CONCEPTUAL QUESTIONS
For
TEXT MINING

Website: www.analytixlabs.co.in
Email: info@analytixlabs.co.in

Natural Language Processing [NLP]

Introduction

According to industry estimates, only 21% of the available data is present in structured form. Data is being generated as we speak, as we tweet, as we send messages on Whatsapp and in various other activities. Majority of this data exists in the textual form, which is highly unstructured in nature.
Few notorious examples include – tweets / posts on social media, user to user chat conversations, news, blogs and articles, product or services reviews and patient records in the healthcare sector. A few more recent ones includes chatbots and other voice driven bots.

Despite having high dimension data, the information present in it is not directly accessible unless it is processed (read and understood) manually or analyzed by an automated system.

In order to produce significant and actionable insights from text data, it is important to get acquainted with the techniques and principles of Natural Language Processing (NLP).

So, if you plan to create chatbots this year, or you want to use the power of unstructured text, this guide is the right starting point. This guide unearths the concepts of natural language processing, its techniques and implementation. The aim of the article is to teach the concepts of natural language processing and apply it on real data set.

Humans are social animals and language is our primary tool to communicate with the society. But, what if machines could understand our language and then act accordingly? Natural Language Processing (NLP) is the science of teaching machines how to understand the language we humans speak and write.

Natural Language Processing is one of the principal areas of Artificial Intelligence. NLP plays a critical role in many intelligent applications such as automated chat bots, article summarizers, multi-lingual translation and opinion identification from data. Every industry which exploits NLP to make sense of unstructured text data, not just demands accuracy, but also swiftness in obtaining results.

Natural Language Processing is a capacious field, some of the tasks in nlp are – text classification, entity detection, machine translation, question answering, and concept identification.

1. Introduction to Natural Language Processing

NLP is a branch of data science that consists of systematic processes for analyzing, understanding, and deriving information from the text data in a smart and efficient manner. By utilizing NLP and its components, one can organize the massive chunks of text data, perform numerous automated tasks and solve a wide range of problems such as – automatic summarization, machine translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation etc.
Before moving further, I would like to explain some terms that are used in the article:

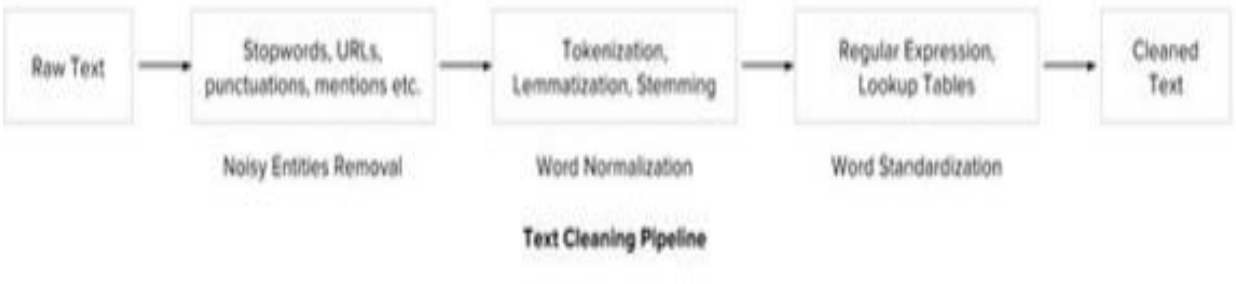
- Tokenization – process of converting a text into tokens
- Tokens – words or entities present in the text
- Text object – a sentence or a phrase or a word or an article

2. Text Preprocessing

Since, text is the most unstructured form of all the available data, various types of noise are present in it and the data is not readily analyzable without any pre-processing. The entire process of cleaning and standardization of text, making it noise-free and ready for analysis is known as text preprocessing.
It is predominantly comprised of three steps:

- Noise Removal
- Lexicon Normalization
- Object Standardization

The following image shows the architecture of text preprocessing pipeline.



2.1 Noise Removal

Any piece of text which is not relevant to the context of the data and the end-output can be specified as the noise.
For example – language stopwords (commonly used words of a language – is, am, the, of, in etc), URLs or links, social media entities (mentions, hashtags), punctuations and industry specific words. This step deals with removal of all types of noisy entities present in the text.
A general approach for noise removal is to prepare a dictionary of noisy entities, and iterate the text object by tokens (or by words), eliminating those tokens which are present in the noise dictionary.

Another approach is to use the regular expressions while dealing with special patterns of noise.

2.2 Lexicon Normalization

Another type of textual noise is about the multiple representations exhibited by single word.
For example – “play”, “player”, “played”, “plays” and “playing” are the different variations of the word – “play”, Though they mean different but contextually all are similar. The step converts all the disparities of a word into their normalized form (also known as lemma). Normalization is a pivotal step for feature engineering with text as it converts the high dimensional features (N different features) to the low dimensional space (1 feature), which is an ideal ask for any ML model.
The most common lexicon normalization practices are:

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

2.3 Object Standardization

Text data often contains words or phrases which are not present in any standard lexical dictionaries. These pieces are not recognized by search engines and models.

Some of the examples are – acronyms, hashtags with attached words, and colloquial slangs. With the help of regular expressions and manually prepared data dictionaries, this type of noise can be fixed

Apart from three steps discussed so far, other types of text preprocessing includes encoding-decoding noise, grammar checker, and spelling correction etc.

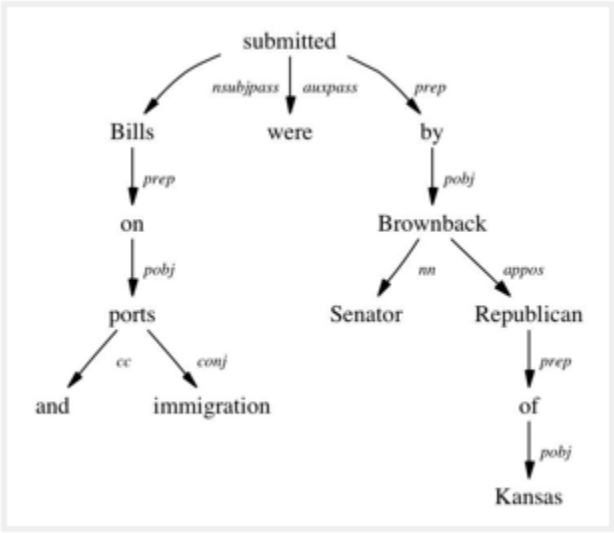
3.Text to Features (Feature Engineering on text data)

To analyse a preprocessed data, it needs to be converted into features. Depending upon the usage, text features can be constructed using assorted techniques – Syntactical Parsing, Entities / N-grams / word-based features, Statistical features, and word embeddings. Read on to understand these techniques in detail.

