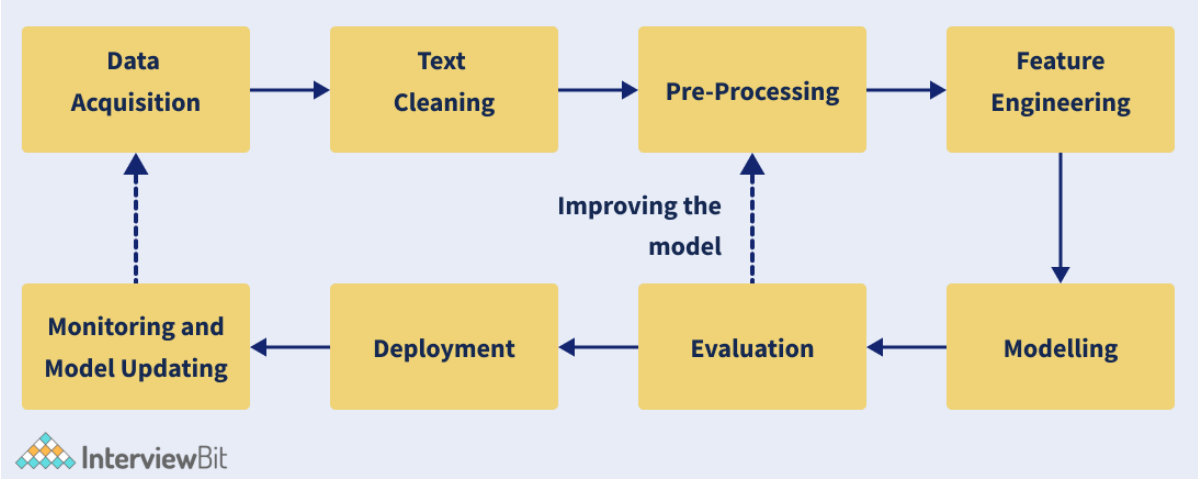
© Copyright by Interviewbit



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:



© Copyright by Interviewbit

- **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.

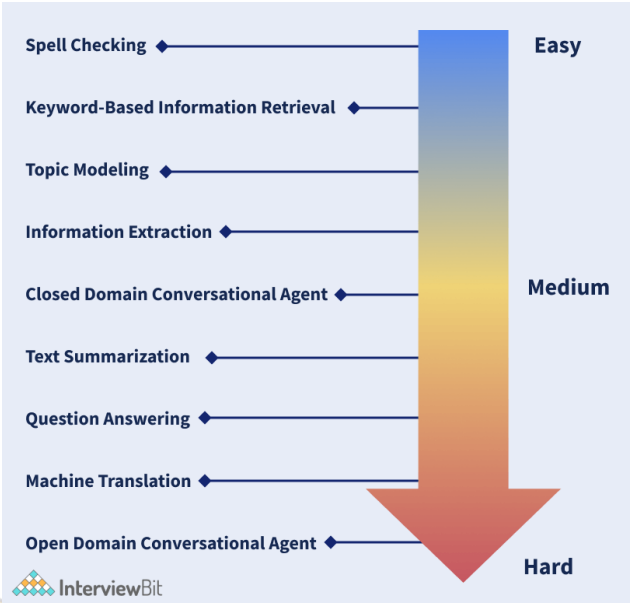
2. What are some of the common NLP tasks?

Some of the common tasks of NLP include:

© Copyright by Interviewbit

- **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.

© Copyright by Interviewbit



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:

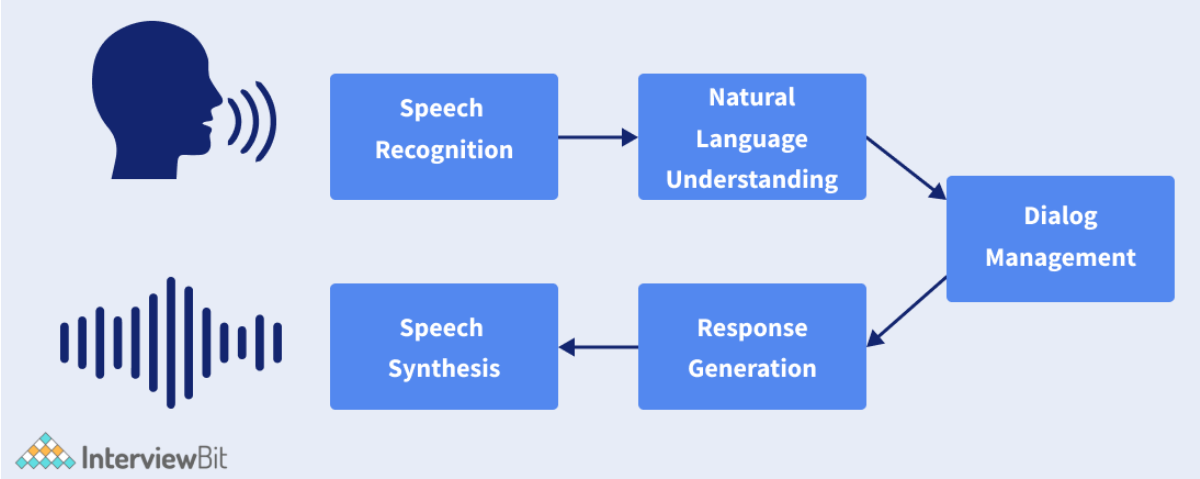
- Heuristics
- Machine learning
- Deep Learning

4. How do Conversational Agents work?

The following NLP components are used in Conversational Agents:

© Copyright by Interviewbit

- **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.

