CBSE | DEPARTMENT OF SKILL EDUCATION

ARTIFICIAL INTELLIGENCE
QUESTION BANK – CLASS 10

CHAPTER 7: NATURAL LANGUAGE PROCESSING

One (01) Mark Questions

1. What is a Chabot?

A chatbot is a computer program that's designed to simulate human conversation through voice commands or text chats or both. Eg: Mitsuku Bot, Jabberwacky etc.

OR

A chatbot is a computer program that can learn over time how to best interact with humans. It can answer questions and troubleshoot customer problems, evaluate and qualify prospects, generate sales leads and increase sales on an ecommerce site.

OR

A chatbot is a computer program designed to simulate conversation with human users. A chatbot is also known as an artificial conversational entity (ACE), chat robot, talk bot, chatterbot or chatterbox.

OR

A chatbot is a software application used to conduct an on-line chat conversation via text or text-to-speech, in lieu of providing direct contact with a live human agent.

2. What is the full form of NLP?

Natural Language Processing

3. While working with NLP what is the meaning of?

a. Syntax
b. Semantics

Syntax: Syntax refers to the grammatical structure of a sentence.
Semantics: It refers to the meaning of the sentence.

4. What is the difference between stemming and lemmatization?

Stemming is a technique used to extract the base form of the words by removing affixes from them. It is just like cutting down the branches of a tree to its stems. For example, the stem of the words *eating, eats, eaten* is *eat*.

Lemmatization is the grouping together of different forms of the same word. In search queries, lemmatization allows end users to query any version of a base word and get relevant results.

OR

Stemming is the process in which the affixes of words are removed and the words are converted to their base form.

In lemmatization, the word we get after affix removal (also known as lemma) is a meaningful one. Lemmatization makes sure that lemma is a word with meaning and hence it takes a longer time to execute than stemming.

OR

Stemming algorithms work by cutting off the end or the beginning of the word, taking into account a list of common prefixes and suffixes that can be found in an inflected word.

Lemmatization on the other hand, takes into consideration the morphological analysis of the words. To do so, it is necessary to have detailed dictionaries which the algorithm can look through to link the form back to its lemma.

5. What is the full form of TFIDF?

Term Frequency and Inverse Document Frequency

6. What is meant by a dictionary in NLP?

Dictionary in NLP means a list of all the unique words occurring in the corpus. If some words are repeated in different documents, they are all written just once as while creating the dictionary.

7. What is term frequency?

Term frequency is the frequency of a word in one document. Term frequency can easily be found from the document vector table as in that table we mention the frequency of each word of the vocabulary in each document.

8. Which package is used for Natural Language Processing in Python programming?

Natural Language Toolkit (NLTK). NLTK is one of the leading platforms for building Python programs that can work with human language data.

9. What is a document vector table?

Document Vector Table is used while implementing Bag of Words algorithm.

In a document vector table, the header row contains the vocabulary of the corpus and other rows correspond to different documents.

If the document contains a particular word it is represented by 1 and absence of word is represented by 0 value.

OR

Document Vector Table is a table containing the frequency of each word of the vocabulary in each document.

10. What do you mean by corpus?

In Text Normalization, we undergo several steps to normalize the text to a lower level. That is, we will be working on text from multiple documents and the term used for the whole textual data from all the documents altogether is known as corpus.

OR

A corpus is a large and structured set of machine-readable texts that have been produced in a natural communicative setting.

OR

A corpus can be defined as a collection of text documents. It can be thought of as just a bunch of text files in a directory, often alongside many other directories of text files.

Two (02) Mark Questions

1. What are the types of data used for Natural Language Processing applications?

Natural Language Processing takes in the data of Natural Languages in the form of written words and spoken words which humans use in their daily lives and operates on this.