3.1 Syntactic Parsing

Syntactical parsing involves the analysis of words in the sentence for grammar and their arrangement in a manner that shows the relationships among the words. Dependency Grammar and Part of Speech tags are the important attributes of text syntactics.

Dependency Trees – Sentences are composed of some words sewed together. The relationship among the words in a sentence is determined by the basic dependency grammar. Dependency grammar is a class of syntactic text analysis that deals with (labeled) asymmetrical binary relations between two lexical items (words). Every relation can be represented in the form of a triplet (relation, governor, dependent). For example: consider the sentence – "*Bills on ports and immigration were submitted by Senator Brownback, Republican of Kansas."* The relationship among the words can be observed in the form of a tree representation as shown:



The tree shows that "submitted" is the root word of this sentence, and is linked by two sub-trees (subject and object subtrees). Each subtree is a itself a dependency tree with relations such as – ("Bills" <-> "ports" <by> "proposition" relation), ("ports" <-> "immigration" <by> "conjugation" relation).

This type of tree, when parsed recursively in top-down manner gives grammar relation triplets as output which can be used as features for many nlp problems like entity wise sentiment analysis, actor & entity identification, and text classification.
Part of speech tagging – Apart from the grammar relations, every word in a sentence is also associated with a part of speech (pos) tag (nouns, verbs, adjectives, adverbs etc). The pos tags defines the usage and function of a word in the sentence.

Part of Speech tagging is used for many important purposes in NLP:

A.Word sense disambiguation: Some language words have multiple meanings according to their usage. For example, in the two sentences below:
I. "Please book my flight for Delhi"
II. "I am going to read this book in the flight"
"Book" is used with different context, however the part of speech tag for both of the cases are different. In sentence I, the word "book" is used as ***verb***, while in II it is used as ***noun***. (Lesk Algorithm is also used for similar purposes)

B.Improving word-based features: A learning model could learn different contexts of a word when used word as the features, however if the part of speech tag is linked with them, the context is preserved, thus making strong features. For example:
Sentence -"book my flight, I will read this book"
Tokens – ("book", 2), ("my", 1), ("flight", 1), ("I", 1), ("will", 1), ("read", 1), ("this", 1)
Tokens with POS – ("book_VB", 1), ("my_PRP$", 1), ("flight_NN", 1), ("I_PRP", 1), ("will_MD", 1), ("read_VB", 1), ("this_DT", 1), ("book_NN", 1)

C. Normalization and Lemmatization: POS tags are the basis of lemmatization process for converting a word to its base form (lemma).

D.Efficient stop word removal: POS tags are also useful in efficient removal of stopwords.
For example, there are some tags which always define the low frequency / less important words of a language. For example: (**IN** – "within", "upon", "except"), (**CD** – "one", "two", "hundred"), (**MD** – "may", "must" etc)

3.2 Entity Extraction (Entities as features)

Entities are defined as the most important chunks of a sentence – noun phrases, verb phrases or both. Entity Detection algorithms are generally ensemble models of rule based parsing, dictionary lookups, pos tagging and dependency parsing. The applicability of entity detection can be seen in the automated chat bots, content analyzers and consumer insights.

Topic Modelling & Named Entity Recognition are the two key entity detection methods in NLP.

A. Named Entity Recognition (NER)

The process of detecting the named entities such as person names, location names, company names etc from the text is called as NER. For example:
Sentence – Sergey Brin, the manager of Google Inc. is walking in the streets of New York.
Named Entities – ("person" : "Sergey Brin"), ("org" : "Google Inc."), ("location" : "New York")
A typical NER model consists of three blocks:

Noun phrase identification: This step deals with extracting all the noun phrases from a text using dependency parsing and part of speech tagging.

Phrase classification: This is the classification step in which all the extracted noun phrases are classified into respective categories (locations, names etc). Google Maps API provides a good path to disambiguate locations, Then, the open databases from dbpedia, wikipedia can be used to identify person names or company names. Apart from this, one can curate the lookup tables and dictionaries by combining information from different sources.

Entity disambiguation: Sometimes it is possible that entities are misclassified, hence creating a validation layer on top of the results is useful. Use of knowledge graphs can be exploited for this purposes. The popular knowledge graphs are – Google Knowledge Graph, IBM Watson and Wikipedia.

B. Topic Modeling

Topic modeling is a process of automatically identifying the topics present in a text corpus, it derives the hidden patterns among the words in the corpus in an unsupervised manner. Topics are defined as "a repeating pattern of co-occurring terms in a corpus". A good topic model results in – "health", "doctor", "patient", "hospital" for a topic – Healthcare, and "farm", "crops", "wheat" for a topic – "Farming".
Latent Dirichlet Allocation (LDA) is the most popular topic modelling technique

C. N-Grams as Features

A combination of N words together are called N-Grams. N grams (N > 1) are generally more informative as compared to words (Unigrams) as features. Also, bigrams (N = 2) are considered as the most important features of all the others.

3.3 Statistical Features

Text data can also be quantified directly into numbers using several techniques described in this section:

A. Term Frequency – Inverse Document Frequency (TF – IDF)

TF-IDF is a weighted model commonly used for information retrieval problems. It aims to convert the text documents into vector models on the basis of occurrence of words in the documents without taking considering the exact ordering. For Example – let say there is a dataset of N text documents, In any document "D", TF and IDF will be defined as –
Term Frequency (TF) – TF for a term "t" is defined as the count of a term "t" in a document "D"
Inverse Document Frequency (IDF) – IDF for a term is defined as logarithm of ratio of total documents available in the corpus and number of documents containing the term T.

TF.IDF – TF IDF formula gives the relative importance of a term in a corpus (list of documents), given by the following formula below. Following is the code using python's scikit learn package to convert a text into tfidf vectors:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

The model creates a vocabulary dictionary and assigns an index to each word. Each row in the output contains a tuple (i,j) and a tf-idf value of word at index j in document i.

B. Count / Density / Readability Features

Count or Density based features can also be used in models and analysis. These features might seem trivial but shows a great impact in learning models. Some of the features are: Word Count, Sentence Count, Punctuation Counts and Industry specific word counts. Other types of measures include readability measures such as syllable counts, smog index and flesch reading ease.

3.4 Word Embedding (text vectors)

Word embedding is the modern way of representing words as vectors. The aim of word embedding is to redefine the high dimensional word features into low dimensional feature vectors by preserving the contextual similarity in the corpus. They are widely used in deep learning models such as Convolutional Neural Networks and Recurrent Neural Networks.

Word2Vec and GloVe are the two popular models to create word embedding of a text. These models takes a text corpus as input and produces the word vectors as output.
Word2Vec model is composed of preprocessing module, a shallow neural network model called Continuous Bag of Words and another shallow neural network model called skip-gram. These models are widely used for all other nlp problems. It first constructs a vocabulary from the training corpus and then learns word embedding representations.

