COMSM0089
Introduction to Data Analytics Coursework

Spring 2022, Lecturers: Edwin Simpson (unit director), Ian Nabney.

Deadline: 13.00 on Wednesday 11th May

Overview

This coursework will take you through the data analytics process for an example scenario, from processing text data to visualising information. As well as implementing data analytics methods and obtaining results, you should aim to demonstrate your understanding of the methods you use and critically evaluate these methods. Your work should also incorporate ideas from the lecture videos and lectorials.

We recommend that you first get a basic implementation for all parts of the required assignment, then start writing your report with some results for all tasks. You can then gradually improve your implementation and results.

Total time required: 40 hours.

Support

The lecturers and teaching assistants are available to answer clarification questions if you are unsure what to do for any part of the coursework. You can ask questions during our Monday labs, post questions to the QA channel on Teams or anonymously to the Blackboard discussion forum. Alternatively, use email to contact Edwin (edwin.simpson@bristol.ac.uk) about questions on tasks 1 and 2 and Ian (ian.nabney@bristol.ac.uk) for the other tasks.

Task 1: Sentiment Classification (max. 22%)

Financial news provides important information for investors, such as positive or negative sentiment towards a company. Your task is to design, implement and evaluate a sentiment classifier for financial news. For this task, we will be working with the ***Financial Phrasebank*** dataset, which contains sentences from English news articles discussing companies listed on the Helsinki stock exchange. Each sentence has one of three labels: positive (2), negative (0), or neutral (1). The dataset can be accessed through the HuggingFace datasets library. Please see the Jupyter notebook *data_loader_demo.ipynb (available on Blackboard)* for example code for loading and splitting it into training and test sets.

The data is described in this paper:

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. Journal of the Association for Information Science and Technology, 65(4), 782-796.

1.1. Implement and train a method for automatically labelling texts in the Financial Phrasebank with their sentiment labels. Refer to the labs, lecture materials and textbook to identify a suitable method. Include the following in your report:

- Briefly explain how your chosen sentiment analysis method works and its main strengths and limitations;
- Describe the features you have chosen and why you chose them, and hypothesise how they will affect your results;
- Explain the preprocessing steps your method requires.

(7 marks)

1.2. Implement, train, and test your method. Briefly document this process in the report. (6 marks)

1.3. Evaluate, interpret and discuss your results, making sure to include the following points:

- Define your performance metrics and state their limitations;
- Show your results using suitable plots, tables and/or a confusion matrix;
- How could you improve the method or experimental process? Consider the errors that your method makes.

(9 marks)

High performance figures are less important for getting high marks than motivating your method well and implementing and evaluating it correctly.

Suggested length of report for task 1: 2 pages.

Task 2: Named Entity Recognition (max. 28%)

Our clients would like to extract information automatically from financial documents about organisations, places, and people. This task is therefore to design and implement named entity recognition using the **SEC-Filings** dataset, containing U.S. financial agreements. The dataset is labelled with the entity tags location (LOC), person (PER), organisation (ORG) and miscellaneous (MISC). Code to load the dataset is provided in the Jupyter notebook *data_loader_demo.ipynb (available on Blackboard)*.

The data is presented in this paper:

Alvarado, J. C. S., Verspoor, K., & Baldwin, T. (2015). Domain adaption of named entity recognition to support credit risk assessment. In Proceedings of the Australasian Language Technology Association Workshop 2015 (pp. 84-90).

2.1. Design a method for tagging named entities in the SEC-Filings dataset. Refer to the labs, lecture materials and textbook to identify a suitable method. Include the following in your report:

- Briefly explain how your chosen named entity recognition method works and its main strengths and limitations;
- Describe the features you have chosen and why you chose them, and hypothesise how they will affect your results;
- Explain the tagging scheme for labelling entities in this dataset.

(7 marks)

2.2. Implement, train, and test your method. Briefly document this process in the report. (6 marks)

2.3. Evaluate, interpret, and discuss your results, making sure to include the following points:

- Explain your choice of performance metrics and their limitations;
- Show your results using suitable plots and/or tables;
- How could you improve the method or experimental process? Consider the errors your method makes.

(8 marks)

2.4. Apply your trained NER tagger to the Financial Phrasebank dataset.

- Compute a sentiment score for each entity that you detect. Briefly explain your method. One way you could compute a score for an organisation is to count the number of positive texts it occurs in and subtract the number of negative documents it occurs in;
- Show your results, for example by listing the five most positive and five most negative organisations, along with their scores.

(7 marks)

Suggested length of report for task 2: 2.5 pages.

Task 3 Information Visualisation Analysis (8%)

Analyse the approach you have used to present your results in tasks 1.3 and 2.3 as defined above.

3.1. Justify the design chosen in terms of key information visualisation principles. (5 marks)

3.2. Define and explain the visual queries that the user carries out when viewing your presentation of results. (3 marks)

Suggested length of report for task 3: less than 1 page.

