NAME: AFNAN ATTAR PRN: F1911203 CLASS: B.E COMP II
SUBJECT:- NLP ASSIGNMENT NO.:- 01

Q1) Explain Natural Language Processing. Why is it hard?

Ans 1. Natural language Processing is a machine learning technology that gives computers the ability to interpret, manipulate and comprehend human language.

2. Natural language Processing or NLP is a part of Computer Science, Human language and Artificial Intelligence.

3. There are several factors that make NLP hard :-

- Case Sensitivity or Case Insensitivity.
- Puncutation marks and numbers need special processing.
- Emoticons, hyperlinks, file extensions etc. also need to be given special attention.
- Ambiguity is present in natural languages. It may exist at the level of word, sentence or meaning.
- let us see some types of ambiguity :-
  - Lexical Ambiguity :- Words having multiple meanings.
  - Syntatic Ambiguity :- Sentence being parsed in different ways.
  - Semantic Ambiguity :- Sentence contains ambiguous words.
  - Anaphoric Ambiguity :- Use of anaphore entities in discousse.

Q2) Differentiate between programming languages and Natural languages.

Ans	Natural languages	Programming languages.
1.	Extensive Vocabulary.	Very few words.
2.	Inherently ambigious.	Unambiguious.
3.	Spoken by people.	Intended for machines.

4.	Grammar emerges spontaneously and is complex & nuanced.	Grammar/Syntax is deliberatly crafted & precise.
5.	Translation of language is an art & requires understanding of culture.	Translation of syntax & semantics is more mechanical.

Q5) What is the concept of tokenization, stemming, lemmatization and POS tagging. Explain all with suitable examples.

Ans 1) Tokenization:

- It is the process of breaking down text into the smallest unit called as tokens. For example:-
- "Hello There World" >> Tokenization → "Hello", "There", "World"
- It can be done on the level of word, character or Sub-word also called as n-gram.

2) Stemming:

- It is the process of finding root words. For example Nicely → Nice ; Dancing → Dance ; Greater → Great.
- However it may produce which does not have a semantic meaning.

3) Lemmatization:-

- It works same as stemming but produces accurate root word most of the times unlike stemming. For example :-
- Intelligence —stemming→ Intelligent ; Intelligence —Lemmatization→ Intelligent.

4) POS Tagging:-

- Each word in a text is labelled with its corresponding part of speech using POS tagging.
- It is used to identify grammatical structure of a sentence to disambiguate words that have multiple meanings.
- Example:-

I	want	an	early	upgrade
↓	↓	↓	↓	↓
Pronoun	verb	Determiner	Adjective	Noun

Page No.
Date

NAME: AFNAN ATTAR PRN: F1911203 CLASS: BE COMP 2
SUBJECT: NLP

Q1) Compare syntatic analysis with semantic analysis.

Ans

Syntatic analysis	Semantic analysis
1. Process of analysing strings of symbols conforming to the rules of formal grammar.	Process of checking if generated parse tree is according to the rules of programming language.
2. It performs syntax analysis.	It performs sematic analysis.
3. It is the second phase of compilation process.	It is the thrid phase of compilation process.
4. It takes tokens as input.	It takes parse tree as input.
5. It gives parse tree as output.	It checks correctness of parse tree.
6. It generates a parse tree.	It generates an annotated syntax tree.

Q2) Elaborate syntatic representation of natural language.

Ans 1. Syntatic representation of a natural language refers to the process of representing the structure of text or sentence by the means of formal grammar or set of rules.

2. Syntax analysis checks the text for meaningfulness comparing to the rules of formal grammar.

3. Some formal notations for syntatic representation are as follows:-

i) Phase-structure grammes :

- It represents a sentence as a hierarchy of phrases.
- Relationship between phrases are represented using rules & symbols.

ii) Dependency grammar:
- Sentence is represented as set of words by relationship of dependency.
- Relationship is represented using arrows/lines.

iii) Transformational Grammar:
- It establishes relationship with different elements in the sentence.

iv) Tree diagram:
- The sentence is represented graphically using Trees.

Qs. Discuss relations among lexemes and their senses.

Ans a] Homonym:
- These are words having same spelling but different meanings.
- Example:- "Bat" may be the cricket bat or the flying mammal.

b] Polysemy:-
- Words or phrases with different but related senses.
- Example:- "Bank" can be river bank or the financial bank.

c] Hyponym:-
- A word with more specific meaning than a general term.
- Example:- "Spoon" is a hyponym of cutlery.

d] Synonym:-
- Relationship between two different lexemes with same meaning.
- Example:- Author - Writer, Fate - Destiny.

e] Wordnet:-
- It is a large lexical database of English words.
- It has three databases: Nouns, Verbs & Adjective + Adverbs.

f] Word Sense Disambiguation (WSD):-
- Word sense Disambiguation is the process of choosing right context.
- WSD algorithms take words in context as an input and output correct word sense in context.