They can be used as feature vectors for ML model, used to measure text similarity using cosine similarity techniques, words clustering and text classification techniques.

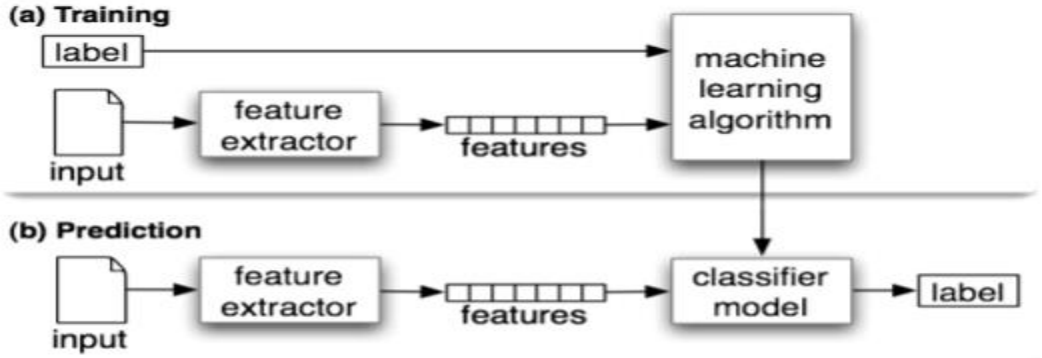
4. Important tasks of NLP

This section talks about different use cases and problems in the field of natural language processing.

4.1 Text Classification

Text classification is one of the classical problems of NLP. Notorious examples include – Email Spam Identification, topic classification of news, sentiment classification and organization of web pages by search engines.
Text classification, in common words is defined as a technique to systematically classify a text object (document or sentence) in one of the fixed category. It is really helpful when the amount of data is too large, especially for organizing, information filtering, and storage purposes.

A typical natural language classifier consists of two parts: (a) Training (b) Prediction as shown in image below. Firstly the text input is processes and features are created. The machine learning models then learn these features and is used for predicting against the new text.



The text classification model are heavily dependent upon the quality and quantity of features, while applying any machine learning model it is always a good practice to include more and more training data. Here are some tips that I wrote about improving the text classification accuracy in one of my previous article.

4.2 Text Matching / Similarity

One of the important areas of NLP is the matching of text objects to find similarities. Important applications of text matching includes automatic spelling correction, data de-duplication and genome analysis etc.

A number of text matching techniques are available depending upon the requirement. This section describes the important techniques in detail.

A. Levenshtein Distance – The Levenshtein distance between two strings is defined as the minimum number of edits needed to transform one string into the other, with the allowable edit operations being insertion, deletion, or substitution of a single character. Following is the implementation for efficient memory computations.

B. Phonetic Matching – A Phonetic matching algorithm takes a keyword as input (person's name, location name etc) and produces a character string that identifies a set of words that are (roughly) phonetically similar. It is very useful for searching large text corpuses, correcting spelling errors and matching relevant names. Soundex and Metaphone are two main phonetic algorithms used for this purpose. Python's module Fuzzy is used to compute soundex strings for different words

C. Flexible String Matching – A complete text matching system includes different algorithms pipelined together to compute variety of text variations. Regular expressions are really helpful for this purposes as well. Another common techniques include – exact string matching, lemmatized matching, and compact matching (takes care of spaces, punctuation's, slangs etc).

D. Cosine Similarity – W hen the text is represented as vector notation, a general cosine similarity can also be applied in order to measure vectorized similarity. Following code converts a text to vectors (using term frequency) and applies cosine similarity to provide closeness among two text.

4.3 Coreference Resolution

Coreference Resolution is a process of finding relational links among the words (or phrases) within the sentences. Consider an example sentence: " Donald went to John's office to see the new table. He looked at it for an hour."

Humans can quickly figure out that "he" denotes Donald (and not John), and that "it" denotes the table (and not John's office). Coreference Resolution is the component of NLP that does this job automatically. It is used in document summarization, question answering, and information extraction.

4.4 Other NLP problems / tasks

- **Text Summarization** – Given a text article or paragraph, summarize it automatically to produce most important and relevant sentences in order.
- **Machine Translation** – Automatically translate text from one human language to another by taking care of grammar, semantics and information about the real world, etc.
- **Natural Language Generation and Understanding** – Convert information from computer databases or semantic intents into readable human language is called language generation. Converting chunks of text into more logical structures that are easier for computer programs to manipulate is called language understanding.
- **Optical Character Recognition** – Given an image representing printed text, determine the corresponding text.
- **Document to Information** – This involves parsing of textual data present in documents (websites, files, pdfs and images) to analyzable and clean format.

5. Important Libraries for NLP (python)

- Scikit-learn: Machine learning in Python
- Natural Language Toolkit (NLTK): The complete toolkit for all NLP techniques.
- Pattern – A web mining module for the tools for NLP and machine learning.
- TextBlob – Easy to use nlp tools API, built on top of NLTK and Pattern.
- spaCy – Industrial strength NLP with Python and Cython.
- Gensim – Topic Modelling for Humans

Feature	Spacy	NLTK	Core NLP
Easy installation	Y	Y	Y
Python API	Y	Y	N
Multi Language support	N	Y	Y
Tokenization	Y	Y	Y
Part-of-speech tagging	Y	Y	Y
Sentence segmentation	Y	Y	Y
Dependency parsing	Y	N	Y
Entity Recognition	Y	Y	Y
Integrated word vectors	Y	N	N
Sentiment analysis	Y	Y	Y

Coreference resolution	N	N	Y	• Stanford Core NLP – NLP services and packages by Stanford NLP Group.

Let's compare Spacy with other famous tools to implement nlp in python – CoreNLP and NLTK.

Feature Availability

Speed: Key Functionalities – Tokenizer, Tagging, Parsing

Package	**Tokenizer**	**Tagging**	**Parsing**
spaCy	0.2ms	1ms	19ms
CoreNLP	2ms	10ms	49ms
NLTK	4ms	443ms	–

Accuracy: Entity Extraction

Package	**Precition**	**Recall**	**F-Score**
spaCy	0.72	0.65	0.69
CoreNLP	0.79	0.73	0.76
NLTK	0.51	0.65	0.58

INTERVIEW QUESTIONS

Q. What is NLP?

Ans: NLP stands for Natural Language Processing. It is a branch of artificial intelligence that focuses on enabling computers to understand, interpret, and generate human language.

Q. What are the key challenges in NLP?

Ans: Some key challenges in NLP include text ambiguity, language understanding, word sense disambiguation, syntactic and semantic parsing, and handling out-of-vocabulary words.

Q. How do you preprocess text data in NLP?

Ans: Text preprocessing includes tasks like tokenization, lowercasing, stop word removal, stemming or lemmatization, and handling special characters or symbols.