2. Differentiate between a script-bot and a smart-bot. (Any 2 differences)

Script-bot	Smart-bot
• A scripted chatbot doesn't carry even a glimpse of A.I • Script bots are easy to make • Script bot functioning is very limited as they are less powerful. • Script bots work around a script which is programmed in them • No or little language processing skills • Limited functionality	• Smart bots are built on NLP and ML. • Smart –bots are comparatively difficult to make. • Smart-bots are flexible and powerful. • Smart bots work on bigger databases and other resources directly • NLP and Machine learning skills are required. • Wide functionality

3. Give an example of the following:

- Multiple meanings of a word
- Perfect syntax, no meaning

- **Example of Multiple meanings of a word –**

His face turns red after consuming the medicine

Meaning - Is he having an allergic reaction? Or is he not able to bear the taste of that medicine?

- **Example of Perfect syntax, no meaning-**

Chickens feed extravagantly while the moon drinks tea.

This statement is correct grammatically but it does not make any sense. In Human language, a perfect balance of syntax and semantics is important for better understanding.

4. What is inverse document frequency?

To understand inverse document frequency, first we need to understand document frequency.

Document Frequency is the number of documents in which the word occurs irrespective of how many times it has occurred in those documents.

In case of inverse document frequency, we need to put the document frequency in the denominator while the total number of documents is the numerator.

For example, if the document frequency of a word "AMAN" is 2 in a particular document then its inverse document frequency will be 3/2. (Here no. of documents is 3)

5. Define the following:

- Stemming
- Lemmatization

Stemming: Stemming is a rudimentary rule-based process of stripping the suffixes ("ing", "ly", "es", "s" etc) from a word.

Stemming is a process of reducing words to their word stem, base or root form (for example, books — book, looked — look).

Lemmatization: Lemmatization, on the other hand, is an organized & step by step procedure of obtaining the root form of the word, it makes use of vocabulary (dictionary importance of words) and morphological analysis (word structure and grammar relations).

The aim of lemmatization, like stemming, is to reduce inflectional forms to a common base form. As opposed to stemming, lemmatization does not simply chop off inflections. Instead it uses lexical knowledge bases to get the correct base forms of words.

OR

Stemming is a technique used to extract the base form of the words by removing affixes from them. It is just like cutting down the branches of a tree to its stems. For example, the stem of the words *eating, eats, eaten* is *eat*.
Lemmatization is the grouping together of different forms of the same word. In search queries, lemmatization allows end users to query any version of a base word and get relevant results.

OR

Stemming is the process in which the affixes of words are removed and the words are converted to their base form.
In lemmatization, the word we get after affix removal (also known as lemma) is a meaningful one. Lemmatization makes sure that lemma is a word with meaning and hence it takes a longer time to execute than stemming.

OR

Stemming algorithms work by cutting off the end or the beginning of the word, taking into account a list of common prefixes and suffixes that can be found in an inflected word.
Lemmatization on the other hand, takes into consideration the morphological analysis of the words. To do so, it is necessary to have detailed dictionaries which the algorithm can look through to link the form back to its lemma.

6. What do you mean by document vectors?

Document Vector contains the frequency of each word of the vocabulary in a particular document.
In document vector vocabulary is written in the top row. Now, for each word in the document, if it matches with the vocabulary, put a 1 under it. If the same word appears again, increment the previous value by 1. And if the word does not occur in that document, put a 0 under it.

7. What is TFIDF? Write its formula.

Term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

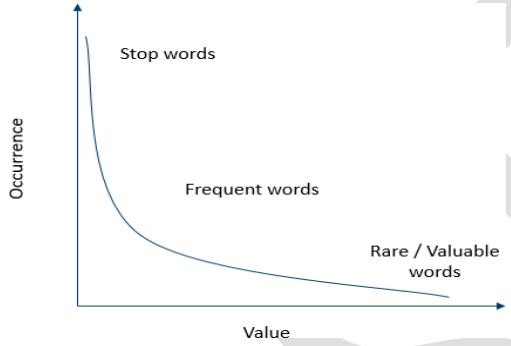
The number of times a word appears in a document divided by the total number of words in the document. Every document has its own term frequency.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

8. Which words in a corpus have the highest values and which ones have the least?

Stop words like - and, this, is, the, etc. have highest values in a corpus. But these words do not talk about the corpus at all. Hence, these are termed as stopwords and are mostly removed at the pre-processing stage only.

