Notes

* NLP is used to convert text into vectors.
* In companies, 2-3 months are spent on text preparation.
* If you don't have data, 10-15 days are spent acquiring it.
* Text preparation is very tedious tasks because data are in multiple languages. So, first filter out data in separate language then convert all languages into English then see how much information will be lost.
* That's why NLP needs hard work and there are fewer resources available in the market.
* If there is a requirement of 4 people in a team, only 2 people are working on that team because there are fewer resources available in the market. If you get that opportunity, grab it because the packages are good in NLP.
* You can see offline as well as online, many companies or websites ask for reviews or comments. These things are collected for further use in NLP.
* For NLP tasks, we use the NLTK and spaCy libraries.
* For doing the same task, we have multiple ways in NLP. For example, to convert uppercase into lowercase, we can use the lower () function of strings, we can use the NLTK or SpaCy library, or we can use a web scraping tool such as Beautiful Soup or Scrapy.
* For web scrapping Beautiful Soup or Scrapy library.
* The good thing about NLP is that the codes are complex, but once you understand the code, the same codes are used in every project / real time because the data cleaning process is the same in all companies.
* We can make a project in NLP in which we pass the whole ‘Games of Thrones’ books and the model will summarize it.
* Learn data pre-processing steps like tokenization, lemmatization, stemming, and the difference between lemmatization and stemming (important interview question) and their codes.
* For Spelling correction, we are using py spell checker and text blob library.
* **After lemmatization, the meaning of the word is not changed. Nouns remain nouns and verbs remain verbs, and the root word has an English dictionary meaning.**
* **After stemming, there may be a chance that the root word does not have an English dictionary meaning. For example, the word "many" would be stemmed to "mani", which does not have an English dictionary meaning.**

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Interview Questions

Q1) What will happen if we do stemming instead of lemmatization?

Q2) Difference between stemming and lemmatization?

Q3) How to remove punctuation marks?  
\* In this question talk about different ways to remove punctuation mark.

Q4) if you don’t want to remove (‘) apostrophe in sentence how you can do?

Q5) Different libraries and ways to perform stemming, lemmatization, tokenization.

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Text Preparation in Detail

\* **Now, let's see how we prepare data in NLP. This is one of the most important parts of NLP because it will create a solid foundation for your NLP knowledge.**

\* **The next step is feature engineering. If the data is not cleaned properly, then the text-to-number conversion will also be poor. This means that the model will not be trained on good data, and it will be unable to perform well. Therefore, good data cleaning is essential.**

\* **Text preparation can involve 10-12 steps for data cleaning. However, it is not necessary to perform all of these steps in every case. The specific steps that are needed will depend on the problem statement, the type of data that is available, and the type of application that is being developed.**

\* You must be aware of the purpose of preparing text. If you are performing data cleaning, you must know why this step is necessary. These are the kinds of questions that are often asked in interviews, such as "Why did you apply this technique?" and "What is the usefulness of this?"

1) Convert into lower case: **After converting text from uppercase to lowercase, the model will treat "he" and "He" as the same word. This is because when text is converted to numbers, each unique word is assigned a separate column. So, if we do not convert the text to lowercase, the model will create two separate columns for "he" and "He." This will increase the training time of the model and also affect its accuracy.**

2) Remove HTML Tags (< >): **When you upload an image or text in comments or reviews, in some cases, the website or image format may not be compatible with the browser version or website. As a result, the data may contain HTML tags (< >). These HTML tags are not required for modelling because they do not contain any meaning. Any information that is contained within these HTML tags (<inside this>) will be removed because it is irrelevant.**

Multiple ways to remove HTML tags:

(i) Use regular expressions.

(ii) Beautiful soup library has multiple function by which we can remove HTML tags.

3) Removing URL’s: **In comment sections and WhatsApp chats, we send URLs to others, but your model cannot open them or understand them. Your model only understands the text that you are sending. Therefore, the URL is irrelevant, so we remove it. We use regex functions to remove URLs.**

4) Removing the punctuation marks: There are 15-20 punctuation marks.

\* We do not remove punctuation marks separately. We have a function that takes a list of punctuation marks as input and removes all of them from the text.

\* If we remove punctuation marks from text that contains abbreviations, such as "B.com" and "He's," the meaning of the text will be lost. This can cause problems with the accuracy of natural language processing (NLP) models.

\* To avoid this problem, we can tell the function to remove all punctuation marks except for periods and apostrophes. This is not a built-in function, but it can be created by using regular expressions.

5) Chat Word Treatment: **While chatting, we use abbreviations such as TTYL, ASAP, and others. These abbreviations are not understood by the model, so we convert them into their full forms. We want your model to be able to understand the full forms of these abbreviations, not the abbreviations themselves.**

6) Spelling Correction: **For spelling correction of data, there are direct libraries available. Some are paid and some are free. We are using free libraries.**

7) Removing the Stop Words: **Stop words** are common words that do not have much meaning on their own. They are often used in both positive and negative reviews, so they are not very helpful for determining the sentiment of a review. For example, the words "is," "am," "are," "the," and "has" are all stop words.

\* **If a customer writes “good”, this is not a stop word, so we understand that the customer is saying something positive.**

\* **We are using a library to remove stop words. The library has a list of all the stop words in English. It checks the data for any stop words and removes them if it finds any.**

\* **We remove stop words to compress the data. The fewer unique words there are, the smaller, faster, and better the model will be.**

\* **The more unique words in the vocabulary, the more time it will take to train the model and the more complex the model will be.**

\* **In real time, the number of unique words in the vocabulary can be in the hundreds of thousands.**

**It is necessary to decrease the number of unique words, otherwise it will create a problem.**

8) Handling emojis and emoticons: **Emojis are images, and we cannot create them using a keypad. However, we can create emotions using a keypad, such as :) and :(.**

There are 2 ways to deal with emojis and emoticons:

(i) Remove them

(ii) **We can convert emojis into text using some pre-defined templates. These templates are complex, but we don't need to remember them because they are used in every project.**

9) Tokenization: Tokenization means break down whole text into sentences or words.

There are 2 types of Tokenization:

(i) Word Tokenization

(ii)Sentence Tokenization

10) Stemming

11) Lemmatization

**\* Stemming and lemmatization are both processes that convert words into their root form. However, there is a key difference between the two: stemming can sometimes produce root words that do not have a corresponding entry in the English dictionary, while lemmatization always produces root words that have an English dictionary meaning.**

\* Lemmatization is used in chatbots where user directly see the output.