Q. What is tokenization?

Ans: Tokenization is the process of breaking text into smaller units called tokens, such as words or subwords, to analyze the text at a granular level.

Q. What are stop words, and why are they removed during preprocessing?

Ans: Stop words are commonly occurring words (e.g., "the," "and," "is") that are filtered out during text preprocessing because they often do not add significant meaning to the text analysis.

Q. Explain TF-IDF (Term Frequency-Inverse Document Frequency)?

Ans: TF-IDF is a numerical representation of the importance of a word in a document relative to a collection of documents. It combines the term frequency (how often a word appears in a document) and the inverse document frequency (how rare the word is across the entire document collection).

Q. What is Word Embedding?

Ans: Word Embedding is a technique used to represent words as dense vectors in a continuous vector space, capturing semantic relationships between words. It helps in numerical representation of words for machine learning algorithms.

Q. Explain the concept of Word2Vec?

Ans: Word2Vec is a popular word embedding model that represents words as dense vectors by predicting the likelihood of words appearing in the context of other words.

Q. What is LSTM (Long Short-Term Memory)?

Ans: LSTM is a type of recurrent neural network designed to handle long-range dependencies in sequential data, making it suitable for processing and understanding textual data.

Q. How does attention mechanism work in NLP models?

Ans: The attention mechanism helps NLP models focus on relevant parts of the input text when generating or understanding the output, improving model performance on tasks like machine translation and summarization.

Q. What is the difference between sentiment analysis and text classification?

Ans: Sentiment analysis aims to determine the sentiment (positive, negative, neutral) expressed in a piece of text, while text classification involves assigning a predefined label or category to a text.

Q. How do you evaluate an NLP model's performance?

Ans: NLP models are evaluated using metrics like accuracy, precision, recall, F1-score, ROC-AUC, or perplexity, depending on the specific task.

Q. What is the BLEU score?

Ans: BLEU (Bilingual Evaluation Understudy) is a metric used to evaluate the quality of machine translation by comparing the generated text to one or more reference translations.

Q. Explain the concept of Transfer Learning in NLP.

Ans: Transfer Learning involves using pre-trained models on large datasets to extract general language knowledge and then fine-tuning these models on specific NLP tasks with smaller datasets, resulting in improved performance and faster training.

Q. What are the main steps in building an NLP pipeline?

Ans: The main steps in building an NLP pipeline include data preprocessing, feature extraction (e.g., word embeddings), model selection and training, and evaluation.

Q. How can you handle out-of-vocabulary words in NLP?

Ans: Out-of-vocabulary words can be handled by replacing them with a special token, using character-level embeddings, or leveraging subword units like byte-pair encoding (BPE) or WordPiece.

Q. What is the difference between rule-based and statistical NLP approaches?

Ans: Rule-based NLP approaches rely on predefined rules or patterns to process and analyze text, while statistical approaches learn patterns and relationships from data using machine learning algorithms.

Q. Explain the concept of Named Entity Recognition (NER).

Ans: Named Entity Recognition is the task of identifying and classifying named entities (such as names, organizations, locations, etc.) in text.

Q. What is the purpose of the POS (Part-of-Speech) tagging?

Ans: POS tagging assigns grammatical tags to each word in a sentence, indicating its part of speech (e.g., noun, verb, adjective) and providing useful information for various NLP tasks.

Q. How can you handle text data with imbalanced classes in classification tasks?

Ans: Handling imbalanced classes can be done by techniques such as undersampling the majority class, oversampling the minority class, or using advanced methods like SMOTE (Synthetic Minority Over-sampling Technique).

Q. What is Naive Bayes algorithm, when we can use this algorithm in NLP?

Ans: 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 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.

Q. Explain Dependency Parsing in NLP?

Ans: 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.

Q. What is text Summarization?

Ans: 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 summarize 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

Q. What is NLTK? How is it different from Spacy?

Ans: 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

Q. What is information extraction?

Ans: 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 include:

- Tagger Module
- Relation Extraction Module
- Fact Extraction Module
- Entity Extraction Module
- Sentiment Analysis Module
- Network Graph Module
- Document Classification & Language Modeling Module

Q. What is Bag of Words?

Ans: 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.

Q. What is Pragmatic Ambiguity in NLP?

Ans: 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.

Q. What is Masked Language Model?

Ans: 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.

Q. What is the difference between NLP and CI(Conversational Interface)?

Ans: The difference between NLP and CI is as follows:

Natural Language Processing (NLP)	Conversational Interface (CI)
NLP attempts to help machines understand and learn how language concepts work.	CI focuses only on providing users with an interface to interact with.
NLP uses AI technology to identify, understand, and interpret the requests of users through language.	CI uses voice, chat, videos, images, and more such conversational aid to create the user interface.

Q. What are the best NLP Tools?

Ans: Some of the best NLP tools from open sources are:

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

Q. What is POS tagging?

Ans: Parts of speech tagging better known as POS tagging refer to the process of identifying specific words in a document and grouping 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 general process used for word mapping is quite ineffective for POS tagging because of the same reason.

Q. What is NES?

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

Q. What is the difference between rule-based and statistical machine translation?

Ans: Rule-based machine translation relies on predefined linguistic rules and dictionaries, while statistical machine translation uses statistical models to learn translation patterns from large parallel corpora.

Q. What is sequence labeling in NLP?

Ans: Sequence labeling is the task of assigning labels to each element in a sequence of inputs, such as part-of-speech tagging, named entity recognition, and chunking.

Q. What is the purpose of word sense disambiguation?

Ans: Word sense disambiguation is the task of determining the correct meaning of a word in context, as many words have multiple meanings. It helps improve the accuracy of NLP applications.

Q. Explain the concept of topic modeling?

Ans: Topic modeling is a technique used to uncover latent topics or themes in a collection of documents. It helps in identifying and organizing the main subjects discussed in the text data.

Q. What are some common applications of NLP?

Ans: Common applications of NLP include sentiment analysis, text classification, machine translation, question-answering systems, chatbots, information extraction, and text summarization.

Q. What is the difference between a generative and discriminative model in NLP?

Ans: Generative models learn the joint probability distribution of the input and output variables, while discriminative models learn the conditional probability of the output given the input.

Q. What is the Perceptron algorithm?

Ans: The Perceptron algorithm is a binary linear classification algorithm used for supervised learning. It iteratively updates the weights based on misclassified samples until convergence.

Q. Explain the concept of word alignment in machine translation?

Ans: Word alignment is the process of aligning words in the source language with their corresponding words in the target language during machine translation, enabling accurate translation.

Q. What is the difference between rule-based and statistical parsing?

Ans: Rule-based parsing relies on grammatical rules to parse sentences, while statistical parsing uses probabilistic models and machine learning algorithms to learn parsing patterns from data.

