Big Data (301046) Assignment

Due: (Part I) Monday 28th May 2018 5:00 pm. (Part II) During class/tutorials in Week 14.

This assignment includes two parts. For Part I, you are given two options. Option 1 involves writing an essay on big data techniques, and Option 2 involves writing a program for movie sentiment classification. For Part II, you will present your Part I work in the lab class with PowerPoint slides. You are given 5 minutes for your presentation including question time. The whole assignment is worth 35% of the total assessment for this unit, with 20% for Part I and 15% for Part II.

Part I. Essay/Program (20 Marks)

You should pick **one and only one** option below. Working on both options will not give you extra credit. Moreover, your Part II presentation will be based on what you did in Part I. Hence you have to choose carefully and wisely based on your experience, expertise and preference. Details of both options can be found below, including task description and marking criteria

Option I. Essay

Task Description

You are required to write an open-topic essay that discusses one of the important techniques used in big data, **for example**, Relational Database Management Systems (RDBMS), NoSQL databases, web APIs and data mining, cloud computing, MapReduce, Hadoop, predictive modelling, etc .

Note that your essay should focus on a single topic area in big data, and provide an in-depth and comprehensive discussion on important issues in the area of your focus. **Depth is much more important than coverage for this assignment**. Hence concentrating on a small topic with your own insight is much better than briefly touching everything superficially. For example, an essay talking about MongoDB or any other specific NoSQL database system and its usefulness in big data applications makes a good topic, whereas one that goes through all different types of general database systems does not, provided that both essays are about the same length.

Doing research is an important step for essay writing, as everything starts with proper reading and finding ideas from the reading materials. In the beginning, you can read the lecture notes to identify an area of your interest for the topic of your essay. However, to gain more understanding and insight, you should not confine your reading to the teaching materials alone. Instead, you need to reach out and look further into the issues of interest by checking out relevant literature, including but not limited to, websites and various online resources, reference books, published journal articles, etc.

Your essay should also be well organised and structured into sections with headings. Each section should focus on a single main point, for e.g. introduction, current techniques and issues, possible future development, summary, etc. A section can usually be further divided into paragraphs depending on the points being covered. You should also include a bibliography or reference section in the end, which contains references to all articles, books and online links that have been cited in the main article. Size 14 font and 1.5 line spacing will be required. Your essay should contain between 2,000 to 3,000 words excluding the reference section. **Moreover and importantly, your essay should not just be about facts and findings from the literature, but**

should also include your own understandings and opinions on the topics discussed backed up by your readings.

Marking Criteria

Your final submission will be marked against the marking criteria below. Marks shown here are in the scale of 100 and only serve as a guideline. Fractional marks may be given for each criterion

Marking Criteria	Level	Mark	Description
Objective (10%)	Good	10	The essay has clear aim and is well focused on a single important topic in big data
	Average	5	Mainly focusing on a single area of study, with some deviations and irrelevant discussions
	Poor	0	Not focusing on a single area of study
Presentation (20%)	Good	20	The essay reads very well, easy to understand and free from grammatical and spelling errors, and use of colloquial English
	Average	10	The essay reads well in general, though there might be certain degree of grammatical errors and lack of clarity in presentation
	Poor	0	The essay is poorly written and can hardly be understood
Originality (20%)	Good	20	The essay contains some original thoughts and analysis from the author alongside discussions on facts and observations
	Average	10	There are some original points made by the author but not sufficient enough
	Poor	0	The essay is basically a reiteration of what is said in the literature and there is virtually no insight or input made on the author's part
Technical Quality (30%)	Excellent	30	The essay is up to a high technical standard and reflects a good level of critical thinking from the author
	Good	20	The essay has good quality in general, but contains some errors and inaccuracies in the discussion
	Average	10	The essay does not meet the requirement on minimum word count, or contains many errors and inaccuracies in discussion
	Poor	0	The essay has poor quality and virtually no technical value
Organisation (10%)	Good	10	The essay is well structured into sections. Each section covers one major point and further divided into paragraphs for clarity. The essay adheres to the formatting specification and word limit.
	Average	5	The essay is structured into sections and paragraphs, although not all of them make sense. The essay partly adheres to the formatting specification and word limits.
	Poor	0	The essay is poorly organised. The essay does not adhere to the formatting specification or word limit.
References (10%)	Sufficient	10	The essay contains sufficient references in bibliography section, which are further cited in the main text
	In-adequate	5	References included but are insufficient; or references not cited in the main text
	Absent	0	Lacking references

Note on originality of work: This is an individual assignment. No group work or collaboration is allowed. All submissions will be examined by the Turnitin system (http://turnitin.com) for plagiarism detection. Any submission that fails the Turnitin test will be given **ZERO** mark and may lead to further investigation on academic misconduct.

