Coursework Administrative Details

Module/Lecture Course:	Natural Language Analysis
Deadline for submission:	Individual reports (1500 words):
Work returned:	TBC
Submission instructions:	Submit all files via duo
Format:	Report as a Word or pdf document. Accompanying data analysis for individual report as a Jupyter notebook. Do not put your name on your report, just your username.
Contribution:	The report contributes 100% to the final mark for the module.

In accordance with University procedures, **submissions that are up to 5 working days late will be subject to a cap of the module pass mark**, and **later submissions will receive a mark of zero**.

Content and skills covered by the assignment:

- Have a strong understanding of how to work with text and transform textual features to numeric features.
- Have an advanced understanding of advanced deep learning models for classifying and generating text.
- Select and implement appropriate feature extraction techniques from text.
- Train and test machine learning and deep learning classification models using real-world data.
- Effective written communication
- Planning, organising and time-management
- Problem solving and analysis

General Requirements

Students are expected to work on the coursework individually.

A dataset consists of news articles will be provided. The news articles cover four separate topics: World, Business, Sports, and Sci/Tech. The dataset contains 10,000 training and 1000 testing examples per topic. Every sample in both the training and testing data files consists of three fields: label (1 for World, 2 for Sports, 3 for Business, 4 for Sci/Tech), article title, and article text.

Students are expected to:

1- Implement natural language processing solutions to build accurate NLP classifiers in order to predict the topic of a given test example.
2- Implement text generation model to generate text for a chosen topic and then use the generated data to test the top performing NLP classifier.

Individual Report [100%]

Each student should separately develop their own NLP models to classify news articles into one of the four topics. Write a report (max 1,500 words) on your findings, which will be assessed as follows:

1) Apply the following feature extraction techniques and explain how they work and their advantages and disadvantages [**20%**]
   a) Term Frequency-Inverse Document Frequency (TF-IDF)
   b) Word2vec
2) Use the features extracted from the step above to train a standard Machine Learning algorithm e.g., SVM, Logistic Regression, Random Forest, and discuss its performance (accuracy) on the testing set [**10%**]
3) Train one Deep Learning model e.g., LSTM, RNN, CNN, using word2vec features, extracted in step1. Explain the architecture of the deep learning model, the hyper-parameters used, and the loss function [**20%**]
4) Analyse and compare the performance and training time results for both ML and DL models. [**10%**]
5) Build a text generation model using the training data of the topic ***'Sports'*** and explain how it works. Use it to generate 100 samples [**15%**].
6) Use the 100 generated samples from step 5 to test the performance of the machine learning and deep learning models developed in steps 2 and 3. Report and discuss the results [**15%**]

7) Academic English writing [**10%**], with good use of technical vocabulary, correct grammar, use of third person (i.e. do not write "I"), appropriate document structure and referencing where relevant.

You should submit your 1,500-word report and also the associated Jupyter notebook used to produce your analysis and graphs.

The report word count should:

- *Include* all the text, including title, preface, introduction, in-text citations, quotations, footnotes and any other item not specifically excluded below.
- *Exclude* diagrams, tables (including tables/lists of contents and figures), equations, executive summary/abstract, acknowledgements, declaration, bibliography/list of references and appendices. However, it is not appropriate to use diagrams or tables merely as a way of circumventing the word limit. If a student uses a table or figure as a means of presenting his/her own words, then this is included in the word count.

Examiners will stop reading once the word limit has been reached, and work beyond this point will not be assessed. Checks of word counts will be carried out on submitted work. Checks may take place manually and/or with the aid of the word count provided via an electronic submission.

Students are strongly advised to use Arial font size 12 for their assignments.

PLAGIARISM and COLLUSION

Your assignment will be put through the plagiarism detection service on duo.

Students suspected of plagiarism, either of published work or work from unpublished sources, including the work of other students, or of collusion will be dealt with according to Computer Science and University guidelines.