Q. What is the purpose of the attention mechanism in sequence-to-sequence models?

Ans: The attention mechanism helps sequence-to-sequence models focus on relevant parts of the input sequence when generating the output sequence, improving translation and summarization tasks.

Q. What is the difference between precision and recall in information retrieval?

Ans: Precision measures the proportion of retrieved documents that are relevant, while recall measures the proportion of relevant documents that are retrieved. Both metrics are important for evaluating retrieval systems.

Q. What are the challenges in building chatbot systems?

Ans: Challenges in building chatbot systems include natural language understanding, context modeling, generating coherent and human-like responses, and handling user queries that deviate from the expected flow.

Q. What is the difference between shallow parsing and deep parsing?

Ans: Shallow parsing (also known as chunking) focuses on identifying and grouping syntactic phrases, while deep parsing aims to create a detailed parse tree representing the complete syntactic structure of a sentence.

Q. What is the purpose of the GloVe algorithm?

Ans: The GloVe (Global Vectors for Word Representation) algorithm is used to learn word embeddings by factorizing the co-occurrence matrix of words, capturing both global and local word relationships.

Q. Explain the concept of attention heads in transformer models.

Ans: Attention heads in transformer models allow the model to attend to different parts of the input simultaneously, enabling the capture of different aspects of the context and improving performance.

Q. What is the purpose of dependency parsing in NLP?

Ans: Dependency parsing aims to determine the grammatical relationships (dependencies) between words in a sentence, representing the syntactic structure as a directed graph.

Q. How can you handle data sparsity in language modeling?

Ans: Data sparsity can be addressed by using techniques like smoothing (e.g., add-k smoothing), backoff models, or incorporating n-grams with lower order models.

Q. Why do we need NLP?

Ans: One of the main reasons why NLP is necessary is because it helps computers communicate with humans in natural language. It also scales other language-related tasks. Because of NLP, it is possible for computers to hear speech, interpret this speech, measure it and also determine which parts of the speech are important.

Q. What must a natural language program decide?

Ans: A natural language program must decide what to say and when to say something.

Q. Where can NLP be useful?

Ans: NLP can be useful in communicating with humans in their own language. It helps improve the efficiency of the machine translation and is useful in emotional analysis too. It can be helpful in sentiment analysis using python too. It also helps in structuring highly unstructured data. It can be helpful in creating chatbots, Text Summarization and virtual assistants.

Q. How to prepare for an NLP Interview?

Ans: The best way to prepare for an NLP Interview is to be clear about the basic concepts. Go through blogs that will help you cover all the key aspects and remember the important topics. Learn specifically for the interviews and be confident while answering all the questions.

Q. Which NLP model gives best accuracy?

Ans: Naive Bayes Algorithm has the **highest accuracy** when it comes to NLP models. It gives up to 73% correct predictions.

Q. What are the major tasks of NLP?

Ans: Translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation are few of the major tasks of NLP. Under unstructured data, there can be a lot of untapped information that can help an organization grow.

Q. Why is NLP so hard?

Ans: There are several factors that make the process of Natural Language Processing difficult. There are hundreds of natural languages all over the world, words can be ambiguous in their meaning, each natural language has a different script and syntax, the meaning of words can change depending on the context, and so the process of NLP can be difficult. If you choose to upskill and continue learning, the process will become easier over time.

Q. How many steps of NLP are there?

The five phases of NLP involve lexical (structure) analysis, parsing, semantic analysis, discourse integration, and pragmatic analysis.

Q. What is the difference between bag-of-words and word embeddings?

Ans: Bag-of-words represents text as a collection of word frequencies, ignoring the order of words, while word embeddings capture semantic relationships between words in a continuous vector space.

Q. What is BERT (Bidirectional Encoder Representations from Transformers)?

Ans: BERT is a pre-trained language model based on transformer architecture that learns contextual word representations by considering both left and right context. It has achieved state-of-the-art results on various NLP tasks.

Q. Explain the concept of beam search in sequence generation tasks.

Ans: Beam search is a search algorithm used in sequence generation tasks to find the most likely sequence of output tokens. It maintains a set of the most probable partial sequences and explores multiple candidate paths.

Objective Questions:

Q. 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

Ans: 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.

Q. 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

Ans: 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 versa.

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

Python code to implement CosineSimlarity function would look like this:

```
def cosine_similarity(x,y):

    return np.dot(x,y)/( np.sqrt(np.dot(x,x)) * np.sqrt(np.dot(y,y)) )

q1 = wikipedia.page('Strawberry')

q2 = wikipedia.page('Pineapple')
```

```
q3 = wikipedia.page('Google')

q4 = wikipedia.page('Microsoft')

cv = CountVectorizer()

X = np.array(cv.fit_transform([q1.content, q2.content, q3.content, q4.content]).todense())

print ("Strawberry Pineapple Cosine Distance", cosine_similarity(X[0],X[1]))

print ("Strawberry Google Cosine Distance", cosine_similarity(X[0],X[2]))

print ("Pineapple Google Cosine Distance", cosine_similarity(X[1],X[2]))

print ("Google Microsoft Cosine Distance", cosine_similarity(X[2],X[3]))

print ("Pineapple Microsoft Cosine Distance", cosine_similarity(X[1],X[3]))
```

```
Strawberry Pineapple Cosine Distance 0.8899200413701714

Strawberry Google Cosine Distance 0.7730935582847817

Pineapple Google Cosine Distance 0.789610214147025

Google Microsoft Cosine Distance 0.8110888282851575
```

Usually Document similarity is measured by how close semantically the content (or words) in the document are to each other. When they are close, the similarity index is close to 1, otherwise near 0.

The **Euclidean distance** between two points is the length of the shortest path connecting them. Usually computed using Pythagoras theorem for a triangle.

Q. 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

Ans: e)

All of the above can be used as features of the text corpus.

Q. 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

Ans: d)

Q. 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

Ans: d)

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

a. True
b. False

Ans: a)

Q. Which one of the following is keyword Normalization techniques in NLP

a. Stemming
b. Part of Speech
c. Named entity recognition
d. Lemmatization

Ans: a) and d)

Part of Speech (POS) and Named Entity Recognition (NER) is not keyword Normalization techniques. Named Entity helps 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.

Q. Which of the below are NLP use cases?

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

Ans: d)

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

Q. 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

Ans: (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, the correct choice is Klog(3)/T

Q. 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)

Ans: b)

Q. 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

Ans: c)

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

Q. In NLP, the process of converting a sentence or paragraph into tokens is referred to as Stemming?

a. True
b. False

Ans: b)

The statement describes the process of tokenization and not stemming, hence it is False.

Q. In NLP, Tokens are converted into numbers before giving to any Neural Network

a. True
b. False

Ans: a)

In NLP, all words are converted into a number before feeding to a Neural Network.

