PART II

This part of the final essay contains one question. It is worth 40 points. Again, 5 points are reserved for clarity of presentation, especially tables and figures. See Q+A session 5 for guidelines on presentation.

The question requires you to write a brief report. It is up to you how you structure the report, but it is advisable to keep introductory material to a minimum, given the word limit. Your report should discuss your methods, your results and the conclusions that you draw from them.

QUESTION C: Describing and Classifying Tweets [40 points]

Many companies monitor social media posts in order to gauge how customers feel about their company and their competitors. For this question, imagine that you have been hired as a consultant by one of the major American airline companies to analyse tweets about airlines. They want to find out how people talk about airlines on Twitter, and then build a predictive tool that can classify tweets in future into 'negative' or 'positive' sentiment toward airlines, to help them respond better to their customers in real time. They have provided you with a dataset of 11,541 tweets about airlines that have been labelled as 'negative' or 'positive' by their staff. The dataset also identifies which airline each tweet is talking about.

Your task is to prepare a brief report that describes the tweets, and recommends a classification method for future tweets. You need to:

i) Use appropriate tools to describe the tweets. In particular, what words are associated with negative or positive sentiment? How does word usage differ across the different airlines?

ii) Use your analysis from i) to build a short dictionary of negative and positive words describing airlines, then use it to classify tweets as 'negative' if they contain more negative than positive language, and 'positive' otherwise [code for creating your own dictionary is provided below]

iii) Use the lasso logit method to classify the tweets into 'negative' and 'positive'

iv) Compare the performance of your classifiers from ii) and iii), and use this analysis to decide which one would be the better classifier for the company to use for future tweets

The dataset for this question is called "tweets" and is contained in the file "tweets.Rda". It contains the following variables:

Variable name	Variable description
text	The text of each tweet
sentiment	Labeled sentiment of each tweet: 1=negative, 0=positive
airline	The airline company featured in the tweet: United, JetBlue, American Airlines, US Airways, Virgin America or Southwest

You should first create a corpus of tweets using the following code:

```
tweetCorpus <- corpus(tweets$text, docvars = tweets)
```

Here is some advice for part ii):

- Your dictionary should contain a minimum of 5 words and a maximum of 15 words in each category
- You are not expected to exhaustively compare the performance of different dictionaries. Instead, simply choose **one** dictionary based on your analysis from i), explaining how you chose the words.

Code for creating a dictionary:
You can create a dictionary called "mydict" in R that contains two categories ('negative' and 'positive') using the following code:

```
neg.words <- c()
pos.words <- c()

mydict <- dictionary(list(negative = neg.words,
                          positive = pos.words))
```

You need to insert your chosen sets of negative and positive words in 'neg.words' and 'pos.words'. This dictionary can then be used with quanteda in exactly the same way as any of the existing built-in dictionaries.