Top 100 NLP Questions

Steve Nouri

Q1. 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.

Q2. 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 a versa.
E.g. cosine angle between two words “Football” and “Cricket” will be closer to 1 as compared to angle between the words “Football” and “New Delhi”

Q3. 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.

Q4. 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)

Q5. 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)

Q6. Dissimilarity between words expressed using cosine similarity will have values significantly higher than 0.5

a. True
b. False

Answer : a)

Q7. Which one of the following are 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) are not keyword Normalization techniques. Named Entity help 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.

Q8. Which of the below are NLP use cases?

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

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

Q9. 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 correct choice is Klog(3)/T

Q10. 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)

Q11. 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

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

Q12. 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.

Q13. 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.

Q14 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

Q15 TF-IDF helps you to establish?

a. most frequently occurring word in the document
b. 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 offset by the number of documents that appear in the corpus.

- TF is the frequency of term divided by a 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.

Q16 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)

Q17 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.

Q18 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.

Q19. In NLP, Words represented as vectors are called as Neural Word Embeddings

a. True
b. False
Answer : a) Word2Vec, GloVe based models build word embedding vectors that are multidimensional.

Q20. 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.

Q21. 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.

Q22. 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.

Q23. Word embeddings capture multiple dimensions of data and are represented as vectors

a. True
b. False

Answer : a)

Q24. In NLP, Word embedding vectors help establish distance between two tokens

a. True
b. False

Answer : a) One can use Cosine similarity to establish distance between two vectors represented through Word Embeddings

Q25. 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.

Q26. 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 pattern 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.

Q27. 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

Q28. 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.

Q29. Same word can have multiple word embeddings possible with _____?

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

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

Q30 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.

Q31. 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.

Q32. Uses unidirectional language model for producing word embedding

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

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

Q33. 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

Answer : 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.

Q34. 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

Q35. Transformer model pays attention to the most important word in Sentence

a. True
b. False

Answer : 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.

Q36. Which NLP model gives the best accuracy amongst the following?

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

Answer : b) 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.

Q37. Permutation Language models is a feature of

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

Answer : 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.

Q38. Transformer XL uses relative positional embedding

a. True
b. False
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.

Q39. What is Naive Bayes algorithm, When we can use this algorithm in NLP?

Naive Bayes algorithm 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.

Q40. 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.

Q41. What is text Summarization?

Text summarization 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

Q42. 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

Q43. 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

Q44. What is Bag of Words?

Bag of Words 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.

Q45. 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.

Q46. What is a 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.

Q48. What are the best NLP Tools?

Some of the best NLP tools from open sources are:

- SpaCy
- TextBlob
- Textacy
- Natural language Toolkit
- Retext
- NLP.js
- Stanford NLP
- CogcompNLP

Q49. What is POS tagging?

Parts of speech tagging better known as POS tagging refers to the process of identifying specific words in a document and group 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 generic process used for word mapping is quite ineffective for POS tagging because of the same reason.

Q50. What is NES?

Name entity recognition is more commonly known as NER is the process of identifying specific entities in a text document which are more informative and have a unique context. These often denote places, people, organisations, 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 categorise them under different predefined classes. This step further helps in extracting information.

Q51 Explain the Masked Language Model?

Masked language modelling is the process in which the output is taken from the corrupted input. This model helps the learners to master the deep representations in downstream tasks. You can predict a word from the other words of the sentence using this model.

Q52 What is pragmatic analysis in NLP?

Pragmatic Analysis: It deals with outside word knowledge, which means knowledge that is external to the documents and/or queries. Pragmatics analysis that focuses on what was described is reinterpreted by what it actually meant, deriving the various aspects of language that require real-world knowledge.

Q53 What is perplexity in NLP?

The word "perplexed" means "puzzled" or "confused", thus Perplexity in general means the inability to tackle something complicated and a problem that is not specified. Therefore, Perplexity in NLP is a way to determine the extent of uncertainty in predicting some text.
In NLP, perplexity is a way of evaluating language models. Perplexity can be high and low; Low perplexity is ethical because the inability to deal with any complicated problem is less while high perplexity is terrible because the failure to deal with a complicated is high.

Q54 What is ngram in NLP?

N-gram in NLP is simply a sequence of n words, and we also conclude the sentences which appeared more frequently, for example, let us consider the progression of these three words:

- New York (2 gram)
- The Golden Compass (3 gram)
- She was there in the hotel (4 gram)

Now from the above sequence, we can easily conclude that sentence (a) appeared more frequently than the other two sentences, and the last sentence(c) is not seen that often. Now if we assign probability in the occurrence of an n-gram, then it will be advantageous. It would help in making next-word predictions and in spelling error corrections.