© Copyright by Interviewbit

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:

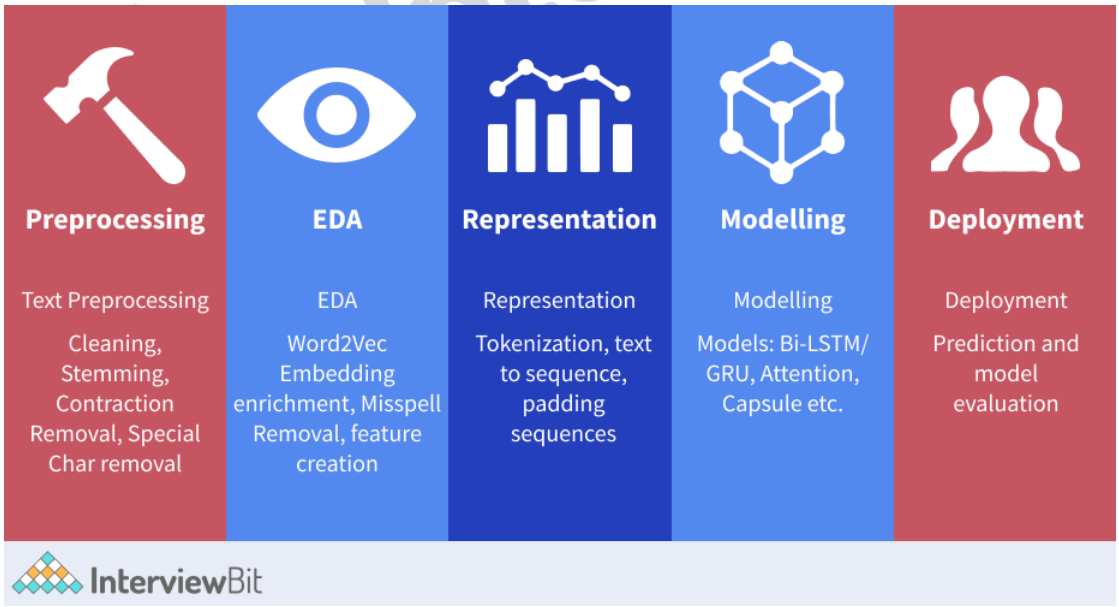
- Named Entity Recognition
- Sentiment Analysis
- Text Summarization
- Aspect Mining
- Topic Modeling

© Copyright by Interviewbit

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”).

© Copyright by Interviewbit

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

© Copyright by Interviewbit

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.

© Copyright by Interviewbit

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

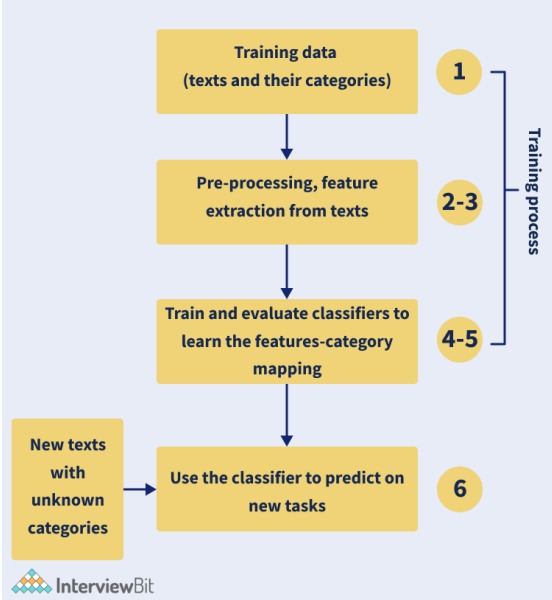
TF-IDF also called **Term Frequency-Inverse Document Frequency** helps us get the importance of a particular word relative to other words in the corpus. It's a common scoring metric in information retrieval (IR) and summarization. TF-IDF converts words into vectors and adds semantic information, resulting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

© Copyright by Interviewbit



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)?

© Copyright by Interviewbit

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?

