FIT5212 Semester 1 in 2020: Assignment 1

Wray Buntine

Marks	Worth 50 marks, and 25% of all marks for the unit
Due Date	Due 8th May, 2020 at 23:59pm
Extension	An extension could be granted under some circumstances. A special consideration application form must be submitted. Please refer to the university webpage on special consideration.
Lateness	For all assessment items handed in after the official due date, and without an agreed extension, a 5% penalty applies to the student's mark for each day after the due date (including weekends) for up to 10 days. Assessment items handed in after 10 days without special consideration will not be considered.
Authorship	This is an individual assignment. All work must be your own. All submissions will be placed through Turnitin.
Submission	Submission is 3 files: one PDF report, and one Jupyter notebook with a PDF print of it. The three files must be submitted via Moodle. All files will go through Turnitin for plagiarism detection.
Programming language	Python in Jupyter

Part 1: Text Classification

The content has been gathered from the popular academic website arXiv.org for articles tagged as computer science content (though some of these are in mathematics or physics categories). Training is 1990-2014 and testing is 2015 plus a bit of 2016. The fields are:

- *ID*: a unique alphanumeric ID.
- *URL*: a working URL if you prepend "http://".
- *Date*: the date in format YYYY-MM-DD.
- *Title*: the full title, though with non-ASCII characters modified and any "," deleted.
- *InfoTheory*: a "1" if it is classified as an Information Theorey article, otherwise "0".
- *CompVis*: a "1" if it is classified as a Computer Vision article, otherwise "0".
- *Math*: a "1" if it is classified as a Mathematics article as well, otherwise "0".
- *Abstract*: the full abstract, though with non-ASCII characters modified.

The *URL*, *Date* and *Title* are provided but not required in the assignment.

The three classes are *InfoTheory*, *CompVis* and *Math*. These can occur in any combination, so an article could be all three at once, two, one or none. Your job is to build three text classifiers that predict these three classes only using the *Abstract* field. You should present two different text classifiers for the three Boolean prediction tasks, and exactly one should be a neural network. That is, you develop three classifiers of the same kind of neural network that predict *InfoTheory*, *CompVis* and *Math*, and another three classifiers of the same kind (which cannot be neural networks) that predict the three Booleans again.

Each set of three classifiers can (and probably should) share processing and data structures. The first set is to be called "Neural Network Method" and the second set called "Machine Learning Method". They should be clearly marked in the notebook file. Each set should be evaluated and a full confusion matrix printed, giving 6 confusion matrices in all (2 methods by 3 Booleans). You may include 1-2 early experiments in your notebook for each of the two methods describing preliminary methods you tested. This would be used for the purposes of discussing your final method and the advantages of it.

Having done the analysis, report on and discuss the results in a section in the PDF report. How well did the algorithms work? What special preprocessing did you do and how did that improve things over other attempts you may have made? How did the neural network compare against the non-neural network? Can anything be said about the confusion matrices about differing proportions?

Part 2: Topic Modelling

The content has been gathered from news sites, containing the term "Monash University", or at least tagged with the label "Monash University" by an external annotator. Your job is to perform appropriate text pre-processing and preparation and then perform two runs of LDA using the `gensim.models.LdaModel()` function call. Select appropriate choice of pre-processing and parameters to develop model outputs that are informative. Use visualisation of some kind to analyse and interpret the output of the topic model. If necessary, read the original news articles

(URLs are in the CSV file). But, be warned the text extraction routine may have failed, so the text at the news URL may be different to the text in the CSV file!

Having done some analysis and interpretation, write a section for the PDF report describing your findings. What sorts of topics do you see? Are they all about Monash University? Are all top topic words comprehensible sets of words? What does this tell you about news about Monash University? Find some articles that are exemplers and use them to illustrate key topics (but don't insert full articles in your report, not enough room, just extract a few lines or the title). Your analysis should serve two purposes: (1) what sorts of news mentioning Monash University is there, and why is it mentioning the university, and (2) to describe how the topic modelling presents this and any advantages or shortcomings of topic modelling for the role in (1). This is a knowledge discovery task rather than a predictive task, so marks will be included for your ability to make novel findings from the topic model.

Submission

All Python code must be included in a single Jupyter notebook that must be submitted. This should have clear headings "Part 1: Text Classification" with two sections "Neural Network Method" and "Machine Learning Method" then followed by "Part 2: Topic Modelling". It should have the students name and ID embedded in the first comment (in Markdown). The name of the file should be "code_012345678.ipynb" where "012345678" is replaced by your own student ID. An example/ skeleton notebook file "code_012345678.ipynb" with appropriate headings is included with the datasets. To complete this, use the export option and export to PDF. Save this as "code_012345678.pdf"

The notebook should:

- have any special or unusual libraries indicated at the top of the file in commented out command form; they must be able to be installed from the standard Python repository,
    - e.g., "# !pip3 install gensim"
- assume the three datasets supplied exist in the current directory
- have been run successfully to completion prior to submission, so the results are all embedded in the notebook

The PDF file should print the last version of the notebook submitted.

All reports and analysis must be written up in a single PDF file no more than 8 pages long. This may duplicate elements of the above notebook. Anything beyond 8 pages will not be marked. The name of the file must be "report_012345678.pdf" where "012345678" is replaced by your own student ID. The 8 pages should be A4 size with standard margins and 11 point font or similar. The report should have two sections, "Part 1: Text Classification" and "Part 2: Topic Modelling".

Therefore, three files are to be submitted, "code_012345678.ipynb", "code_012345678.pdf" and "report_012345678.pdf" where "012345678" is replaced by your own student ID.