Q55 Explain differences between AI, Machine Learning and NLP

Artificial Intelligence	Machine Learning	Natural Language Processing
It is the technique to create smarter machines	Machine Learning is the term used for systems that learn from experience.	This is the set of system that has the ability to understand the language
AI includes human intervention	Machine Learning purely involves the working of computers and no human intervention.	NLP links both computer and human languages.
Artificial intelligence is a broader concept than Machine Learning	ML is a narrow concept and is a subset of AI.	

Q56 Why self-attention is awesome?

“In terms of computational complexity, self-attention layers are faster than recurrent layers when the sequence length n is smaller than the representation dimensionality d, which is most often the case with sentence representations used by state-of-the-art models in machine translations, such as word-piece and byte-pair representations.” — from Attention is all you need

Q57 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.

Q58 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.

Q60 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 {ε} is a regular language, then ε 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.

Q61 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:

Q62 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

Q63. There have some various common elements of natural language processing. Those elements are very important for understanding NLP properly, can you please explain the same in details with an example?

Answer:

There have a lot of components normally using by natural language processing (NLP). Some of the major components are explained below:

- Extraction of Entity: It actually identifying and extracting some critical data from the available information which help to segmentation of provided sentence on identifying each entity. It can help in identifying one human that it's fictional or real, same kind of reality identification for any organization, events or any geographic location etc.
- The analysis in a syntactic way: it mainly helps for maintaining ordering properly of the available words.

Q64 In the case of processing natural language, we normally mentioned one common terminology NLP and binding every language with the same terminology properly. Please explain in details about this NLP terminology with an example?

Answer:

This is the basic NLP Interview Questions asked in an interview. There have some several factors available in case of explaining natural language processing. Some of the key factors are given below:

- Vectors and Weights: Google Word vectors, length of TF-IDF, varieties documents, word vectors, TF-IDF.
- Structure of Text: Named Entities, tagging of part of speech, identifying the head of the sentence.
- Analysis of sentiment: Know about the features of sentiment, entities available for the sentiment, sentiment common dictionary.
- Classification of Text: Learning supervising, set off a train, set of validation in Dev, Set of define test, a feature of the individual text, LDA.
- Reading of Machine Language: Extraction of the possible entity, linking with an individual entity, DBpedia, some libraries like Pikes or FRED.

Q65 Explain briefly about word2vec

Word2Vec embeds words in a lower-dimensional vector space using a shallow neural network. The result is a set of word-vectors where vectors close together in vector space have similar meanings based on context, and word-vectors distant to each other have differing meanings. For example, apple and orange would be close together and apple and gravity would be relatively far.

There are two versions of this model based on skip-grams (SG) and continuous-bag-of-words (CBOW).

Q66 What are the metrics used to test an NLP model?

Accuracy, Precision, Recall and F1. Accuracy is the usual ratio of the prediction to the desired output. But going just be accuracy is naive considering the complexities involved.

Q67 What are some ways we can preprocess text input?

Here are several preprocessing steps that are commonly used for NLP tasks:

- case normalization: we can convert all input to the same case (lowercase or uppercase) as a way of reducing our text to a more canonical form
- punctuation/stop word/white space/special characters removal: if we don't think these words or characters are relevant, we can remove them to reduce the feature space
- lemmatizing/stemming: we can also reduce words to their inflectional forms (i.e. walks → walk) to further trim our vocabulary
- generalizing irrelevant information: we can replace all numbers with a <NUMBER> token or all names with a <NAME> token

Q68 How does the encoder-decoder structure work for language modelling?

The encoder-decoder structure is a deep learning model architecture responsible for several state of the art solutions, including Machine Translation.
The input sequence is passed to the encoder where it is transformed to a fixed-dimensional vector representation using a neural network. The transformed input is then decoded using another neural network. Then, these outputs undergo another transformation and a softmax layer. The final output is a vector of probabilities over the vocabularies. Meaningful information is extracted based on these probabilities.

Q69 What are attention mechanisms and why do we use them?

This was a followup to the encoder-decoder question. Only the output from the last time step is passed to the decoder, resulting in a loss of information learned at previous time steps. This information loss is compounded for longer text sequences with more time steps.
Attention mechanisms are a function of the hidden weights at each time step. When we use attention in encoder-decoder networks, the fixed-dimensional vector passed to the decoder becomes a function of all vectors outputted in the intermediary steps.
Two commonly used attention mechanisms are additive attention and multiplicative attention. As the names suggest, additive attention is a weighted sum while multiplicative attention is a weighted multiplier of the hidden weights. During the training process, the model also learns weights for the attention mechanisms to recognize the relative importance of each time step.

Q70 How would you implement an NLP system as a service, and what are some pitfalls you might face in production?

This is less of a NLP question than a question for productionizing machine learning models. There are however certain intricacies to NLP models.

