Assignment - 3

1. Define N-gram.

Ans An n-gram is a contiguous sequence of n words. For instance, "edpresso shots" is a 2-gram.

The concept of n-grams is commonly found in NLP and data science.

An n-gram of size 1 is also referred to as a "unigram", size 2 is a "bigram", and size 3 is a "trigram".

1. "Educative" is a unigram (1-gram)
2. "Edpresso shots" is a bigram (2-gram)
3. "Here is another example" is a 4-gram.

"This is a sentence"

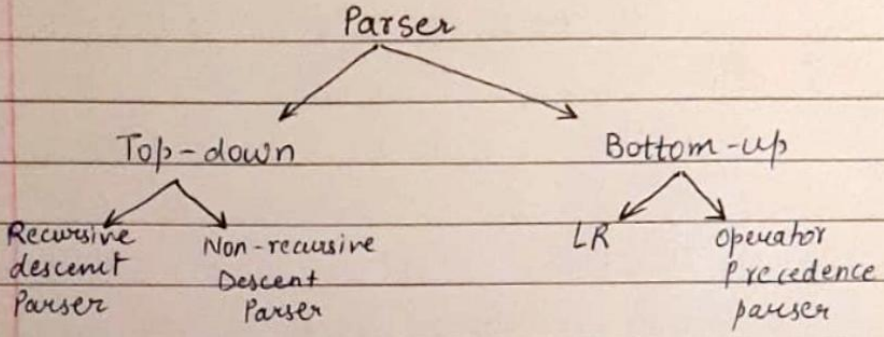
Unigrams	Bigrams	Trigrams
"This"	"This is"	"This is a"
"is"	"is a"	"is a sentence".
"a"	"a sentence"	
"sentence"		

N-grams may be used to create probabilistic language models called n-grams models. N-gram models predict the occurence of a word based on its N-1 previous word.

N-grams have a wide variety of uses. Some applications of n-grams in NLP include auto-completion of sentences, auto spell-check, and semantic analysis

2. Explain different parsing methods in detail.

Ans NLP provides us with two basic parsing techniques viz. Top-down Parsing and Bottom-up parsing. Their name describes the direction in which parsing process advances.



Top-down Parser

The top-down parser is the parser that generates parse for the given input string with the help of grammar productions by expanding the non-terminals i.e. it starts from the start symbol and ends on the terminal.

1. Recursive descent parser is also known as the Brute force parser or the backtracking parser.
2. Non-Recursive descent parser is also known as LL(1) parser or predictive parser or without backtracking parser or dynamic parser.

Bottom-up Parser

Bottom-up parser is the parser that generates

the parse tree for the given input string with the help of grammar productions by compressing the non-terminals i.e. it starts from non-terminals & ends on the start symbol.

1.) LR parser generates the parse tree for the given string by using unambiguous grammar.

LR parser is of four types :-

a) LR (0)

b) SLR(1)

c) LALR (1)

d) CLR (1)

2) Operator precedence parser generates the parse tree from given grammar and string but the only condition is two consecutive non-terminals.

3. What is Treebank.

Ans A syntactically processed corpus that contains annotations of natural language data at various linguistic levels. A treebank provides mainly the morphosyntactic and syntactic structure of the utterances within the corpus and consists of a bank of linguistic trees.

A text-corpus in which each sentence is annotated with syntactic structure. Syntactic structure is commonly represented as a tree structure. Treebanks can be used in corpus linguistics for studying syntactic

phenomena or in computational linguistics for training or testing parsers.

4. Explain the following applications of NLP using case study:

i) Sentiment Analysis

Sentiment analysis is a natural language processing technique used to determine wheter data is positive, negative or neutral. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, & understand customer needs. Sentiment analysis focuses on the polarity of a text (+ve, -ve, neutral) but it also goes beyond polarity to detect specific feelings & emotions, urgency & even intentions.

Most popular types of sentiment analysis:

1) Graded Sentiment Analysis
2) Emotion detection
3) Aspect-based Sentiment Analysis
4) Multilingual Sentiment Analysis

Sentiment Analysis is important, since human expres their thoughts and feelings more openly than ever before, sentiment analysis is fast

becoming an essential tool to monitor & understand sentiment in all types of data. The overall benefits of sentiment analysis include:

- Sorting Data at Scale
- Real Time Analysis
- Consistent Criteria

ii) Question Answering System.

Ans QA is a computer science discipline with in the fields of information retrieval and NLP, which is concerned with building systems that automatically answer questions posed by human in a natural language.

Question answering is very dependent on a good search corpus - for without documents containing the answer, there is little any question answering system can do. It thus makes sense that larger collection sizes generally lend well to better question answering performance, unless the question domain is orthogonal to the collection. The notion of data redundancy in massive collection such as the web, means that nuggets of information are likely to be phrased in many different ways in differing contexts and documents, leading to two benefits.

1. By having the right information appear in many forms, the burden on the question answering

system to perform complex NLP techniques to understand the text is lessened.

2.) Correct answers can be filtered from false positives by replying on the correct answers to appear more times in the documents than instances of incorrect ones.

iii) Text Classification.

Ans Text classification is the processing of labeling or organizing text data into groups. It forms a fundamental part of Natural Language processing. In the digital age that we live in we are surrounded by text on our social media accounts, in commercials, on websites, Ebooks, etc. The majority of this text data is unstructured, so classifying this data can be extremely useful.

Text classification has a wide array of applications. Some popular uses are:

- Spam detection in emails.
- Sentiment analysis of online reviews.
- Topic labeling documents like research papers.
- Language detection like in Google Translate.
- Age / Gender identification of anonymous users.
- Tagging online content.