Graphic Era
HILL UNIVERSITY
Established by an Act of the State Legislature of Uttarakhand (Adhiniyam Sankhya 12 of 2011)
University under section 2(f) of UGC Act, 1956

MOOCS REPORT

ON

BASIC

INTRODUCTION TO NLP

SUBMITTED TO :

MR.AMIT JUYAL

(ASSISTANT PROFESSOR)

SEC B

SUBMITTED BY

Vivek kumar

MCA2

2201458 (68)

ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to my teacher "**Mr.Amit Juyal**" who gave me the golden opportunity to do this MOOCs on "**Basic Introduction to Natural Language Processing**". And I would also like to share my gratitude to all my teachers who have done such hard work for us. They supported usin any conditions and always guided and gave strength to achieve the goals.

I would also like to acknowledge my classmates and friends whoalso helped me.

I am indebted to my family members for their love, support and encouragement.

Lastly, I would like to give special thanks and greetings to my fellow mates for giving me some required information, valuable advices and suggestions to complete the report in a comprehensive manner.

Vivek Kumar

MCA2 (68 **)**

CERTIFICATE

About course

The mentor of this course is Prof.Sampriti Chatterjee.
The duration of the course is of 1weeks and this course is offered by Great Learning.

NLP course starts by introducing you to NLP and Python.This course help us to learn about data pre-processing and learn to work with different types of documents using Python. We will understand tokenization, its needs, and its implementation through this NLP course. It go through a hands-on session online and word tokenization implementation using Python programming. Understand stemming, lemmatization, and stopwords better with the hands-on sessions on their implementation using Python.

The mission of Great Learning is to educate its students and cultivate their capacity for life-long learning, to foster independent and original research, and to bring the benefits of discovery to the world.

INDEX

INTRODUCTION:

What is Natural language processing?

NLP stands for natural language processing which is basically used to understand and interpret human language to the machine. In short it is the automatic way to manipulate the natural language, like speech and text, by software for further analysis to get the required information from them

Examples of Natural language processing:

- Predictive text
- Email filters
- Data analysis
- Language translation
- Smart assistants

APPLICATION OF NLP:

Natural Language Processing a subset technique of Artificial Intelligence which is used to narrow the communication gap between the Computer and Human.

APPLICATION:

- SPEECH RECOGNIZATION
- VOICE RECOGNIZATION
- SENTIMENTAL ANALYSIS
- SPAM DETECTION
- LANGUAGE TRANSLATION
- SMART ASSISTANTS
- CHATBOTS
- EMAIL FILTERING
- PREDICTIVE TEXT
- AUTOMATIC SUMMARIZATION
- SOCIAL MEDIA MONITORING

SPEECH RECOGNIZATION:

Speech recognition, also known as automatic speech recognition (ASR), computer speech recognition, or speech-to-text, is a capability which enables a program to process human speech into a written format. While it's commonly confused with voice recognition, speech recognition focuses on the translation of speech from a verbal format to a text one whereas voice recognition just seeks to identify an individual user's voice.

VOICE RECOGNIZATION:

NLP and Voice Recognition are complementary but different. Voice Recognition focuses on processing voice data to convert it into a structured form such as text. NLP focuses on understanding the meaning by processing text input. Voice Recognition can work without NLP , but NLP cannot directly process audio inputs. voice recognition just seeks to identify an individual user's voice.

SENTIMENTAL ANALYSIS:

The daily conversations, the posted content and comments, book, restaurant, and product reviews, hence almost all the conversations and texts are full of emotions. Understanding these emotions is as important as understanding the word-to-word meaning. We as humans can interpret emotional sentiments in writings and conversations, but with the help of

natural language processing, computer systems can also understand the sentiments of a text along with its literal meaning.

SPAM DETECTION:

Our purpose is to detect spam by using various algorithms and measuring their accuracy to find the best fitting algorithm. Using Natural Language Processing for Spam Detection in Email is one of the spam detection through NLP.

