Natural Language Processing

Natural language processing aims to create machines that interpret and respond to text or voice input in the same manner that people do—and respond with text or speech of their own.

- ## Definition

  - Natural language processing (NLP) is a subject of computer science specifically, a branch of artificial intelligence (AI) concerning the ability of computers to understand text and spoken words in the same manner that humans can.

  - 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 Task

  - Speech Recognition
  - Part of speech tagging
  - Word sense disambiguation
  - Named entity recognition
  - Sentiment analysis
  - Natural Language Generation

- ## Important NLP Tools

  - CoreNLP from Stanford group
  - NLTK python library
  - Textblob
  - Gensim
  - Spacy
  - Textacy
  - Pytorch_NLP

- ## NLP Cases

  - Spam detection
  - Machine translation
  - Virtual agents and chatbots

- Social media content analysis
- Text summarization

Important Interview questions:

1. What is Natural Language Processing (NLP)?

Natural Language Processing, or NLP, is an automated method for understanding and analysing natural languages, as well as extracting essential information from them using machine learning algorithms.

2. List out the important components in NLP?

- Entity extraction - It involves segmenting a sentence to identify and extract entities, such as a person (real or fictional), organization, geographies, events, etc.
- Syntactic analysis - It refers to the proper ordering of words.
- Pragmatic analysis - Pragmatic Analysis is part of the process of extracting information from text.

3. Define the terminology in NLP?

NLP Terminology is based on the following factors:

Weights and Vectors: TF-IDF, length (TF-IDF, doc), Word Vectors, Google Word Vectors

Text Structure: Part-Of-Speech Tagging, Head of sentence, Named entities

Sentiment Analysis: Sentiment Dictionary, Sentiment Entities, Sentiment Features

Text Classification: Supervised Learning, Train Set, Dev(=Validation) Set, Test Set, Text Features, LDA.

Machine Reading: Entity Extraction, Entity Linking, dbpedia, FRED (lib) / Pikes

4. What is tokenization in NLP?

The process of splitting the words and sentences is called tokenization.

5. What is the difference between stemming and lemmatization?

Both stemming and lemmatization are keyword normalizing approaches that try to reduce morphological variance between words in a phrase.

Stemming	Lemmatization

This method includes deleting the affixes that have been added to a word, leaving only the rest of the term. Eg: beautiful- beauty	The process of turning an inflected word into its lemma is known as lemmatization. Eg: beautiful- beauti

6. How would you extract the features from a corpus for NLP?

- Bag of words
- Word embedding

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

LSI is a technique that examines a collection of documents to determine the statistical coexistence of words that appear in the same context. It provides information about the issues covered in such documents.

8. What is 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.

Higher the perplexing denotes the less information the language model conveys.

9. Which algorithm in the NLP supports the bidirectional context?

Bidirectional Encoder Representations and Transformers (BERT) is used in the NLP which supports the bidirectional context.

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?

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 prior words.

13. What is named entity recognition?

Named entity recognition is a natural language processing technique that can automatically scan entire articles and pull out some fundamental entities in a text and classify them into predefined categories.

14. Do you have any idea about fasttext?

FastText is a library created by the Facebook Research Team for efficient learning of word representations and sentence classification. It can give the vector representations for the words not present in the dictionary (OOV words) since these can also be broken down into character n-grams.

15. What is Glove?

GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.

16. What is context2vec?

Context2vec is an unsupervised model for learning generic context embedding of wide sentential contexts, using a bidirectional LSTM. ... In contrast to word2vec that use context modeling mostly internally and considers the target word embeddings as their main output, the focus of context2vec is the context representation.

17. Explain how we can do parsing

Parsing is the method to identify and understand the syntactic structure of a text. It is done by analysing 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.

18. 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: I am going to play cricket

Unigram	I	Am	Going	To	Play	Cricket
Bi-gram	I am	am going	going to	to play	Play cricket	
Tri-gram	I am going	am going to	going to play	To play cricket		

19. What is tokenization in NLP? How to tokenize a sentence?

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

20. 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.

21. What is web scraping in NLP and mention the libraries used to do it?

Web scraping refers to the extraction of data from a website. In most cases, this is done using software tools such as web scrapers. Once the data is scraped, you'd usually then export it in a more convenient format such as an Excel spreadsheet or JSON.

Libraries used requests, NLTK and Beautifulsoup.