Without diving too much into the productionization aspect, an ideal Machine Learning service will have:

- endpoint(s) that other business systems can use to make inference
- a feedback mechanism for validating model predictions
- a database to store predictions and ground truths from the feedback
- a workflow orchestrator which will (upon some signal) re-train and load the new model for serving based on the records from the database + any prior training data
- some form of model version control to facilitate rollbacks in case of bad deployments
- post-production accuracy and error monitoring

Q71 How can we handle misspellings for text input?

By using word embeddings trained over a large corpus (for instance, an extensive web scrape of billions of words), the model vocabulary would include common misspellings by design. The model can then learn the relationship between misspelled and correctly spelled words to recognize their semantic similarity.

We can also preprocess the input to prevent misspellings. Terms not found in the model vocabulary can be mapped to the "closest" vocabulary term using:

- edit distance between strings
- phonetic distance between word pronunciations
- keyword distance to catch common typos

Q72 Which of the following models can perform tweet classification with regards to context mentioned above?

A) Naive Bayes
B) SVM
C) None of the above
Solution: (C)

Since, you are given only the data of tweets and no other information, which means there is no target variable present. One cannot train a supervised learning model, both svm and naive bayes are supervised learning techniques.

Q73 You have created a document term matrix of the data, treating every tweet as one document. Which of the following is correct, in regards to document term matrix?

1. Removal of stopwords from the data will affect the dimensionality of data
2. Normalization of words in the data will reduce the dimensionality of data

3. Converting all the words in lowercase will not affect the dimensionality of the data

A) Only 1
B) Only 2
C) Only 3
D) 1 and 2
E) 2 and 3
F) 1, 2 and 3
Solution: (D)
Choices A and B are correct because stopword removal will decrease the number of features in the matrix, normalization of words will also reduce redundant features, and, converting all words to lowercase will also decrease the dimensionality.

Q74 Which of the following features can be used for accuracy improvement of a classification model?

A) Frequency count of terms
B) Vector Notation of sentence
C) Part of Speech Tag
D) Dependency Grammar
E) All of these
Solution: (E)
All of the techniques can be used for the purpose of engineering features in a model.

Q75 What percentage of the total statements are correct with regards to Topic Modeling?

1. It is a supervised learning technique
2. LDA (Linear Discriminant Analysis) can be used to perform topic modeling
3. Selection of number of topics in a model does not depend on the size of data
4. Number of topic terms are directly proportional to size of the data

A) 0
B) 25
C) 50
D) 75
E) 100
Solution: (A)
LDA is unsupervised learning model, LDA is latent Dirichlet allocation, not Linear discriminant analysis. Selection of the number of topics is directly proportional to the size of the data, while number of topic terms is not directly proportional to the size of the data. Hence none of the statements are correct.

Q76 In Latent Dirichlet Allocation model for text classification purposes, what does alpha and beta hyperparameter represent-

A) Alpha: number of topics within documents, beta: number of terms within topics False
B) Alpha: density of terms generated within topics, beta: density of topics generated within terms False
C) Alpha: number of topics within documents, beta: number of terms within topics False
D) Alpha: density of topics generated within documents, beta: density of terms generated within topics True
Solution: (D)
Option D is correct

Q77 What is the problem with ReLu?

- Exploding gradient(Solved by gradient clipping)
- Dying ReLu — No learning if the activation is 0 (Solved by parametric relu)
- Mean and variance of activations is not 0 and 1.(Partially solved by subtracting around 0.5 from activation. Better explained in fastai videos)

Q78 What is the difference between learning latent features using SVD and getting embedding vectors using deep network?

SVD uses linear combination of inputs while a neural network uses nonlinear combination.

Q79 What is the information in the hidden and cell state of LSTM?

Hidden stores all the information till that time step and cell state stores particular information that might be needed in the future time step.
Number of parameters in an LSTM model with bias
$4(mh+h^2+h)$ where m is input vectors size and h is output vectors size a.k.a. hidden
The point to see here is that mh dictates the model size as m>>h. Hence it's important to have a small vocab.
Time complexity of LSTM
seq_length*hidden²
Time complexity of transfomer
seq_length²*hidden
When hidden size is more than the seq_length(which is normally the case), transfomer is faster than LSTM.

Q80 When is self-attention not faster than recurrent layers?

When the sequence length is greater than the representation dimensions. This is rare.

Q81 What is the benefit of learning rate warm-up?

Learning rate warm-up is a learning rate schedule where you have low (or lower) learning rate at the beginning of training to avoid divergence due to unreliable gradients at the beginning. As the model becomes more stable, the learning rate would increase to speed up convergence.

Q82 What's the difference between hard and soft parameter sharing in multi-task learning?