LANGUAGE TRANSLATION:

There are as many languages in this world as there are cultures, but not everyone understands all these languages. As our world is now a global village owing to the dawn of technology, we need to communicate with other people who speak a language that might be foreign to us. Natural Language processing helps us by translating the language with all its sentiments.

SMART ASSISTANTS:

In today's world, every new day brings in a new smart device, making this world smarter and smarter by the day. And this advancement is not just limited to machines. We have advanced enough technology to have smart assistants, such as Siri, Alexa, and Cortana. We can talk to them like we talk to normal human beings, and they even respond to us in the same way.

All of this is possible because of Natural Language Processing. It helps the computer system understand our language by

breaking it into parts of speech, root stem, and other linguistic features. It not only helps them understand the language but also in processing its meaning and sentiments and answering back in the same way humans do.

DOCUMENT ANALYSIS:

Another one of NLP's applications is document analysis. Companies, colleges, schools, and other such places are always filled to the brim with data, which needs to be sorted out properly, maintained, and searched for. All this could be done using NLP. It not only searches a keyword but also categorizes it according to the instructions and saves us from the long and hectic work of searching for a single person's information from a pile of files. It is not only limited to this but also helps its user to inform decision-making on claims and risk management.

PREDICTIVE TEXT:

A similar application to online searches is predictive text. It is something we use whenever we type anything on our smartphones. Whenever we type a few letters on the screen, the keyboard gives us suggestions about what that word might be and when we have written a few words, it starts suggesting what the next word could be. These predictive texts might be a little off in the beginning.

Still, as time passes, it gets trained according to our texts and starts to suggest the next word correctly even when we have not written a single letter of the next word. All this is done using

NLP by making our smartphones intelligent enough to suggest words and learn from our texting habits.

AUTOMATIC SUMMARIZATION:

With the increasing inventions and innovations, data has also increased. This increase in data has also expanded the scope of data processing. Still, manual data processing is time taking and is prone to error. NLP has a solution for that, too, it can not only summarize the meaning of information, but it can also understand the emotional meaning hidden in the information. Thus, making the summarization process quick and impeccable.

SENTIMENT ANALYSIS:

The daily conversations, the posted content and comments, book, restaurant, and product reviews, hence almost all the conversations and texts are full of emotions. Understanding these emotions is as important as understanding the word-to-word meaning. We as humans can interpret emotional sentiments in writings and conversations, but with the help of natural language processing, computer systems can also understand the sentiments of a text along with its literal meaning.

CHATBOTS:

With the increase in technology, everything has been digitalized, from studying to shopping, booking tickets, and customer service. Instead of waiting a long time to get some short and instant answers, the chatbot replies instantly and accurately. NLP gives these chatbots conversational capabilities, which help them respond appropriately to the customer's needs instead of just bare-bones replies.

Chatbots also help in places where human power is less or is not available round the clock. Chatbots operating on NLP also have emotional intelligence, which helps them understand the customer's emotional sentiments and respond to them effectively.

SOCIAL MEDIA MONITORING:

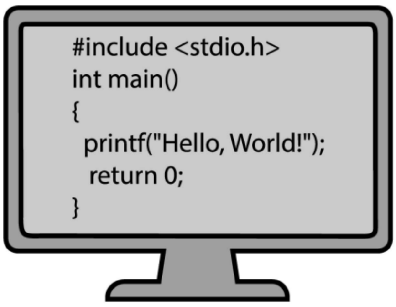
Nowadays, every other person has a social media account where they share their thoughts, likes, dislikes, experiences, etc., which tells a lot about the individuals. We do not only find information about individuals but also about the products and services. The relevant companies can process this data to get information about their products and services to improve or amend them. NLP comes into play here. It enables the computer system to understand unstructured social media data, analyze it and produce the required results in a valuable form for companies.

