MS4S09 Assignment 1: 2020/21

Submission deadline: **9pm 26th January 2021**

Contribution to module: **40%**

It is a requirement that this assignment be completed using R.

A copy of your R code containing all tasks should be submitted via Blackboard Assessment in advance of the deadline. The R file containing all your code should also be sent to penny.holborn@southwales.ac.uk prior to the deadline.

Please ensure your code is commented appropriately, easy to follow and concise. Marks will be deducted for inefficient, ineffective, and uncommented code.

You are provided with a dataset titled as airline_coursework and saved in CSV format.

The data set consists of tweets which mention either of the airlines: Virgin US or United Airlines. It also has a pre-defined labelled sentiment class.

The dataset comprises of 8 variables as follows:

Variable Name	Description
tweet_id	An identifier for the tweet
sentiment	Sentiment label for the tweet either 'Positive', 'Neutral', or 'Negative'
airline	Name of the airline tweet is addressed to
retweet_count	#times this tweet has been retweeted
text	The text of the tweet
tweet_created	The date and time tweet was created
tweet_location	The location where tweet was created
user_timezone	The tweet creator's timezone

The aim is to analyse the tweets' and provide insights into the general trends and patterns of tweets addressed to both the airlines. To achieve this a series of specific tasks have been outlined.

Task A – Text Mining (25%)

You are required to utilise the pre-processing and Text Mining techniques shown to you in lectures in order to prepare and draw insight from the text provided and produce informative visualisations.

Task B – Sentiment Analysis (20%)

You are required to utilise the Sentiment Analysis techniques shown to you in lectures in order to make comparisons between airlines over time and also assess the "sentiment" provided in the data.

Task C – Topic Modelling (15%)

You are required to utilise the Topic Modelling techniques shown to you in lectures in order to cluster word groups and similar expressions that best characterise the tweets in order to identify hidden trends within the text.

Task D – Further exploration (15%)

You are required to utilise any further techniques shown to you in lectures or from your own research in order to draw meaningful insight from the text, looking to utilise all variables within the dataset.

Task E - Demonstration (25%)

You are required to deliver a 10-minute demonstration of your code summarising your main findings of each task. This demonstration will take place virtually to the Lecturer, following the submission deadline.

The Lecturer who will interact and ask questions to gauge understanding. At this point, you may be asked to provide information or explain parts of your code. The demonstration is used to test that you a) understand your code and b) can explain the algorithms utilised.

Marking Guidelines

	80-100	70-79	60-69	50-59	40-49	30-39	0-29
	Exceptional First	**First**	**Upper 2nd**	**Lower 2nd**	**Third**	**Narrow Fail**	**Fail**
Data pre-processing	Sophisticated pre-processing of data.	Comprehensive pre-processing of data.	Adequate pre-processing of data.	Pre-processing of data is attempted but with some flaws.	Limited pre-processing of data.	Inadequate pre-processing of data.	No pre-processing of data.
Analysis	Sophisticated text mining techniques utilised.	Comprehensive text mining techniques utilised	Adequate text mining techniques utilised.	Text Mining techniques attempted but with some flaws.	Limited text mining techniques utilised.	Inadequate text mining techniques utilised.	No text mining techniques utilised.
Key results and correctness of content	Unanticipated results and implementations presented. Appropriate, substantial, correct and sophisticated nature.	Comprehensive results and implementations, presented and employed well. Appropriate, substantial and correct.	Expected results and implementations presented. All appropriate, largely correct, with few flaws.	Not all expected results and implementations presented. All appropriate, largely correct, with few flaws	Few or simple results and implementations presented. Much appropriate material, but flawed.	Seriously flawed results or no implementation. Appropriate but seriously flawed material.	No results or implementation. Incorrect or inappropriate content.
Demonstration	Able to execute and explain the program clearly. Demonstrates an excellent level of understanding and explanation.	Able to execute and explain the program with no aid or guidance. Has a good level of understanding.	Able to explain the program with minor errors. Has a good but not complete understanding of the code.	Able to explain the program with some errors. Has some understanding of the code.	Able to explain the program with major errors. Makes an effort to explain the program but unable to do so.	Unable to explain in any detail the program that has been created Little awareness of the tasks or sections of code.	Unable to explain the program at all. No awareness of the purpose / location of any set tasks or sections of code.