Hard sharing is where we train for all the task at the same time and update our weights using all the losses whereas soft sharing is where we train for one task at a time.

Q83 What's the difference between BatchNorm and LayerNorm?

BatchNorm computes the mean and variance at each layer for every minibatch whereas LayerNorm computes the mean and variance for every sample for each layer independently. Batch normalisation allows you to set higher learning rates, increasing speed of training as it reduces the unstability of initial starting weights.

Q84 Difference between BatchNorm and LayerNorm?

BatchNorm — Compute the mean and var at each layer for every minibatch
LayerNorm — Compute the mean and var for every single sample for each layer independently

Q85 Why does the transformer block have LayerNorm instead of BatchNorm?

Looking at the advantages of LayerNorm, it is robust to batch size and works better as it works at the sample level and not batch level.

Q86 What changes would you make to your deep learning code if you knew there are errors in your training data?

We can do label smoothening where the smoothening value is based on % error. If any particular class has known error, we can also use class weights to modify the loss.

Q87 What are the tricks used in ULMFiT? (Not a great questions but checks the awareness)

- LM tuning with task text
- Weight dropout
- Discriminative learning rates for layers
- Gradual unfreezing of layers
- Slanted triangular learning rate schedule

This can be followed up with a question on explaining how they help.

Q88 Tell me a language model which doesn't use dropout

ALBERT v2 — This throws a light on the fact that a lot of assumptions we take for granted are not necessarily true. The regularisation effect of parameter sharing in ALBERT is so strong that dropouts are not needed. (ALBERT v1 had dropouts.)

Q89 What are the differences between GPT and GPT-2? (From Lilian Weng)

- Layer normalization was moved to the input of each sub-block, similar to a residual unit of type "building block" (differently from the original type "bottleneck", it has batch normalization applied before weight layers).
- An additional layer normalization was added after the final self-attention block.
- A modified initialization was constructed as a function of the model depth.
- The weights of residual layers were initially scaled by a factor of 1/√n where n is the number of residual layers.
- Use larger vocabulary size and context size.

Q90 What are the differences between GPT and BERT?

- GPT is not bidirectional and has no concept of masking
- BERT adds next sentence prediction task in training and so it also has a segment embedding

Q91 What are the differences between BERT and ALBERT v2?

- Embedding matrix factorisation(helps in reducing no. of parameters)
- No dropout
- Parameter sharing(helps in reducing no. of parameters and regularisation)

Q92 How does parameter sharing in ALBERT affect the training and inference time?

No effect. Parameter sharing just decreases the number of parameters.

Q93 How would you reduce the inference time of a trained NN model?

- Serve on GPU/TPU/FPGA
- 16 bit quantisation and served on GPU with fp16 support
- Pruning to reduce parameters
- Knowledge distillation (To a smaller transformer model or simple neural network)
- Hierarchical softmax/Adaptive softmax
- You can also cache results as explained here.

Q94 Would you use BPE with classical models?

Of course! BPE is a smart tokeniser and it can help us get a smaller vocabulary which can help us find a model with less parameters.

Q95 How would you make an arxiv papers search engine? (I was asked — How would you make a plagiarism detector?)

Get top k results with TF-IDF similarity and then rank results with

- semantic encoding + cosine similarity
- a model trained for ranking

Q96 Get top k results with TF-IDF similarity and then rank results with

- semantic encoding + cosine similarity
- a model trained for ranking

Q97 How would you make a sentiment classifier?

This is a trick question. The interviewee can say all things such as using transfer learning and latest models but they need to talk about having a neutral class too otherwise you can have really good accuracy/f1 and still, the model will classify everything into positive or negative.
The truth is that a lot of news is neutral and so the training needs to have this class. The interviewee should also talk about how he will create a dataset and his training strategies like the selection of language model, language model fine-tuning and using various datasets for multi-task learning.

Q98 What is the difference between regular expression and regular grammar?

A regular expression is the representation of natural language in the form of mathematical expressions containing a character sequence. On the other hand, regular grammar is the generator of natural language, defining a set of defined rules and syntax which the strings in the natural language must follow.

Q99 Why should we use Batch Normalization?

Once the interviewer has asked you about the fundamentals of deep learning architectures, they would move on to the key topic of improving your deep learning model's performance.
Batch Normalization is one of the techniques used for reducing the training time of our deep learning algorithm. Just like normalizing our input helps improve our logistic regression model, we can normalize the activations of the hidden layers in our deep learning model as well:

Q100 How is backpropagation different in RNN compared to ANN?

In Recurrent Neural Networks, we have an additional loop at each node:
This loop essentially includes a time component into the network as well. This helps in capturing sequential information from the data, which could not be possible in a generic artificial neural network.
This is why the backpropagation in RNN is called Backpropagation through Time, as in backpropagation at each time step.