BUT, How to implement these techniques?

For this python programming language is used.

What is Python?

Python is a popular high level, object oriented and interpreted language



- High level
- Object oriented
- Interpreted

Python is commonly used for developing websites and software, task automation, data analysis, and data visualization. Since it's relatively easy to learn, Python has been adopted by many non-programmers such as accountants and scientists, for a variety of everyday tasks, like organizing finances.

Python is a general-purpose language, meaning it can be used to create a variety of different programs and isn’t specialized for any specific problems. This versatility, along with its beginner-friendliness, has made it one of the most-used programming languages today.

Here are some of the reasons why Python is one of the best choices for natural language processing projects:

- Python's transparent semantics and syntax make it an excellent choice for projects.
- Python developers can enjoy solid support for integration with other languages and tools to build machine learning models.
- Python offers a versatile collection of NLP tools and libraries that enable developers to handle different NLP tasks, including sentiment analysis, POS tagging, document classification, topic modeling, word vectors, and more.

FEATURES OF PYTHON:

- **Easy To Learn and Readable Language**

  Python is extremely easy to learn. Its syntax is super simple and the learning curve of Python is very smooth. It is extremely easy to learn and code in Python and the indentation used instead of curly braces in Python makes it very easy to read Python code.

- **Interpreted Language**

Python is an interpreted language (an interpreted language is a programming language that is generally interpreted, without compiling a program into machine instructions. It is one where the instructions are not directly executed by

the target machine, but instead, read and executed by some other program known as the interpreter)

- **Dynamically Typed Language**

  Python is a dynamically typed language. In other words, in Python, we do not need to declare the data types of the variables which we define. It is the job of the Python interpreter to determine the data types of the variables at runtime based on the types of the parts of the expression. Though it makes coding easier for programmers, this property might create runtime errors.

- **Open Source And Free**

  Python is an open-source programming language and one can download it for free from Python's official website. The community of Python users is constantly contributing to the code of Python in order to improve it.

- **High-Level Language:**

  A high-level language (HLL) is a programming language that enables a programmer to write programs that are more or less independent of a particular type of

computer. These languages are said to be high-level since they are very close to human languages and far away from machine languages. Unlike C, Python is a high-level language.

- **Object Oriented Programming Language**

  Python supports various programming paradigms like structured programming, functional programming, and object-oriented programming. However, the most important fact is that the Object-Oriented approach of Python allows its users to implement the concepts of Encapsulation, Inheritance, Polymorphism, etc.

- **Large Community Support**

  With one of the biggest communities on StackOverflow and Meetup, Python has gained popularity over the years. If we need any kind of help related to Python, the huge community is always there to answer our queries. A lot of questions about Python have already been answered on these sites and Python users can reference them as per requirement.

- **Platform Independent**

Platform independence is yet another amazing feature of Python. In other words, it means that if we write a program in Python, it can run on a variety of platforms, for instance, Windows, Mac, Linux, etc. We do not have to write separate Python code for different platforms.

BENEFITS OF PYTHON:

1. Web development using python
2. Python is simple and beginner friendly language
3. Length of the program is short.
4. Mathematical computation can be done is easily

Important libraries in Python:

- NUMPY (To Solve Numerical Problems)
- PANDAS (Data Preprocessing)
- MATPLOTLIB (Data Visualization)
- SEABORN(Data Visualization)
- NLTK(Used For NLP Techniques)
- SPACY(Used For NLP Techniques)

Roadmap to learn natural language processing:

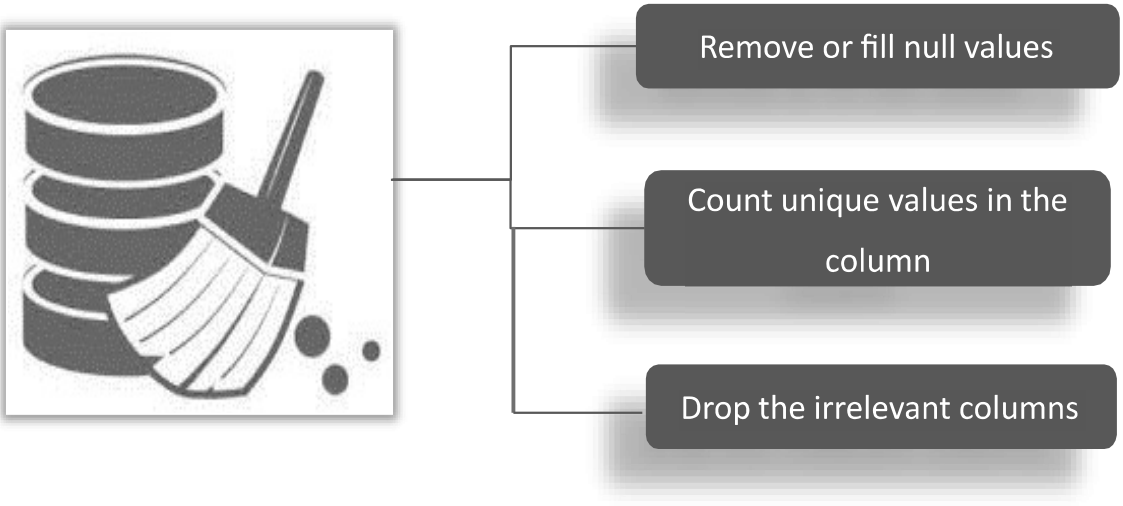
Natural Language Processing (NLP) is the area of research in Artificial Intelligence focused on processing and using Text and Speech data to create smart machines and create insights. One of nowadays most interesting NLP application is creating machines able to discuss with humans about complex topics.

REQUIREMENT:

- PREPROCESSING TECHNIQUES
- MODELLING TECHNIQUES

DATA PREPROCESSING:

Data preprocessing is a way to develop informative data from raw data by removing noise and unwanted attributes



- Remove or fill the null values in the data to get appropriate informative data.
- Take the count of unique data to understand and evaluate the dataset.
- Remove all the unnecessary columns

PREPROCESSING TECHNIQUES:

Some of the most common techniques which are applied in order to prepare text data for inference are:

- **Tokenization:** is used to segment the input text into its constituents words (tokens). In this way, it becomes easier to then convert our data into a numerical format.
- **Stop Words Removal:** is applied in order to remove from our text all the prepositions (eg. “an”, “the”, etc…) which can just be considered as a source of noise in our data (since they do not carry additional informative information in our data).
- **Stemming:** is finally used in order to get rid of all the affixes in our data (eg. prefixes or suffixes). In this way, it can in fact become much easier for our algorithm to

not consider as distinguished words which have actually similar meaning (eg. insight ~ insightful).

TOKENIZATION:

Splitting the sentence into words.

Tokenization method is used to split a phrase, sentence, paragraph, or an entire text document into smaller units. By doing that we can get the individual words or terms. Each of these smaller units are called tokensThe first step of the NLP process is gathering the data (a sentence) and breaking it into understandable parts (words)

NEED OF TOKENIZATION:

- It helps to interpret the meaning of the text by analyzing the wordspresent in the text
- Count the number of words in the text

sentence = "Books are on the table"

Output: ['Books', 'are', 'on', 'the', 'table']

Tokenization is **breaking the raw text into small chunks**. Tokenization breaks the raw text into words, sentences called tokens
These tokens help in understanding the context or developing the model for the NLP. The tokenization helps in interpreting

the meaning of the text by analyzing the sequence of the words. Tokenization **replaces a sensitive data element, for example, a bank account number, with a non-sensitive substitute, known as a token.** The token is a randomized data string that has no essential or exploitable value or meaning.