Rare or valuable words occur the least but add the most importance to the corpus. Hence, when we look at the text, we take frequent and rare words into consideration.



9. Does the vocabulary of a corpus remain the same before and after text normalization? Why?

No, the vocabulary of a corpus does not remain the same before and after text normalization. Reasons are –

- In normalization the text is normalized through various steps and is lowered to minimum vocabulary since the machine does not require grammatically correct statements but the essence of it.
- In normalization Stop words, Special Characters and Numbers are removed.
- In stemming the affixes of words are removed and the words are converted to their base form.

So, after normalization, we get the reduced vocabulary.

10. What is the significance of converting the text into a common case?

In Text Normalization, we undergo several steps to normalize the text to a lower level. After the removal of stop words, we convert the whole text into a similar case, preferably lower case. This ensures that the case-sensitivity of the machine does not consider same words as different just because of different cases.

11. Mention some applications of Natural Language Processing.

Natural Language Processing Applications-

- Sentiment Analysis.
- Chatbots & Virtual Assistants.
- Text Classification.
- Text Extraction.
- Machine Translation
- Text Summarization
- Market Intelligence
- Auto-Correct

12. What is the need of text normalization in NLP?

Since we all know that the language of computers is Numerical, the very first step that comes to our mind is to convert our language to numbers.

This conversion takes a few steps to happen. The first step to it is Text Normalization.

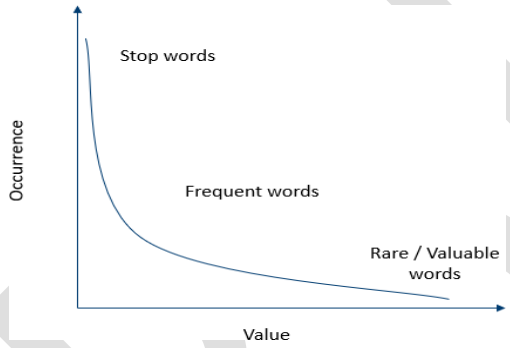
Since human languages are complex, we need to first of all simplify them in order to make sure that the understanding becomes possible. Text Normalization helps in cleaning up the textual data in such a way that it comes down to a level where its complexity is lower than the actual data.

13. Explain the concept of Bag of Words.

Bag of Words is a Natural Language Processing model which helps in extracting features out of the text which can be helpful in machine learning algorithms. In bag of words, we get the occurrences of each word and construct the vocabulary for the corpus.

Bag of Words just creates a set of vectors containing the count of word occurrences in the document (reviews). Bag of Words vectors are easy to interpret.

14. Explain the relation between occurrence and value of a word.



plot of occurrence of words versus their value

As shown in the graph, occurrence and value of a word are inversely proportional. The words which occur most (like stop words) have negligible value. As the occurrence of words drops, the value of such words rises. These words are termed as rare or valuable words. These words occur the least but add the most value to the corpus.

15. What are the applications of TFIDF?

TFIDF is commonly used in the Natural Language Processing domain. Some of its applications are:

- Document Classification - Helps in classifying the type and genre of a document.
- Topic Modelling - It helps in predicting the topic for a corpus.
- Information Retrieval System - To extract the important information out of a corpus.
- Stop word filtering - Helps in removing the unnecessary words out of a text body.

16. What are stop words? Explain with the help of examples.

"Stop words" are the most common words in a language like "the", "a", "on", "is", "all". These words do not carry important meaning and are usually removed from texts. It is possible to remove stop words using Natural Language Toolkit (NLTK), a suite of libraries and programs for symbolic and statistical natural language processing.