Q. Identify the odd one out

a. nltk
b. scikit learn
c. SpaCy
d. BERT

Ans: d)

All the ones mentioned are NLP libraries except BERT, which is a word embedding.

Q. TF-IDF helps you to establish?

a. most frequently occurring word in document
b. the most important word in the document

Ans: 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 is offset by the number of documents that appear in the corpus.

- TF is the frequency of terms divided by the 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.

Suppose that we have term count tables of a corpus consisting of only two documents, as listed here:

Term	**Document 1 Frequency**	**Document 2 Frequency**
This	1	1
is	1	1
a	2	
Sample	1	
another		2
example		3

The calculation of tf–idf for the term “this” is performed as follows:

for "this"

---

tf("this", d1) = 1/5 = 0.2

tf("this", d2) = 1/7 = 0.14

idf("this", D) = log (2/2) =0

hence tf-idf

tfidf("this", d1, D) = 0.2* 0 = 0

tfidf("this", d2, D) = 0.14* 0 = 0

for "example"

---

tf("example", d1) = 0/5 = 0

tf("example", d2) = 3/7 = 0.43

idf("example", D) = log(2/1) = 0.301

tfidf("example", d1, D) = tf("example", d1) * idf("example", D) = 0 * 0.301 = 0

tfidf("example", d2, D) = tf("example", d2) * idf("example", D) = 0.43 * 0.301 = 0.129

In its raw frequency form, TF is just the frequency of the “this” for each document. In each document, the word “this” appears once; but as document 2 has more words, its relative frequency is smaller.

An IDF is constant per corpus, and accounts for the ratio of documents that include the word “this”. In this case, we have a corpus of two documents and all of them include the word “this”.So TF–IDF is zero for the word “this”, which implies that the word is not very informative as it appears in all documents.

The word “example” is more interesting – it occurs three times, but only in the second document. To understand more about NLP, check out these NLP projects.

Q. 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

Ans: d)

Q. 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

Ans: 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.

Q. 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

Ans: 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]]

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.*

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

a. True
b. False

Ans: a)

Word2Vec, GloVe based models build word embedding vectors that are multidimensional.

Q. 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

Ans: 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.

Q. In NLP, Bidirectional context is supported by which of the following embedding

a. Word2Vec
b. BERT
c. GloVe
d. All the above

Ans: 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.

Q. 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

Ans: 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.

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

a. True
b. False

Ans: a)

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

a. True
b. False

Ans: a)

One can use Cosine similarity to establish the distance between two vectors represented through Word Embeddings

Q. 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

Ans: a)

Statement b) is a bias as it buckets Woman into Homemaker, whereas statement a) is not a biased statement.

Q. 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

Ans: b)

Open AI's GPT is able to learn complex patterns 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.

Q. Transformer architecture was first introduced with?

a. GloVe
b. BERT
c. Open AI's GPT
d. ULMFit

Ans: c)

ULMFit has an LSTM based Language modeling architecture. This got replaced into Transformer architecture with Open AI's GPT.

Q. Which of the following architecture can be trained faster and needs less amount of training data

a. LSTM-based Language Modelling
b. Transformer architecture

Ans: b)

Transformer architectures were supported from GPT onwards and were faster to train and needed less amount of data for training too.

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

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

Ans: c)

ELMo word embeddings support the same word with multiple embeddings, this helps in using the same word in a different context and thus captures the context than just the meaning of the word unlike in GloVe and Word2Vec. Nltk is not a word embedding.

Word2Vec	Supported word embeddings (one word one embedding)
GloVe	Supported word embeddings (one word one embedding)
ELMo	Trained on a massive dataset to generate word embeddings. Same word multiple word embeddings based on the context it is in using a bidirectional LSTM arch. Two independent LSTM models one is left to right, and another right to left.
ULMFIt	Introduced Transfer Learning, LSTM based bidirectional arch. But two independent models one is left to right and another right to left.
GPT, GPT-2	Transformer based architecture with attention models. Sentences are trained - Unidirectional Left to right
BERT (Bidirectional Encoder Representations from Transformers)	• Bidirectional context-based model using Transfer learning, Masking words in sentences, and predicting based on previous and next sentence context. • BERT uses WordPiece embedding with 30K token vocabulary and learned positional embeddings with supported sequence length up to 512 tokens
XLNET (Generalized Autoregressive Pretraining for Language Understanding)	• Improves on BERT by addressing the dependency between masked words. It doesn't mask any word in the sentence, unlike BERT. • Enables Learning bidirectional context by maximizing the expected likelihood over all the permutations

Q. 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

Ans: c)
BERT uses token, segment and position embedding.

Q. ____ 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

Ans: 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.

Q. ____ Uses unidirectional language model for producing word embedding.

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

Ans: b)

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

Q. 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

Ans: 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.

Q. 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

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

a. True
b. False

Ans: 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.

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

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

Ans: b) XLNET

XLNET has given best accuracy amongst all the models. It has outperformed BERT on 20 tasks and achieves state of the art results on 18 tasks including sentiment analysis, question answering, natural language inference, etc.

Q. Permutation Language models is a feature of ____

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

Ans: 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.

Q. Transformer XL uses relative positional embedding.

a. True
b. False

Ans: 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.

Q. Which of the following techniques can be used for the purpose of keyword normalization, the process of converting a keyword into its base form?
Ans:

1. Lemmatization
2. Levenshtein
3. Stemming
4. Soundex

A) 1 and 2
B) 2 and 4
C) 1 and 3
D) 1, 2 and 3
E) 2, 3 and 4
F) 1, 2, 3 and 4

Solution: (C)

Q. N-grams are defined as the combination of N keywords together. How many bi-grams can be generated from given sentence?

Ans:

"Analytics Labs is a great source to learn data science"

A) 7
B) 8
C) 9
D) 10
E) 11

Solution: (C)

Q. How many trigrams phrases can be generated from the following sentence, after performing following text cleaning steps?

Ans:

- Stopword Removal
- Replacing punctuations by a single space

"#Analytics-Labs is a great source to learn @data_science."

A) 3
B) 4
C) 5
D) 6
E) 7

Solution: (C)

Q. Which of the following regular expression can be used to identify date(s) present in the text object:
"The next meetup on data science will be held on 2017-09-21, previously it happened on 31/03, 2016"

Ans:

A) \d{4}-\d{2}-\d{2}
B) (19|20)\d{2}-(0[1-9]|1[0-2])-[0-2][1-9] C) (19|20)\d{2}-(0[1-9]|1[0-2])-([0-2][1-9]|3[0-1])
D) None of the above

Solution: (D)

Question Context 5-6:
You have collected a data of about 10,000 rows of tweet text and no other information. You want to create a tweet classification model that categorizes each of the tweets in three buckets – positive, negative and neutral.

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

Ans:

A) Naive Bayes
B) SVM
C) None of the above

Solution: (C)

Q. 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?
Ans:

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)

Q. Which of the following features can be used for accuracy improvement of a classification model?
Ans:

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.

