Coursework

**Module code: CSMBD16 Big Data Analytics**

**Lecturer responsible: Frederic Stahl**

**Work to be submitted on-line via Blackboard by 8th January 2018 12:00 hours. Work will be marked and returned by 29nd January 2018.**

## DEVELOPING AN AUTOMATIC TEXT CLASSIFICATION MODEL

## Your task

The goal of this coursework is to derive an automatic text classification model using KNIME/KNIME Labs. The sample data for this task can be downloaded here: http://clair.si.umich.edu/clair/clairlib/10newsgroup.tar.gz. The data comprises texts from 10 different “news” categories.

Your aim is to develop a workflow/model that is able to separate (classify) all the categories. If you struggle to separate all 10 categories you can try to break the problem/data up into fewer categories. But your overall aim is to have a classification model that is capable to separate all 10 categories.

You must try alternative approaches in terms of algorithm settings and/or combinations of algorithms (i.e. different classifiers and different pre-processing methods). You must also evaluate your model using adequate classification evaluation techniques available in KNIME.

You are required to submit a report outlining your approach, discussing your decisions and settings. You must also outline the alternative approaches you tried and discuss why you didn’t include them in your final model. In addition you must include an evaluation of your model in the report.

Your report should be approximately 6 pages long, including figures and excluding references (if applicable).

**MARKING CRITERIA**

1. The table below shows what is typically expected of the work to obtain a given mark. The assignment carries 50% of the total course marks.

|  |  |
| --- | --- |
| **Classification Range** | **Typically the work should meet these requirements** |
| **First Class (>= 70%)** | The coursework demonstrates:   * An exceptional understanding of the principles of unstructured data analysis. * Demonstration of professional technical writing skills and style. * Excellent technical skills in implementing, testing and tuning the text classification workflow. * Clear presentation of the coursework. |
| **Upper Second (60..69)** | The coursework demonstrates:   * Good and deep understanding of the principles of unstructured data analysis. * Demonstration of appropriate technical writing skills and style. * Excellent technical skills in implementing, testing and tuning the text classification workflow. * Clear presentation of the coursework. |
| **Lower Second (50..59)** | The coursework demonstrates:   * Basic understanding of the principles of unstructured data analysis. * Demonstration of moderate technical writing skills and style. * Excellent technical skills in implementing, testing and tuning the text classification workflow. * Clear presentation of the coursework. |
| **Third (40..49)** | The coursework demonstrates:   * Satisfactory understanding of the principles of unstructured data analysis. * Demonstration of satisfactory technical writing skills and style. * Excellent technical skills in implementing, testing and tuning the text classification workflow. * Clear presentation of the coursework. |
| **Fail (<40)** | The coursework fails to demonstrate understanding of unstructured data analysis. Techniques and technical skills to implement and evaluate a text mining workflow. |

1. The table below shows the mark scheme for the assessment

|  |  |
| --- | --- |
| **Part of Submission** | **Marks Available** |
| **Model Design:** Are the design decisions sensible and fit for purpose for the text classification task? | 55 |
| **Structure:** How suitable is the adopted structure of the report, does it comprise an abstract, introduction, evaluation and discussion/conclusions? | 15 |
| **Experimentation and Evaluation:** Are the experiments conducted described in sufficient detail and adequate for the purpose of the model? | 20 |
| **Format and Referencing:** Is the format of the report adequate and has an appropriate referencing style been used? | 10 |