17. Differentiate between Human Language and Computer Language.

Humans communicate through language which we process all the time. Our brain keeps on processing the sounds that it hears around itself and tries to make sense out of them all the time.

On the other hand, the computer understands the language of numbers. Everything that is sent to the machine has to be converted to numbers. And while typing, if a single mistake is made, the computer throws an error and does not process that part. The communications made by the machines are very basic and simple.

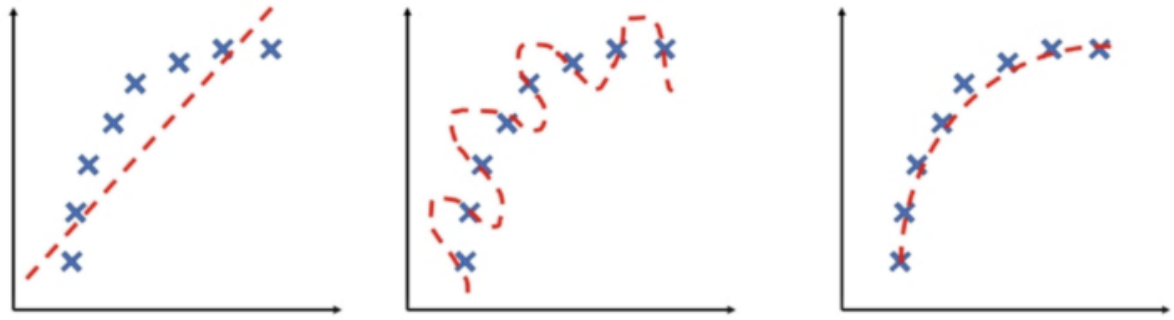
Four 04 Mark Questions

1. Create a document vector table for the given corpus:

Document 1: We are going to Mumbai
Document 2: Mumbai is a famous place.
Document 3: We are going to a famous place.
Document 4: I am famous in Mumbai.

We	Are	going	to	Mumbai	is	a	famous	place	I	am	in
1	1	1	1	1	0	0	0	0	0	0	0
0	0	0	0	1	1	1	1	1	0	0	0
1	1	1	1	0	0	1	1	1	0	0	0
0	0	0	0	1	0	0	1	0	1	1	1

2. Classify each of the images according to how well the model's output matches the data samples:



Here, the red dashed line is model's output while the blue crosses are actual data samples.

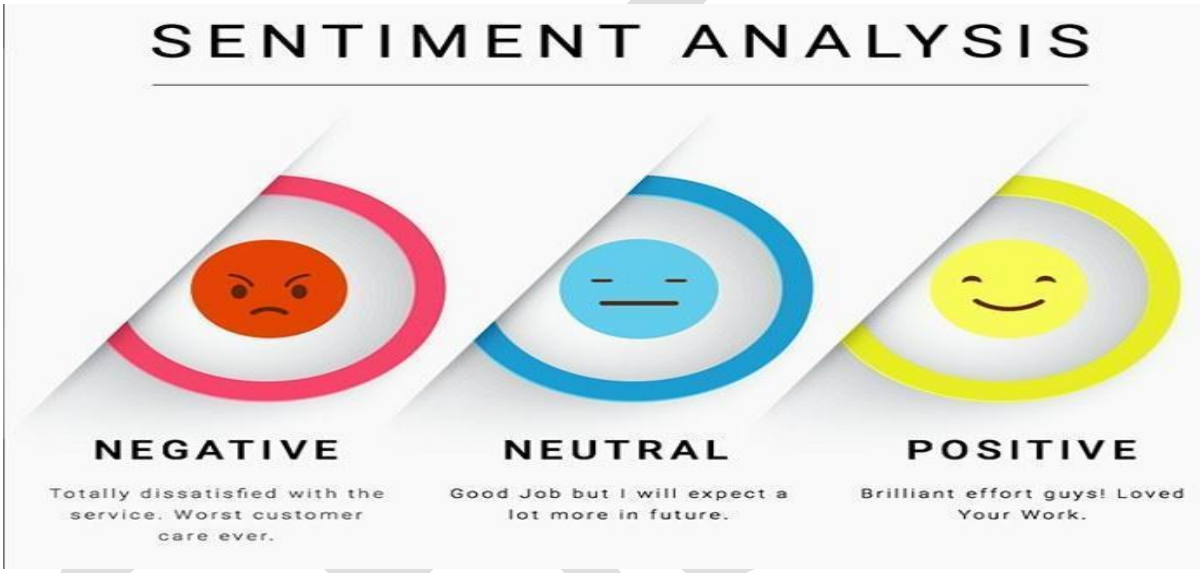
- The model's output does not match the true function at all. Hence the model is said to be under fitting and its accuracy is lower.
- In the second case, model performance is trying to cover all the data samples even if they are out of alignment to the true function. This model is said to be over fitting and this too has a lower accuracy
- In the third one, the model's performance matches well with the true function which states that the model has optimum accuracy and the model is called a perfect fit.