Page No.
Date

NAME:- AFNAN ATTAR PRN: F1911203 CLASS:- BE COMP 2
SUBJECT:- NLP ASSIGNMENT No. :- 03

Q1) What is label encoding?

Ans 1. Label encoding is a technique that is used to convert categorical columns into numerical ones so that they can be fitted by machine learning models which only take numerical data.

2. Example :- Suppose we have a column name height that consists of dataset that has elements as Tall, Medium, ~~too~~ Short.

3. After applying label encoding, height column is converted into numerical column.

Height	Height
Tall	0
Medium	1
Short	2

Here, 0 is the label for Tall, 1 is for medium and 2 is for short.

4. Label encoding may lead to priority issues.

Q2) Which are lemmatization methods? Explain any one.

Ans The Lemmatization methods are: (i) WordNet. (ii) TextBlob (iii) WordNet(POS tag). (iv) TextBlob(POS tag). (v) spaCy. (vi) TreeTagger. (vii) Pattern. (viii) Genism. (ix) Stanford CoreNLP.

Let us see Genism :-

1. Genism is designed to handle large text collections using data streaming.
2. Its lemmatization facilities are based on the pattern package in python.
3. genism.utils.lemmatization() function can be used for performing Lemmatization.
4. We can use this lemmatizer from pattern to extract UTF8

encoded tokens in their base form i.e lemma.

5. It only considers noun, verbs, adjectives & adverbs.

6. For example :- are/is/being $\xrightarrow{\text{Lemmatization}}$ be.

saw $\xrightarrow{\text{Lemmatization}}$ see.

Q3. What is the need of text cleaning? How it is done?

Ans 1. Text cleaning refers to the process of removing or transforming certain parts of the text so that the text becomes more easily understandable for NLP models that are learning the text, helping NLP models perform better by reducing noise in text data.

2. Common methods to do text cleaning :-

i) Lowercasing the data : Convert input text to same casing format so that it converts 'DATA' → 'data'.

ii) Removing Punctuations :- Punctuation removal process helps us treat each text equally. Example "data!" → 'data'.

iii) Removing Numbers :- Sometimes numbers may not hold vital information in text depending upon use cases.

iv) Removing Extra Space :- Removing extra space is good as extra memory to store whitespace is not wasted.

v) Replacing repetitions of punctuations :- We can use regular expressions to remove repeated punctuations. Example :- "data!!!" becomes "data".

vi) Removing emoji's :- Removing emoji's may be useful as it may not hold any valuable information. Example :- "The moon is beautiful 😍" → "The moon is beautiful".

vii) Removing emoticons :- Many text data on twitter & Instagram contains emoticons which may not be indispensable. Example :- "May all beings be happy and free <3" → "May all beings be happy and free".

ctually we

Page No.
Date

NAME: AFNAN ATTAR PRN: F19112003 CLASS: BE COMPT
SUBJECT:- NLP ASSIGNMENT No.:- 04

Q1) What is language modeling? Explain any one language model in detail.

Ans 1. Language modelling is the use of various statistical and probabilistic techniques to determine the probability of a given sequence of words occuring in a sentence.

2. Language models analyze bodies of text data to provide a basis for their word prediction.

3. Some common statistical language modelling types are:
(i) N-gram (ii) Unigram (iii) Bidirectional (iv) Exponential (v) Continuous space

