NLP Interview Questions

Table of Contents

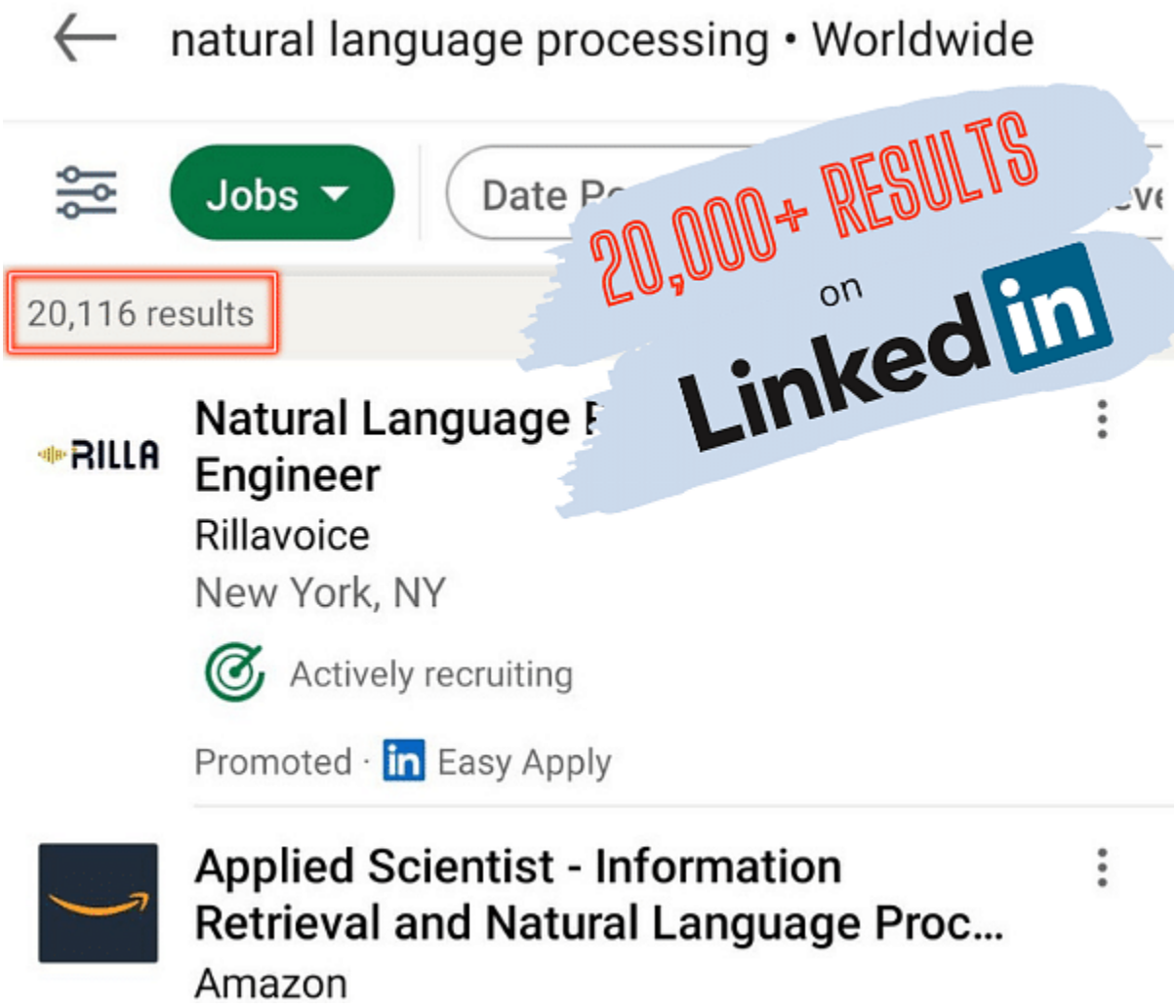
Here is a list of NLP research engineer interview questions with answers that will help you ace all kinds of NLP Interview Questions. The interview questions in NLP have been divided into subgroups for your convenience. So get your tickets of time and take the giant leap towards landing your dream job of becoming an NLP Engineer.

Most people start their mornings with an energetic morning walk and a bit of grocery shopping. In the grocery store, if there is an item they want to find and cannot understand the foreign language written on the packaging, they quickly take out their phone. They open the Google Translate app, and voila! They are able to decide whether to buy the product or not. The app is of great help for someone with specific allergies and people living in foreign countries.

The Google Translate app is an excellent example of Natural Language Processing (NLP) applications. And, if such apps fill you with the zeal of designing something more intelligent, then NLP is the field for you. NLP is a subdomain of Artificial Intelligence that deals with making a machine (a computer) understand the way human beings speak and write the language in their everyday lives.

NLP has recently gained popularity because it can automate tasks for organizations and save time for them. It is an exciting technology that is here to stay for a long time. Thus, it'd be a great option to consider becoming a NLP Research Engineer , Data Scientist, Machine Learning Engineer as a career option to explore given the vast opportunities.

As per Fortune Business Insights, the global artificial intelligence market is expected to climb $266.92 Billion by 2027. A survey conducted by Gartner revealed in 2019 that 37% of the surveyed companies have started implementing AI in their day-to-day tasks, thus signifying a 270% increase in the last four years (w.r.t. 2019). Do a quick search on LinkedIn, and don't be surprised to notice that there are about 20000+ jobs for NLP Engineer/Researcher.



All these stats suggest that now is the perfect time to explore a career in AI and Machine Learning. And if NLP is the subdomain that thrills you and you have already made up your mind for it, then you are on the right page to prepare for your next dream job role that requires NLP skills.

50 NLP Interview Questions and Answers

Below you'll find those NLP interview questions answers that most recruiters ask. These interview questions in NLP are primarily straightforward and are often asked at the beginning of a data science or machine learning interview.

Common NLP Interview Questions with Answers

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.

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.

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 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

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.

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

$$TF*IDF(w)=nw\times lognm$$

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 $X = x_1, x_2, x_3,....,x_n$, the perplexity is given by,

$$PP(X)=P(x1,x2,\ldots,xN)-1N$$

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 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.

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 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

5. Removing Stopwords

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.

11. Use the minimum distance algorithm to show how many editing steps it will take for the word 'intention' to transform into 'execution'.

or

Calculate the Levenshtein distance between two sequences 'intention' and 'execution'.

The image above can be used to understand the number of editing steps it will take for the word intention to transform into execution.

1. The first step is deletion (d) of 'I.'
2. The next step is to substitute (s) the letter 'N' with 'E.'
3. Replace the letter 'T' with 'X.'
4. The letter E remains unchanged, and the letter 'C' is inserted (i).
5. Substitute 'U' for the letter 'N.'

Thus, it will take five editing steps for transformations, and the Levenshtein distance is five.

12. What are the full listing hypothesis and minimum redundancy hypothesis?

These are the two hypotheses relating to the way humans store words of a language in their memory.

Full Listing Hypothesis: This hypothesis suggests that all humans perceive all the words in their memory without any internal morphological structure. So, words like tire, tiring, tired are all stored separately in the mental lexicon.

Minimum Redundancy Hypothesis: This hypothesis proposes that only the raw form of the words (morphemes) form the part of the mental lexicon. When humans process a word like tired, they recall both the morphemes (tire-d).

The interview round is of course the most important round that an applicant must focus on. But, without any hands-on experience in solving real-world problems, it would be difficult for you to clear the technical rounds. Check out these solved end-to-end NLP Projects from our repository that will guide you through the exciting applications of NLP in the tech world.