Work with the “Supercar racing” data mentioned in the slides. You need to pre-process the text and conduct text analysis.

Part 1:

* Pre-process the text
  + Read the *tweets\_sport.csv* data into Jupyter
  + Tokenize the text
  + Remove punctuation and stop words
    - adding your own stop words
  + Stemming and lemmatization
    - *Q1: What is the difference between stemming and lemmatization?*

Basic concept:

The basic concept behind stemming and lemmatization is the same as both are Natural Language Processing (NLP) techniques are used for reducing the inflectional and derivational forms of word to their base or root form

Difference:

|  |  |
| --- | --- |
| Stemming | lemmatization |
| Stemming is the process of reducing the word to its base form by removing suffix or prefix of the word. | Lemmatization is the process of reducing the word to its base form by taking into account the context and the part of speech of the word |
| The result of stemming may not be valid or be an actual word in the language. | The result of lemmatization is always a valid and an actual word in the language. |
| Stemming technique is faster. | Lemmatization technique is slow as compare to stemming. |
| Stemming technique is not reliable as it can give words that do not belong in the language. | Lemmatization technique is reliable as it always gives words that are part of the language. |
| Stemming technique do not create accurate result every time. | Lemmatization technique always create an accurate result. |

* Feature engineering
  + Bag of words
  + N-gram
  + TF-IDF
  + Word2Vec
    - *Q2: What is the difference between bag of words and N-gram?*

Basic concept:

The basic concept behind bag of words and N-gram is the same as both are Natural Language Processing (NLP) techniques are used to represent text data.

Difference:

|  |  |
| --- | --- |
| bag of words | N-gram |
| Bag of Words is a text representation technique that represents text as a collection of words without considering the order in which they appear in the document or text. | N-gram is a text representation technique that represents text as a collection of contiguous sequences containing N-words. |
| Bag of Words contain less contextual information then N-gram. | N-gram contain more contextual information then Bag of Words. |
| Bag of Words represent each word separately. | N-gram represent words in contiguous sequences containing N-word. |
| Bag of words take more space as it represents each word separately. | N-gram take less space as it represents each word in contiguous sequence. |
| Bag of words generate a spares representation of text. | N-gram generate a denser representation of text. |

* Conduct the following major text analysis
  + Name identify recognition
  + Sentiment analysis
  + Document summarization
* *Please go through the codes provided for the above analysis carefully and have a deep understanding of them*
* *Write down your answers on the notebook and in a separate word file*

Part 2:

Supervised learning:

About 1000 Tweets (*tweets\_sport\_positive\_labeled.csv*) have been manually identified as having positive sentiment or not. Please build a supervised learning model to predict the tweet sentiment for the unidentified tweets (*tweets\_sport\_positive\_to\_predict.csv*).

* Please use:
  + - Feature extraction: Bag of words, TFIDF and N-gram
    - And Random forest to predict
* Please use the best model (out of BOW, TFIDF and N-gram) to predict the sentiment of the unidentified tweets
  + *Q3: What are advantages of sentiment analysis based on supervised machine learning compared to unsupervised sentiment analysis?*

sentiment analysis:

sentiment analysis is a Natural Language Processing (NLP) technique used to

extract subjective information such as emotions from the text. It analyse a the given text and decided if the overall sentiment behind the text is negative, positive or neutral.

|  |  |
| --- | --- |
| supervised machine learning | unsupervised sentiment analysis |
| Supervised sentiment analysis models can achieve higher accuracy level in sentiment analysis as the trained data is labelled. | Unsupervised sentiment analysis models can be more for some applications only. |
| Supervised sentiment analysis models can be trained on specific domains which make it more flexible. | Unsupervised sentiment analysis models cannot be trained on specific domains which make it less flexible. |
| Due to label data the supervised sentiment analysis model are not cost effective. | Unsupervised sentiment analysis model is cost effective. |

*Final Delivery: .ipynb file output including answers to Q1-Q3 and predicted y (.csv file) and word document with answers to A1-Q3*