Q. What percentage of the total statements are correct with regards to Topic Modeling?
Ans:

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.

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

Ans:

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

Q. In a corpus of N documents, one document is randomly picked. The document contains a total of T terms and the term "data" appears K times.

What is the correct value for the product of TF (term frequency) and IDF (inverse-document-frequency), if the term "data" appears in approximately one-third of the total documents?

A) KT * Log(3)
B) K * Log(3) / T
C) T * Log(3) / K
D) Log(3) / KT

Solution: (B)

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

Question Context 12 to 14:

Refer the following document term matrix

Term	Document						
	d1	*d2*	*d3*	*d4*	*d5*	*d6*	*d7*
t1	2	1	0	0	0	0	0
t2	1	2	0	0	0	0	1
t3	3	1	0	0	1	1	0
t4	0	0	1	2	1	1	1
t5	0	0	1	1	1	1	1
t6	0	0	1	1	0	0	0

Q. Which of the following documents contains the same number of terms and the number of terms in the one of the document is not equal to least number of terms in any document in the entire corpus.

Ans:

A) d1 and d4
B) d6 and d7
C) d2 and d4
D) d5 and d6

Solution: (C)

Both of the documents d2 and d4 contains 4 terms and does not contain the least number of terms which is 3.

Q. Which are the most common and the rarest terms of the corpus?

Ans:

A) t4, t6
B) t3, t5
C) t5, t1
D) t5, t6

Solution: (A)

T5 is most common terms across 5 out of 7 documents, T6 is rare term only appears in d3 and d4

Q. What is the term frequency of a term which is used a maximum number of times in that document?

A) t6 – 2/5
B) t3 – 3/6
C) t4 – 2/6
D) t1 – 2/6

Solution: (B)

t3 is used max times in entire corpus = 3, tf for t3 is 3/6

Q. Which of the following techniques is not a part of flexible text matching?

Ans:

A) Soundex
B) Metaphone
C) Edit Distance
D) Keyword Hashing

Solution: (D)

Except Keyword Hashing all other are the techniques used in flexible string matching

Q. True or False: Word2Vec model is a machine learning model used to create vector notations of text objects. Word2vec contains multiple deep neural networks

Ans:

A) TRUE
B) FALSE

Solution: (B)

Word2vec also contains preprocessing model which is not a deep neural network

Q. Which of the following statement is(are) true for Word2Vec model?

Ans:

A) The architecture of word2vec consists of only two layers – continuous bag of words and skip-gram model

B) Continuous bag of word is a shallow neural network model

C) Skip-gram is a deep neural network model

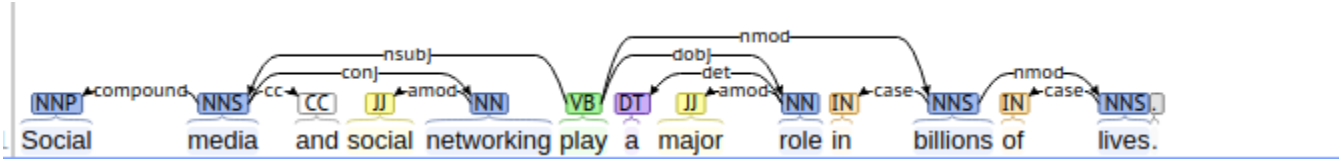
D) Both CBOW and Skip-gram are deep neural network models

E) All of the above

Solution: (C)

Word2vec contains the Continuous bag of words and skip-gram models, which are deep neural nets.

Q. With respect to this context-free dependency graphs, how many sub-trees exist in the sentence?



A) 3

B) 4

C) 5

D) 6

Solution: (D)

Subtrees in the dependency graph can be viewed as nodes having an outward link, for example: Media, networking, play, role, billions, and lives are the roots of subtrees

Q. What is the right order for a text classification model components?

Ans:

1. Text cleaning
2. Text annotation
3. Gradient descent
4. Model tuning
5. Text to predictors

A) 12345

B) 13425

C) 12534

D) 13452

Solution: (C)

A right text classification model contains – cleaning of text to remove noise, annotation to create more features, converting text-based features into predictors, learning a model using gradient descent and finally tuning a model.

Q. Polysemy is defined as the coexistence of multiple meanings for a word or phrase in a text object. Which of the following models is likely the best choice to correct this problem?

Ans:

A) Random Forest Classifier
B) Convolutional Neural Networks
C) Gradient Boosting
D) All of these

Solution: (B)

CNNs are popular choice for text classification problems because they take into consideration left and right contexts of the words as features which can solve the problem of polysemy

Q. Which of the following models can be used for the purpose of document similarity?

Ans:

A) Training a word 2 vector model on the corpus that learns context present in the document
B) Training a bag of words model that learns occurrence of words in the document
C) Creating a document-term matrix and using cosine similarity for each document
D) All of the above

Solution: (D)

word2vec model can be used for measuring document similarity based on context. Bag of words and document term matrix can be used for measuring similarity based on terms.

Q. What are the possible features of a text corpus?

Ans:

1. Count of word in a document
2. Boolean feature – presence of word in a document
3. Vector notation of word
4. Part of Speech Tag
5. Basic Dependency Grammar
6. Entire document as a feature

A) 1
B) 12
C) 123
D) 1234
E) 12345
F) 123456

Solution: (E)

Except for entire document as the feature, rest all can be used as features of text classification learning model.

Q. While creating a machine learning model on text data, you created a document term matrix of the input data of 100K documents. Which of the following remedies can be used to reduce the dimensions of data –

1. Latent Dirichlet Allocation
2. Latent Semantic Indexing
3. Keyword Normalization

A) only 1
B) 2, 3
C) 1, 3
D) 1, 2, 3
Solution: (D)
All of the techniques can be used to reduce the dimensions of the data.

Q. Google Search's feature – "Did you mean", is a mixture of different techniques. Which of the following techniques are likely to be ingredients?
Ans:

1. Collaborative Filtering model to detect similar user behaviors (queries)
2. Model that checks for Levenshtein distance among the dictionary terms
3. Translation of sentences into multiple languages

A) 1
B) 2
C) 1, 2
D) 1, 2, 3

Solution: (C)
Collaborative filtering can be used to check what are the patterns used by people, Levenshtein is used to measure the distance among dictionary terms.

Q. While working with text data obtained from news sentences, which are structured in nature, which of the grammar-based text parsing techniques can be used for noun phrase detection, verb phrase detection, subject detection and object detection.
Ans:
A) Part of speech tagging
B) Dependency Parsing and Constituency Parsing
C) Skip Gram and N-Gram extraction
D) Continuous Bag of Words
Solution: (B)
Dependency and constituent parsing extract these relations from the text

Q. Social Media platforms are the most intuitive form of text data. You are given a corpus of complete social media data of tweets. How can you create a model that suggests the hashtags?
Ans:

A) Perform Topic Models to obtain most significant words of the corpus
B) Train a Bag of Ngrams model to capture top n-grams – words and their combinations

C) Train a word2vector model to learn repeating contexts in the sentences
D) All of these

Solution: (D)
All of the techniques can be used to extract most significant terms of a corpus.