3. Explain how AI can play a role in sentiment analysis of human beings?

The goal of sentiment analysis is to identify sentiment among several posts or even in the same post where emotion is not always explicitly expressed.

Companies use Natural Language Processing applications, such as sentiment analysis, to identify opinions and sentiment online to help them understand what customers think about their products and services (i.e., "I love the new iPhone" and, a few lines later "But sometimes it doesn't work well" where the person is still talking about the iPhone) and overall *

Beyond determining simple polarity, sentiment analysis understands sentiment in context to help better understand what's behind an expressed opinion, which can be extremely relevant in understanding and driving purchasing decisions.



4. Why are human languages complicated for a computer to understand? Explain.

The communications made by the machines are very basic and simple. Human communication is complex. There are multiple characteristics of the human language that might be easy for a human to understand but extremely difficult for a computer to understand.

For machines it is difficult to understand our language. Let us take a look at some of them here:

Arrangement of the words and meaning - There are rules in human language. There are nouns, verbs, adverbs, adjectives. A word can be a noun at one time and an adjective some other time. This can create difficulty while processing by computers.

Analogy with programming language- Different syntax, same semantics: 2+3 = 3+2 Here the way these statements are written is different, but their meanings are the same that is 5. Different semantics, same syntax: 2/3 (Python 2.7) ≠ 2/3 (Python 3) Here the statements written have the same syntax but their meanings are different. In Python 2.7, this statement would result in 1 while in Python 3, it would give an output of 1.5.

Multiple Meanings of a word - In natural language, it is important to understand that a word can have multiple meanings and the meanings fit into the statement according to the context of it.

Perfect Syntax, no Meaning - Sometimes, a statement can have a perfectly correct syntax but it does not mean anything. In Human language, a perfect balance of syntax and semantics is important for better understanding.
These are some of the challenges we might have to face if we try to teach computers how to understand and interact in human language.

5. What are the steps of text Normalization? Explain them in brief.

Text Normalizationin Text Normalization, we undergo several steps to normalize the text to a lower level.

Sentence Segmentation - Under sentence segmentation, the whole corpus is divided into sentences. Each sentence is taken as a different data so now the whole corpus gets reduced to sentences.

Tokenisation- After segmenting the sentences, each sentence is then further divided into tokens. Tokens is a term used for any word or number or special character occurring in a sentence. Under tokenisation, every word, number and special character is considered separately and each of them is now a separate token.
Removing Stop words, Special Characters and Numbers - In this step, the tokens which are not necessary are removed from the token list.
Converting text to a common case -After the stop words removal, we convert the whole text into a similar case, preferably lower case. This ensures that the case-sensitivity of the machine does not consider same words as different just because of different cases.
Stemming In this step, the remaining words are reduced to their root words. In other words, stemming is the process in which the affixes of words are removed and the words are converted to their base form.
Lemmatization -in lemmatization, the word we get after affix removal (also known as lemma) is a meaningful one.