LEMMATIZATION:

Lemmatization is **the process of grouping together different inflected forms of the same word**. It's used in computational linguistics, natural language processing (NLP) and chatbots. Lemmatization links similar meaning words as one word, making tools such as chatbots and search engine queries more effective and accurate.

The goal of lemmatization is to reduce a word to its root form, also called a *lemma*.

Lemmatization helps to do the morphological analysis of the words. It is important to have the knowledge about the detailed dictionaries which the algorithm can refer to link the formback to its lemma.

Lemmatization takes a word and breaks it down to its lemma. For example, the verb "walk" might appear as "walking," "walks" or "walked." Inflectional endings such as "s," "ed" and "ing" are removed. Lemmatization groups these words as its lemma, "walk."

Form	Morphological information	Lemma

Helps	Third person singular number, present tense help	Help
Helping	Ing form of the verb	Help

Applications of lemmatization

Lemmatization is commonly applied in the following areas:

- Artificial intelligence (AI).
- Big data analytics.
- Chatbots.
- Search queries.
- Sentiment analysis.

Lemmatization can be applied in a number of different circumstances. For example, in search queries, lemmatization lets end users query any version of a base word and get relevant results. Because search engine algorithms use lemmatization, the user can query any inflectional form of a word and get relevant results. For example, if the user queries the plural form of a word such as "routers," the search engine knows to also return relevant content that uses the singular form of the same word -- "router."

STEMMING:

Stemming is the way to reduce a word to its word stem that affixes to suffixes and prefixes. In simple term, This algorithms work by cutting off the end or the beginning of the word while taking into account a list of common prefixes and suffixes that can be found in an inflected word.

Form	Suffix	stem
Cats	-s	cat
Birds	-s	bird

NEED OF STEMMING:

- Less input dimensions
- Machine Learning techniques work better with it
- Make training data more dense
- Reduce the size of the dictionary
- Helps to normalize the word in the document

Difference between stemming and lemmatization

In linguistics, lemmatization is closely related to stemming, as both strip prefixes and suffixes that have been added to a word's base form.

Stemming algorithms cut off the beginning or end of a word using a list of common prefixes and suffixes that might be part of an inflected word. This process is generally indiscriminate and can result in base forms of a word with incorrect spelling or meaning. Stemming operates without any contextual

knowledge, meaning that it can't discern between similar words with different meanings.

For example, the stem of "studies" and "studying" would be "studi" and "study," while in lemmatization the base form would be "study" for both "studies" and "studying." But both lemmatization and stemming would still have the same base form for the word "walking," for example. While being less accurate, stemming is easier to implement and runs faster. An example of stemming and lemmatization is shown as follows:

Stemming:

Study → Studi

Studying → Studi

Studies → Studi

Studied → Studi

Studier → Studier

Lemmatization:

Study → Study

Studying → Study

Studies → Study

Studied → Study

Studier → Study

With stemming, most inflections of the word "study" become "studi" compared to lemmatization where most outputs become "study."

Lemmatization is more complex than stemming, as lemmatization requires words to be categorized by a part of speech as well as by inflected form. This can become quite complicated in languages other than English, whose only inflected forms are singular or plural, verb tense and comparative or superlative forms of adverbs and adjectives.

Topic	Stemming	Lemmatization
Goal	Reduce inflectional forms (Stemming refers to the crude heuristic process which chops offthe ends of the words in order toachieve the goal correctly)	Reduce inflectional forms(Lemmatization refers to do the things properly with the help of a vocabulary and morphological analysis of words)
Implementation	stemmers are typically easier to implement and run faster compareto lemmatization	Lemmatization is difficult to implement

LOWER CASING:

Converting a word to lower case (NLP -> nlp). Words like *Book* and *book* mean the same but when not converted to the lower case those two are represented as two different words in the vector space model (resulting in more dimensions).

```
sentence = "Books are on the table."
sentence = sentence.lower()
```

```
print(sentence)
```

Output: books are on the table.