Q. While working with context extraction from a text data, you encountered two different sentences: The tank is full of soldiers. The tank is full of nitrogen. Which of the following measures can be used to remove the problem of word sense disambiguation in the sentences?
Ans:

A) Compare the dictionary definition of an ambiguous word with the terms contained in its neighborhood
B) Co-reference resolution in which one resolute the meaning of ambiguous word with the proper noun present in the previous sentence
C) Use dependency parsing of sentence to understand the meanings

Solution: (A)
Option 1 is called Lesk algorithm, used for word sense disambiguation, rest others cannot be used.

Q. Collaborative Filtering and Content Based Models are the two popular recommendation engines, what role does NLP play in building such algorithms.
A) Feature Extraction from text
B) Measuring Feature Similarity
C) Engineering Features for vector space learning model
D) All of these

Solution: (D)
NLP can be used anywhere where text data is involved – feature extraction, measuring feature similarity, create vector features of the text.

Q. Retrieval based models and Generative models are the two popular techniques used for building chatbots. Which of the following is an example of retrieval model and generative model respectively.
Ans:

A) Dictionary based learning and Word 2 vector model
B) Rule-based learning and Sequence to Sequence model
C) Word 2 vector and Sentence to Vector model
D) Recurrent neural network and convolutional neural network

Solution: (B)
choice 2 best explains examples of retrieval based models and generative models

Q. What is the major difference between CRF (Conditional Random Field) and HMM (Hidden Markov Model)?
Ans:
A) CRF is Generative whereas HMM is Discriminative model
B) CRF is Discriminative whereas HMM is Generative model

C) Both CRF and HMM are Generative model
D) Both CRF and HMM are Discriminative model
Solution: (B)
Option B is correct

Q. Natural Language Processing (NLP) is field of

a) Computer Science
b) Artificial Intelligence
c) Linguistics
d) All of the mentioned
Solution: d

Q. NLP is concerned with the interactions between computers and human (natural) languages.
a) True
b) False

Solution: A)
Explanation: NLP has its focus on understanding the human spoken/written language and convert that interpretation into machine understandable language.

Q. One of the main challenge/s of NLP Is _____

a) Handling Ambiguity of Sentences
b) Handling Tokenization
c) Handling POS-Tagging
d) All of the mentioned
Solution: a
Explanation: There are enormous ambiguity exists when processing natural language.

Q. Modern NLP algorithms are based on machine learning, especially statistical machine learning.

a) True
b) False
Solution: a

Q. Choose form the following areas where NLP can be useful.
a) Automatic Text Summarization
b) Automatic Question-Answering Systems
c) Information Retrieval
d) All of the mentioned
Solution: d

Q. The major tasks of NLP include
a) Automatic Summarization
b) Discourse Analysis
c) Machine Translation
d) All of the mentioned

Solution: d
Explanation: There is even bigger list of tasks of NLP.

Q. Coreference Resolution is
a) Anaphora Resolution
b) Given a sentence or larger chunk of text, determine which words ("mentions") refer to the same objects ("entities")
c) All of the mentioned
d) None of the mentioned
Solution: b
Explanation: Anaphora resolution is a specific type of coreference resolution.

Q. Machine Translation
a) Converts one human language to another
b) Converts human language to machine language
c) Converts any human language to English
d) Converts Machine language to human language
Solution: a
Explanation: The best known example of machine translation is Google translator.

Q. The more general task of coreference resolution also includes identifying so-called "bridging relationships" involving referring expressions.
a) True
b) False
Solution: a
Explanation: Refer the definition of Coreference Resolution.

Q. Morphological Segmentation
a) Does Discourse Analysis
b) Separate words into individual morphemes and identify the class of the morphemes
c) Is an extension of propositional logic
d) None of the mentioned

Solution: b

Q. Given a stream of text, Named Entity Recognition determines which pronoun maps to which noun.
a) False
b) True
Solution: a
Explanation: Given a stream of text, Named Entity Recognition determines which items in the text maps to proper names.

Q. Natural Language generation is the main task of Natural language processing.
a) True
b) False
View Answer

Solution: a
Explanation: Natural Language Generation is to Convert information from computer databases into readable human language.

Q. OCR (Optical Character Recognition) uses NLP.
a) True
b) False
View Answer
Solution: a
Explanation: Given an image representing printed text, determines the corresponding text.

Q. Parts-of-Speech tagging determines
a) part-of-speech for each word dynamically as per meaning of the sentence
b) part-of-speech for each word dynamically as per sentence structure
c) all part-of-speech for a specific word given as input
d) all of the mentioned

Solution: d
Explanation: A Bayesian network provides a complete description of the domain.

Q. Parsing determines Parse Trees (Grammatical Analysis) for a given sentence.
a) True
b) False

Solution: a
Explanation: Determine the parse tree (grammatical analysis) of a given sentence. The grammar for natural languages is ambiguous and typical sentences have multiple possible analyses. In fact, perhaps surprisingly, for a typical sentence there may be thousands of potential parses (most of which will seem completely nonsensical to a human).

Q. IR (information Retrieval) and IE (Information Extraction) are the two same things.
a) True
b) False

Solution: b
Explanation: Information retrieval (IR)
This is concerned with storing, searching and retrieving information. It is a separate field within computer science (closer to databases), but IR relies on some NLP methods (for example, stemming). Some current research and applications seek to bridge the gap between IR and NLP.
Information extraction (IE)
This is concerned in general with the extraction of semantic information from text. This covers tasks such as named entity recognition, Coreference resolution, relationship extraction, etc.

Q. Many words have more than one meaning; we have to select the meaning which makes the most sense in context. This can be resolved by
a) Fuzzy Logic
b) Word Sense Disambiguation
c) Shallow Semantic Analysis

d) All of the mentioned

Solution: b
Explanation: Shallow Semantic Analysis doesn't cover word sense disambiguation.

Q. Given a sound clip of a person or people speaking, determine the textual representation of the speech.
a) Text-to-speech
b) Speech-to-text
c) All of the mentioned
d) None of the mentioned

Solution: b
Explanation: NLP is required to linguistic analysis.

Q. Speech Segmentation is a subtask of Speech Recognition.
a) True
b) False

Solution: a
Explanation: None.

Q. In linguistic morphology, _____ is the process for reducing inflected words to their root form.
a) Rooting
b) Stemming
c) Text-Proofing
d) Both Rooting & Stemming

Solution: b
Explanation: None.