Option 2. Program

Task Description

For this task, you will create a complete program to perform sentiment classification for movie reviews. You will use a large movie review dataset containing a set of 25,000 movie reviews for training, and 25,000 for testing. You can visit the following website for more information about the dataset and the downloading link to the zip file.

http://ai.stanford.edu/~amaas/data/sentiment/

Download the dataset and unzip it to the local directory. Enter the **aclImdb/** directory created by the zip file, you will find five items

`train/` - feature files and raw text files for the training set

`test/` - feature files and raw text files for the testing set

`imdb.vocab` - expected rating for each token

`imdbEr.txt` - text tokens for each feature index

`README` - the readme file for more information on the dataset

For the purpose of this task, you only need the following files under the **train/** and **test/** directories.

`train/labeledBow.feat` feature vector file for the training set

`test/labeledBow.feat` feature vector file for the testing set

Each feature vector file contains 25,000 lines, each line represents label value and feature vector of word occurrences for the corresponding movie review in the training/testing set. For e.g., the following is the first line of **train/labeledBow.feat**

`9 0:9 1:1 2:4 … 47304:1`

This means that the first review gets a rating of 9, **0:9** for 9 occurrences of the word "the" (the first token in imdb.vocab), **1:1** for 1 occurrence of the word "and", **2:4** for 4 occurrences of the word "a", where "the", "and", "a" are the first three tokens in **imdb.vocab** file, and the last token **47304:1** for one occurrence of the word "pettiness", the 47305th token in **imdb.vocab**.

The above input vector basically calculates the number of occurrences for each word/token appearing in the raw text. The features are highly sparse, i.e. the majority of the entries are zero with only a few non-zero values. Your program should be able to read data from the training/testing files and parse them into the label vector and feature vectors for all 25,000 input examples. The input files (*.feat) are in a format called libsvm / svmlight and can be read into a matrix using the sklearn.datasets.load_svmlight_files function (see http://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_svmlight_files.html). To then use the data for training classification model, you need to perform feature normalisation to the feature vectors. A recommended normalisation scheme for text data is the TF-IDF scheme. More information can be found in the lecture notes or online resources.

After reading the file, parsing the data and computing the TF-IDF metric, you need to train a classification model to differentiate between movies with positive and negative feedbacks depending on the reviews and ratings. This is a standard binary classification problem by treating all reviews with >5 rating scores as the positive class and those with <=5 rating scores as the negative class. You can use any classification model you prefer, including but not limited to the Support Vector Machine (SVM), Decision Trees, Random Forest, K-Nearest Neighbour (K-NN), and the Naïve Bayes classifier. Due to the size of the training and testing set, you are best advised to employ an efficient classification model (e.g. Linear SVM, Decision Tree, Random Forest, Naïve Bayes). All of these models have been implemented in the scikit-learn Python package. **Check lecture notes and scikit-learn online document for further information.**

Note the performance of any classification model depends heavily on the choice of parameters to control the bias and variance trade-off, e.g. regularisation parameter C for linear SVM, tree depth for Decision Tree, depth and number of trees for Random Forest. Hence in addition to classifier training, you also need to implement proper function for parameter selection in your program that chooses the optimal parameters for the classification problem. This can be done by the cross validation procedure discussed in the lectures. You can also find resources online for discussions on parameter selection and cross validation.

To compare the performances of different classification models, you need to implement at least two different types of models and compare their predictive performances on the movie reviews dataset in your program in order to get full mark for this task.

Marking Criteria

Your program will be marked against both functional and operational requirements. Functional requirements accounts for 80% of the mark, which measure how well your program achieves the expected functionality and are further broken down into the following list of criteria. Again, note that fractional marks might be given for each of the criterion listed below.

- **File Reading (10%)** Your program can successfully read the text feature vectors and label values from the text file into the memory.
- **Pre-processing of input data (10%)** Your program can apply
- **Parameter selection (15%)** Your program should include the cross validation function for selecting the optimal parameters for classification model training.
- **Model Training (30%)** Your program should be able to complete classifier training on the training dataset, using at least two classification models mentioned above. You will get a maximum of 20% if only one classification model is implemented in your program. Moreover, parameters for each classification model should be properly selected using the cross validation technique mentioned above.
- **Prediction and display of results (15%)** Your program should make predictions on the testing dataset and save the prediction results into a csv file in the following format.

```
0, Positive
```

1, Positive

…

24999, Negative

Here the first number on each line denotes the index of testing data (from 0 to 24999 for the first up to the 25000th reviews), and the second text gives prediction result (positive or negative for each review).