STOP WORDS REMOVAL:

Stop words are very commonly used words (a, an, the, etc.) in the documents. These words do not really signify any importance as they do not help in distinguishing two documents. Stop words are basically a set of commonly used words in any language, not just English.

Stop words are available in abundance in any human language. By removing these words, we remove the low-level information from our text in order to give more focus to the important information. In order words, we can say that the removal of such words does not show any negative consequences on the model we train for our task.

Removal of stop words definitely reduces the dataset size and thus reduces the training time due to the fewer number of tokens involved in the training.

We do not always remove the stop words. The removal of stop words is highly dependent on the task we are performing and the goal we want to achieve. For example, if we are training a model that can perform the sentiment analysis task, we might not remove the stop words. Before removing stop words, research a bit about your task and the problem you are trying to solve, and then make your decision.

Movie review: "The movie was not good at all."

Text after removal of stop words: "movie good"

We can clearly see that the review for the movie was negative. However, after the removal of stop words, the review became positive, which is not the reality. Thus, the removal of stop words can be problematic here

Need to Remove;

- They **provide no meaningful information**, especially if we are building a text classification model. Therefore, we have to remove stopwords from our dataset.

- As the frequency of stop words are too high, removing them from the corpus results in **much smaller data in terms of size**. Reduced size results in **faster computations on text data** and the text classification model need to deal with a **lesser number of features** resulting in a robust model
- **Stop words**, which are highly occurring words in the document such as ‘a’, ‘an’,’the’,’is’,’was’,’will’,’would’ etc.

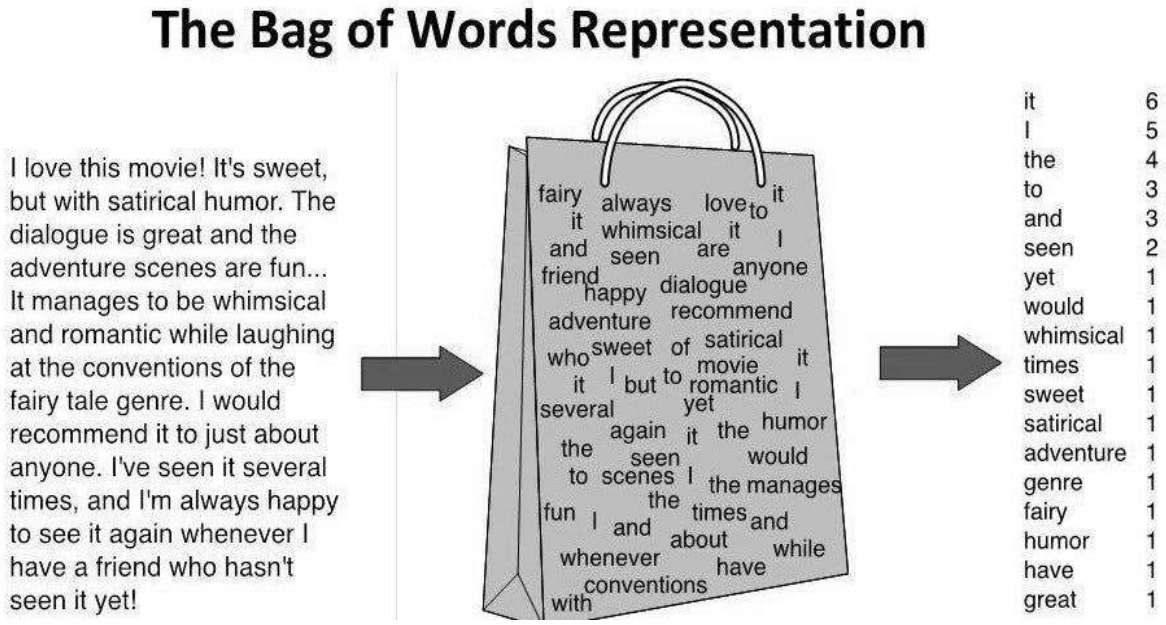
- **Significant words** are those words that have a moderate frequency in the document and **add actual meaning to the text**. These words are more important than stop words.