Task 4: Information Visualisation (42%)

4.1. Use Tableau to create plots that enable the user to explore ***the Bookshop dataset*** that was used in lab 3. You should enable the user to answer these questions:

- Is there a link between the number of hours per day that an author writes and their total output (in terms of the total number of pages in their books)?
- Is there a link between ratings and sales at the book level?
- Show where authors live on a world map, how many work in each country (using a visual representation), and in a tooltip provide information on the average price for books written in that country.

In about two pages, write a short description of the visualization techniques you used and a justification for your choices. You should refer to the principles of info vis, relevant aspects of human perception and cognition, and the scientific literature where appropriate.

(32 marks: 22 marks for the visualization; 10 marks for the description and justification).

4.2. Using appropriate levels and types of validation (as in Chapter 4 of Munzner and the lectures from week 2), assess the quality of your visualization by making appropriate measurements and observations of the other students in your group (the groups will be defined separately) in an

analytic task using your visualisation. The lab class on 25th April will be dedicated to this activity, so you will need a complete visualization by then. Your report on this should be no more than one page. (10 marks).

Implementation

Text Analytics: The lab notebooks provide useful example code and we recommend using Python 3 with the libraries used in the labs. You may use other libraries if preferred and you can write your code in either Jupyter notebooks or standard Python files.

Information Visualisation: We recommend using Tableau and applying what you have learned in the labs and lectorials.

Report Formatting

- Maximum of 10 pages
- References do not count toward the page limit
- We recommend using the template from COLING 2020 if writing the report in Latex[1], or following the same formatting style if using Word or another application.
- No less than 11pt font
- Single line spacing
- A4 page format
- Aim for quality rather than quantity: you do not have to use the maximum number of pages and will receive higher marks if you write concisely and clearly.
- The text in your figures must be big enough to read without zooming in.

Citations and References

Make sure to cite a relevant source when you introduce a method or discuss results from previous work. The preferred style is given in the COLING 2020 style guide above. The details of the cited papers must be given at the end in the references section (no page limits on the references list). Please only include papers that you discuss in the main body of the report.

Google Scholar and similar tools are useful for finding relevant papers. The 'cite' link provides bibtex code for use with latex and references that you can copy, but beware they often contains errors.

Submission

- Deadline: 13.00 (GMT+1) on 11th May.
- On Blackboard under the "assessment, submission and feedback" link.

Please upload the following **three files:**

1. Your report as a **PDF with filename <student_number>.pdf,** where <student_number> is your student number (not your username).
2. Your code inside a **single zip file with filename <student_number>.zip.** Please remove datasets and other large files to minimise the upload size – we only need the code itself.

[1] Latex is the most common tool for writing published papers in AI/ML/NLP research. It separates writing the content from formatting. A good way to get started with Latex is to use https://www.overleaf.com/.

3. A **packaged** Tableau workbook (use this link to find out more) with filename **<student_number>.twbx** containing your solution to Task 4. This enables us to run the workbook in Tableau reliably.

We will briefly review your Python code by eye – we do not need to run it. Your marks will be based on the contents of your report, with the code used to check how you carried out the experiments described in your report. We will **not** give marks for the coding style, comments, or organisation of the code.

Please do not include your name in the report text itself: to ensure fairness, we mark the reports anonymously.

Please check that your submission follows these guidelines before uploading, otherwise **you may lose marks.**

Assessment Criteria

Your coursework will be evaluated based on your submitted report containing the presentation of methods, results and discussions for each task. To gain high marks your report will need to demonstrate a thorough understanding of the tasks and the methods used, backed up by a clear explanation (including figures) of your results and error analysis. The exact structure of the report and what is included in it is your decision and you should aim to write it in a professional and objective manner. Marks will be awarded for appropriately including concepts and techniques from the lectures.

Avoiding Academic Offences

Please re-read the university's plagiarism rules to make sure you do not unknowingly break any rules. Do not copy text directly from your sources – always rewrite in your own words and provide a citation.

Academic offences include submission of work that is not your own, falsification of data/evidence or the use of materials without appropriate referencing. Note that sharing your report with others is also not allowed. These offences are all taken very seriously by the University.

Suspected offences will be dealt with in accordance with the University's policies and procedures. If an academic offence is suspected in your work, you will be asked to attend an interview with senior members of the school, where you will be given the opportunity to defend your work. The plagiarism panel can apply a range of penalties, depending on the severity of the offence. These include a requirement to resubmit work, capping of grades and the award of no mark for an element of assessment.

Extenuating Circumstances

If the completion of your assignment has been significantly disrupted by serious health conditions, personal problems, periods of quarantine, or other similar issues, you can apply for consideration of extenuating circumstances in accordance with the normal university policy and processes. Students should apply for consideration of extenuating circumstances as soon as possible when the problem occurs. Please see the details here.