Your program should also calculate the test error for each classification model and display them on screen by comparing the prediction results against true labels for the testing data, which can be extracted from the test feature file **test/labeledBow.feat** following the same method for obtaining labels for training data discussed above

In addition to function requirements, your program should also meet the operational and style requirements, which can be broken down into the following list of criteria.

- **Readability (5%)** Comments should be included in your program to explain the main idea of your design; use meaningful variable and function names; do not declare variables that are not used in the program
- **Modularity (5%)** Your program should make use of functions or classes wherever possible to achieve modular design.
- **Useability (10%)** Your program should be easy to use by the user. These include displaying messages for user interaction, performing adequate input validation, and allowing the users to choose the locations of the data file for model training.

Part II. Presentation (15%)

For Part II, you will create PowerPoint slides to present the work you did in Part I during week 14's lecture and practical classes. You will be allocated 5 minutes including question time to go through the content of your essay or explain the design and result of your program to the audience (tutor and rest of the class) followed by questions from them. **It is important to note that your work for Part I and Part II should be consistent and cover the same content**[1]**. Failure to do so will lead to a 0 mark for Part II.** Also, your performances in both delivering the presentation and answering the questions will be taken into consideration for marking.

Important instructions for class attendance:

Due to the number of enrolled students, the presentations will take place in both the lecture time and both practical class times. A schedule of the presentations will be announced during week 13. Every effort will be made to schedule students within their nominated practical class otherwise during the lecture time. Make sure you attend the class you have been allocated from the beginning to ensure participation of your presentation and completion of allocated peer assessment tasks. As the presentations contain a peer assessment component, you are required to be present for entire class duration. Leaving early or arriving late is not permitted. **You will lose your marks for Part II if you fail to meet these attendance requirements.**

Marking instructions:

The mark for your presentation will be determined by **peer assessment**, i.e. weighted average of the marks given by the tutor and other students. To implement this, each student has to mark four randomly selected presentations given by other students in class. Equivalently, each presentation will be marked by four peer students and the tutor. Tutor's mark and students' marks will count 30% and 70% respectively in the final mark. Your mark for Part II also depends on your obligations on peer assessment, and you will lose 10% of the mark for each presentation you failed to mark. To complete the marking, you need to fill out the marking sheet (one for each presentation you mark) attached in Appendix A, where you can also find the detailed marking criteria.

To facilitate with the marking process, make sure that the marks you give for peer assessment are objective and discriminative. That is, you should give high marks to students delivering excellent presentations, and lower marks for unsatisfactory ones. Note that your peer assessment marks will not be considered in determining the final marks if you consistently give the same good (or bad) marks to all students or show clear signs of bias.

1This means if you submit an essay on NoSQL database for Part I, you should talk about NoSQL for your Part II presentation but not anything else. If you write the program for Part I, you should talk about your program for your presentation and nothing else.

Presentation Marking Sheet **Presenter Student Number: _____**

Grader Student Number: _____

	Excellent (2.5-3)	Satisfactory (1-2)	Poor (0-0.5)	Mark
Content	Comprehensive coverage of the topic being discussed with good focus	The topic well covered in general, though the focus or coverage could be further improved	Poor coverage of the topic, lacking both completeness and focus	
Clarity	All materials clearly presented, all concepts were well explained, the audience was engaged	Majority of the materials were clearly presented, though there might be some ambiguities	Poorly covered materials, technical concepts were poorly explained to the audience	
Understanding	The presenter shows extensive knowledge of the topic and well understands all technical details	The presenter shows good knowledge of the topic but does not understand all technical details	The presenter barely knows what he/she talks about	
Preparation	Well organised slides, well prepared talk with good timing, all questions being well answered	Well prepared in general, though questions might not be well answered	Poor preparation, poor timing, just reading off slides	
Impressiveness	Very impressive presentation (I will give the job to the presenter if this is a presentation for job interview)	Overall quality is good, but does not impress me that much (I will be reluctant to give the job to the presenter)	Overall impression is negative (definitely no job given to the presenter if I were the employer)	

1. List up to three things that you like most about the presentation.

_____

_____

_____

_____

2. List up to three things that you dislike most about the presentation.

_____

_____

_____

_____

3. Further comments.

_____

_____

_____

_____

_____

_____