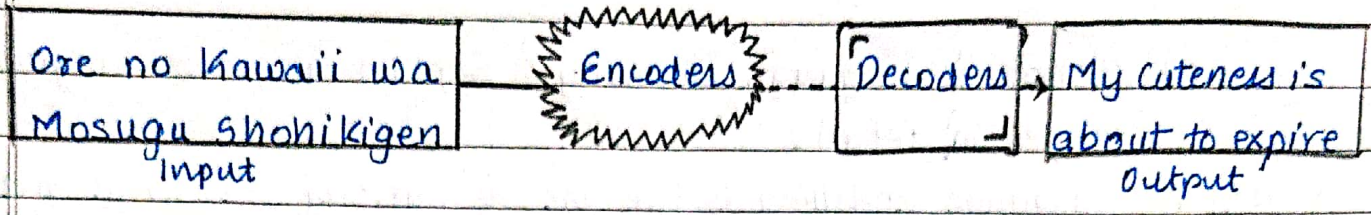
- **N-gram :-**

1. N-gram are relatively simple approach to language models.
2. They create a probability distribution for a sequence of n.
3. N can be any number and determines the size of the "gram".
4. For example if n = 5, a gram can be "never gonna give you up", the model then assigns probabilities using sequences of n size.
5. Some types of n-grams are unigram [n = 1], bigrams [n = 2], trigrams [n = 3] and so on.
6. Basically n can be thought of amount of context the model is told to consider.
7. n = 4, 5 works really well for european language.
8. N-gram approximation: $P(w_1^n) = \prod_{k=1}^{n} P(w_k \mid w_{k-N+1}^{k-1})$

Q2) What is the transformer model in NLP and how it works?

Ans 1. A transformer is a deep learning model that adopts the mechanism of self-attention, differentially weighting the significance of each part of the input data.

2. It is used primarily in NLP and computer vison.
3. It has an encoder-decoder architecture :-



4. In reality, we use stacks of encoders & decoders not just single.
5. On a high level, the encoders map an input sequence into an abstract continuous representation that holds all the learned information of that input.
6. The decoder generates single output while being fed previous output.

Q3) What is topic modelling?

Ans 1. Topic modelling is a machine learning technique that automatically analyses text data to determine cluster words for a set of documents.
2. Topic modelling does not require training and hence is an easy way to analyze our data quickly.
3. Topic modelling could be used to identify the topics of a set of customer reviews by detecting patterns & recurring words.
4. For example, consider a reeview about our software "Supla Shot":- A great thing about supla shot is that "it's free to use" as long you're not "charging" for the event. There is a "fee" if you are "charging" for the event - "2.5% plus a 0.99$ transaction fee".
5. Here words under double quotes are identified as topics and our topic modelling system can group these words with other reviews that talk about similar things.
6. Topic modelling simply involves counting words and grouping similar word patterns to infer topics within unstructured data.

sy operate

Page No.
Date

NAME:- AFNAN ATTAR PRN:- F19112003 CLASS:- BE COMP II
SUBJECT:- NLP ASSIGNMENT No.:- 05

Q1) What is morphological analysis of words?

Ans 1. Morphological analysis is a method for exploring possible solutions. It is the analysis of a word based on the meaningful parts contained within.

2. Some words cannot be broken down into multiple meaningful parts, but many words are composed of more than one meaningful unit, called as morphemes.

3. In linguistics, morphology is the identification, analysis and description of the structure of a given language's morphemes and other linguistic units, such as root words, affixes etc.

4. Some types of morphemes are :-

- Free morphemes : Can appear with other lexemes or may standalone too. Example:- "Free", "Go".
- Bound morphemes : Appear together with other morphemes to form a lexeme. For example:- hapiness [happy + ness]
- Inflectional morphemes :- Modify a word's tense, number aspect etc. without deriving a word in new grammatical category. [Dog → Dogs].
- Zero morphemes :- It is a morpheme that is realized by a phonologically null afix i.e empty strings of phonological segments.
- Derivational morphemes : They modify a word or create a new word which gets its own entry in the dictionary. They change a word's grammatical category unlike inflectional morphemes. For example:- Beauty + ful = Beautiful.

Q2 What is a word in morphology?

Ans The word in morphology is a speech sound or combination of

sounds, or its representation in writing, that symbolizes and communicates a meaning and may consists of a single morpheme or a combination of morphemes.

Q3. How many morphemes are in a word?

Ans 1. Every word must have at least one morpheme, but it may have more than one.

2. For example the word "uncharacteristically" has 4 morphemes.
   i) "Charater" is the base or free morphemes.
   ii) "Characteristic" is the bound morpheme.
   iii) "Uncharacteristically" is another bound morpheme.
   iv) "characteristically" is a bound morpheme.

Q4. What is the difference between a word and a morpheme?

Ans 1. Words are potentially complex units, composed of even more basic units called morphemes.

2. The main difference is that while a word can stand alone, a morpheme may or may not be able to stand alone.

Q5. Are morphemes phonemes?

Ans 1. A phoneme is the basic unit of phonology, it is the smallest unit of sound that may cause change of meaning within a language, but does not have any meaning by itself.

2. Morphemes, the basic unit of morphology are the smallest meaningful unit of language.

3. Thus, a morpheme is a series of phonemes that has a special meaning.