© Copyright by Interviewbit

Natural Language Processing (NLP)	Natural Language Understanding
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 transformin them into structured text.
NLP focuses on interpreting language in its most literal sense, such as what was said.	NLU, on the other hand, concentrat extracting context and meaning, or 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:

© Copyright by Interviewbit

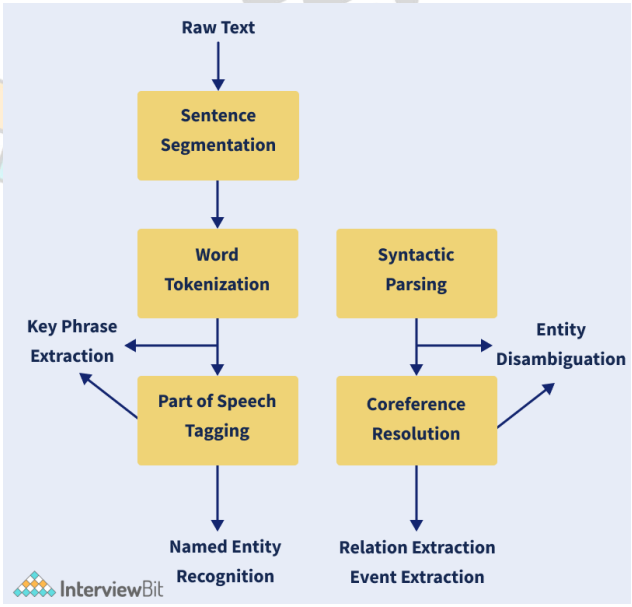
- **Accuracy:** When the output variable is categorical or discrete, accuracy is used. It is the percentage of correct predictions made by the model compared to the total number of predictions made.
- **Precision:** Indicates how precise or exact the model's predictions are, i.e., how many positive (the 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.

© Copyright by Interviewbit

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.

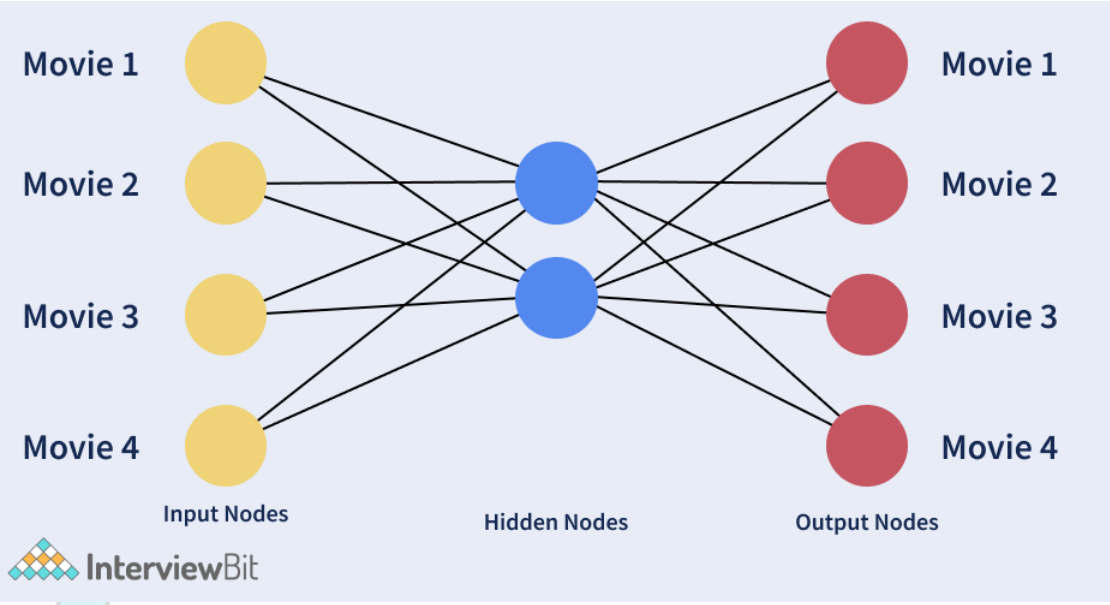


23. What do you mean by Autoencoders?

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.

© Copyright by Interviewbit

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.

© Copyright by Interviewbit

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.

© Copyright by Interviewbit

References and Resources:

- Natural Language Processing with Python – Book by Edward Loper, Ewan Klein, and Steven Bird (Published by: O'Reilly Media, Inc.)
- Practical Natural Language Processing – By Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, Harshit Surana (Published by: O'Reilly Media, Inc.)
- Natural Language Processing in Action: Understanding, Analyzing, and Generating Text with Python – Book by Cole Howard, Hannes Hapke, and Hobson Lane

© Copyright by Interviewbit

Links to More Interview Questions

C Interview Questions	Php Interview Questions	C Sharp Interview Questions
Web Api Interview Questions	Hibernate Interview Questions	Node Js Interview Questions
Cpp Interview Questions	Oops Interview Questions	Devops Interview Questions
Machine Learning Interview Questions	Docker Interview Questions	Mysql Interview Questions
Css Interview Questions	Laravel Interview Questions	Asp Net Interview Questions
Django Interview Questions	Dot Net Interview Questions	Kubernetes Interview Questions
Operating System Interview Questions	React Native Interview Questions	Aws Interview Questions
Git Interview Questions	Java 8 Interview Questions	Mongodb Interview Questions
Dbms Interview Questions	Spring Boot Interview Questions	Power Bi Interview Questions
Pl Sql Interview Questions	Tableau Interview Questions	Linux Interview Questions
Ansible Interview Questions	Java Interview Questions	Jenkins Interview Questions

© Copyright by Interviewbit