- **Rarely occurring words** are those words that occur with very less frequency and have relatively lesser importance than significant words. Rarely occurring words may or may not be helpful in understanding the context of the text.

Modelling Techniques:

Bag of Words

Bag of Words is a technique used in Natural Language Processing and in order to create new features for training classifiers This technique is implemented by constructing a histogram counting all the words in our document (not taking into account the word order and syntax rules). Bag of Words model is used to preprocess the text or documentations. It converts the documents into a bag of words, which keeps a count of the total occurrences of most frequently used words. Bag-of-Words is one of the most used methods to transform tokens into a set of features



One of the main problems which can limit the efficacy of this technique is the presence of prepositions, pronouns, articles, etc… in our text. In fact, these can all be considered as words which are likely to appear frequently in our text even

without necessarily be really informative in finding out what are the main characteristics and topics in our document.

In order to solve this type of problem, a technique called "Term Frequency-Inverse Document Frequency" (TFIDF) is commonly used. TFIDF aims to rescale the words count frequency in our text by considering how frequently each of the words in our text appears overall in a large sample of texts. Using this technique, we will then reward words (scaling up their frequency value) which appear quite commonly in our text but rarely in other texts, while punishing words (scaling down their frequency value) which appear frequently in both our text and other texts (such as prepositions, pronouns, etc…).

What is TF-IDF?

- This helps to measure the score in order to get the information retrieval (IR) or summarization.
- TF-IDF stands for Term Frequency and Inverse Document Frequency,
- TF-IDF is also used to reflect how relevant a term is in a given document
- Procedure to calculate TF-IDF by multiplying two metrics:

- How many times a word appears in a document,

- And the inverse document frequency of the word across a set of documents

Why do we need TF-IDF?

- 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 offset by the number of documents that appear in the corpus.

- TF is the frequency of term divided by a 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.

What is word embedding?

Word Embeddings vectors are one of the most common way to encode words as vectors of numbers those vectors can ben fed in into the Machine

Learning models for inference and also it helps to establish the distance between two tokens.

NLP Modeling is the process of recreating excellence. We can model any human behavior by mastering the beliefs, the physiology and the specific thought processes (that is the strategies) that underlie the skill or behavior.

THREE PHASES OF MODELLING:

Phase 1: Observing the model

his involves fully imagining yourself in someone else's reality by using what NLP calls a second position shift.

The focus is on "what" the person does (behavior and physiology), "how" they do it (internal thinking strategies) and "why" they do it (supporting beliefs and assumptions).

Phase 2: Find the difference that makes the difference

Traditional learning adds pieces of a skill one bit at a time until we have them all. The drawback to this method is we don't know which bits are essential. By contrast, modeling which is the basis of accelerated learning, gets all the elements and then subtracts to find what is necessary.

Phase 3: Design a method to teach the skill

Until you have all the relevant pieces of a skill and the necessary sequence, you cannot teach it effectively. We

currently teach many skills with extra background information and pieces muddying the waters.

Rehearsal of the natural sequence of the skill is important. If you tried to make a cake by putting it in the oven before mixing the ingredients together, it would be yucky. Yet we think we can teach separate elements of skills out of sequence and out of context and succeed.

What is sentiment analysis?

Sentiment Analysis is a technique which is commonly used to understand the positive, negative or neutral sentiment about a particular topic.

It is an unsupervised Machine Learning technique.

BIBLIOGRAPHY:

https://nlp-mentor.com/nlp-modeling/

https://www.mygreatlearning.com/academy/learn-for-free/courses/introduction-to-natural-language-processing

https://www.ibm.com/topics/natural-language-processing

https://www.deeplearning.ai/resources/natural-language-processing/