Work with the “Supercar racing” data mentioned in the slides. You need to conduct topic modelling

Part 1:

* Read the tweets\_sport.csv data into Jupyter
* Conduct topic modelling analysis
  + Generate the key words for top 3 topics
  + Generate the maps for top 3 topics
  + Combining them with the tweets, interpret the results

**Result Interpretation** After examining the results of the sentiment analysis on tweets about the Bathurst 1000 event, we discovered three key topics that were frequently brought up in the discussions. Based on the sentiment ratings assigned to each topic, which offer a numerical assessment of the positivity or negativity of the tweets.

* Topic 1 focuses on certain racers, including Morris, Chaz, Mostert, and Paul, as well as the anticipation for the competition. The sentiment score for the topic was 18.7, which shows that tweets about it were generally positive.
* Topic 2 focus on the general discussion about the Bathurst 1000 race, along with praise and support for the competitors' efforts. 'amp', 'like', 'day', 'race','start', 'believe', 'car', 'thanks', and 'home' are important words in this topic. The topic's sentiment score was 7.5, indicating that tweets about it tended to have a slightly positive score.
* Topic 3 focus on the technical elements of racing, such as automobile performance, safety precautions, and race strategy. 'Car','supercars', 'fuel', 'year', 'whincup','sport','make','safety', and'svg' are some of the key words in this topic. The sentiment score for this subject was 13.6, which means that tweets about it were generally positive.

The generated topics offer insightful information on the attitudes and conversations around the Bathurst 1000 event. Event planners, teams, and sponsors may improve the entire audience experience of the race event by analyzing these areas to better understand the audience involvement and make data-driven choices.

* *Write down your result explanations on the notebook and in the word file (.docx)*

Part 2: Combining topic modelling with sentiment analysis

* Generate sentiment scores for all the tweets (using *SentimentIntensityAnalyzer* from *NKTK* as shown in tute 8)
* Calculate the sentiment score for the 2 largest topics for each event day.
  + *Tips*
    1. *Get the sentiment scores weighted by the topic relevance probability (e.g., topic1\_setiment=sentiment score \* topic1 relevance score; topic2\_setiment=sentiment score \* topic2 relevance score*
    2. *Use the . groupby function (see codes in Tute 7) and variables month and day to generate the mean sentiment for the 2 topics for each day, e.g.,*

group\_tweet\_data = topics.groupby([pd.Grouper('month'),pd.Grouper('day')]).agg(topic1\_senti\_day

=('topic1\_senti', 'mean'),topics2\_senti\_day=('topic2\_senti','mean'))

**Result Interpretation** This code uses NLTK's SentimentIntensityAnalyzer to calculate the sentiment score of tweets. It then applies topic modelling using LDA on groups of tweets by event day. For each group, it selects the top two topics and calculates the sentiment score for each topic. Finally, it prints the sentiment scores for each topic for each event day.

The output shows the sentiment score for the top two topics of each event day. The sentiment score is calculated based on the compound score of the tweets that belong to a specific topic. A higher score indicates a more positive sentiment, while a lower score indicates a more negative sentiment.

For example, on event day 1, the sentiment score for topic 0 is 9.3031, and the sentiment score for topic 2 is 6.2282. This means that tweets belonging to topic 0 had a more positive sentiment compared to tweets belonging to topic 2 on that day. Similarly, on event day 12, the sentiment score for topic 4 is 514.4254, while the sentiment score for topic 3 is 370.2965. This indicates that tweets belonging to topic 4 had a much more positive sentiment compared to tweets belonging to topic 3 on that day.

* *Write down your result explanations on the notebook and in the word file (.docx)*