With this we have normalized our text to tokens which are the simplest form of words present in the corpus. Now it is time to convert the tokens into numbers. For this, we would use the Bag of Words algorithm

6. Through a step-by-step process, calculate TFIDF for the given corpus and mention the word(s) having highest value.

Document 1: We are going to Mumbai
Document 2: Mumbai is a famous place.
Document 3: We are going to a famous place.
Document 4: I am famous in Mumbai.

Term Frequency
Term frequency is the frequency of a word in one document. Term frequency can easily be found from the document vector table as in that table we mention the frequency of each word of the vocabulary in each document.

We	Are	Going	to	Mumbai	is	a	famous	Place	I	am	in
1	1	1	1	1	0	0	0	0	0	0	0
0	0	0	0	1	1	1	1	1	0	0	0
1	1	1	1	0	0	1	1	1	0	0	0
0	0	0	0	1	0	0	1	0	1	1	1

Inverse Document Frequency

The other half of TFIDF which is Inverse Document Frequency. For this, let us first understand what does document frequency mean. Document Frequency is the number of documents in which the word occurs irrespective of how many times it has occurred in those documents. The document frequency for the exemplar vocabulary would be:

We	Are	going	to	Mumbai	is	a	Famous	place	I	am	in
2	2	2	2	3	1	2	3	2	1	1	1

Talking about inverse document frequency, we need to put the document frequency in the denominator while the total number of documents is the numerator. Here, the total number of documents are 3, hence inverse document frequency becomes:

We	Are	going	to	Mumbai	is	a	Famous	Place	I	am	in
4/2	4/2	4/2	4/2	4/3	4/1	4/2	4/3	4/2	4/1	4/1	4/1

The formula of TFIDF for any word W becomes:
TFIDF(W) = TF(W) * log (IDF(W))

The words having highest value are – Mumbai, Famous

7. Normalize the given text and comment on the vocabulary before and after the normalization:
Raj and Vijay are best friends. They play together with other friends. Raj likes to play football but Vijay prefers to play online games. Raj wants to be a footballer. Vijay wants to become an online gamer.

Normalization of the given text:
Sentence Segmentation:

1. *Raj and Vijay are best friends.*
2. *They play together with other friends.*
3. *Raj likes to play football but Vijay prefers to play online games.*
4. *Raj wants to be a footballer.*
5. *Vijay wants to become an online gamer.*

Tokenization:

Raj and Vijay are best friends.	*Raj*	*and*	*Vijay*	*are*	*best*	*friends*	*.*
They play together with other friends	*They*	*play*	*Together*	*with*	*other*	*friends*	*.*

Same will be done for all sentences.

Removing Stop words, Special Characters and Numbers:

In this step, the tokens which are not necessary are removed from the token list.

So, the words and, are, to, an, (Punctuation) will be removed.

Converting text to a common case:

After the stop words removal, we convert the whole text into a similar case, preferably lower case.

Here we don't have words in different case so this step is not required for given text.

Stemming:

In this step, the remaining words are reduced to their root words. In other words, stemming is the process in which the affixes of words are removed and the words are converted to their base form.

Word	*Affixes*	*Stem*
Likes	*-s*	*Like*
Prefers	*-s*	*Prefer*
Wants	*-s*	*want*

In the given text Lemmatization is not required.

Given Text

Raj and Vijay are best friends. They play together with other friends. Raj likes to play football but Vijay prefers to play online games. Raj wants to be a footballer. Vijay wants to become an online gamer.

Normalized Text

Raj and Vijay best friends They play together with other friends Raj likes to play football but Vijay prefers to play online games Raj wants to be a footballer Vijay wants to become an online gamer