COMSM0089
Introduction to Data Analytics Coursework

Spring 2023, Lecturers: Edwin Simpson (unit director), Ian Nabney.

Deadline: 13.00 on Wednesday 24th May

Overview

This coursework will take you through several data analytics tasks involving processing text and 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 provide clarifications about what you are required to do for any part of the coursework. You can ask questions during our lab sessions or 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 task 3.

Task 1: Sentiment Classification (max. 31%)

Sentiment analysis of news and social media provides important information for investors to help them make informed business decisions. Your task is to design, implement and evaluate a sentiment classifier for financial news headlines and social media posts. For this task, we will be working with the ***FiQA Sentiment Analysis*** dataset, which contains English news headlines and social media posts related to finance.

We provide a copy of the data and a 'data_loader_demo' Jupyter notebook containing code for loading the data in our Github repository, https://github.com/uob-TextAnalytics/intro-labs-public. Each instance in the dataset has a continuous sentiment score from -1 to 1, which our data loader notebook maps to one of three discrete labels: positive (2), negative (0), or neutral (1). Further information about the data is available on the FiQA website: https://sites.google.com/view/fiqa/home. In the tasks below, you can use any data you wish, including other data provided by FiQA, to build your sentiment classifier.

The data is described in this paper:

Maia, Macedo & Handschuh, Siegfried & Freitas, Andre & Davis, Brian & McDermott, Ross & Zarrouk, Manel & Balahur, Alexandra. (2018). WWW'18 Open Challenge: Financial Opinion Mining and Question Answering. WWW '18: Companion Proceedings of the The Web Conference 2018. 1941-1942. 10.1145/3184558.3192301.

1.1. Implement and train a method for automatically classifying texts in the FiQA sentiment analysis dataset as positive, neutral or negative. Refer to the labs, lecture materials and textbook to identify a suitable method. In your report:

- Briefly explain how your chosen method works and its main strengths and limitations;
- Describe the preprocessing steps and the features you use to represent each text instance;
- Explain why you chose those features and preprocessing steps and hypothesise how they will affect your results;
- Briefly describe your software implementation.

(10 marks)

1.2. Evaluate your method, then interpret and discuss your results. Include the following points:

- Define your performance metrics and state their limitations;
- Describe the testing procedure (e.g., how you used each split of the dataset);
- Show your results using suitable plots or tables;
- How could you improve the method or experimental process? Consider the errors that your method makes.

(9 marks)

1.3. Can you identify common themes or topics associated with negative sentiment or positive sentiment in this dataset?

- Explain the method you use to identify themes or topics;
- Show your results (e.g., by listing or visualising example topics or themes);
- Interpret the results and summarise the limitations of your approach.

(12 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: 2.5 pages.

Task 2: Named Entity Recognition (max. 19%)

In scientific research, information extraction can help researchers to discover relevant findings from across a wide body of literature. As a first step, your task is to build a tool for named entity recognition in scientific journal article abstracts. We will be working with the ***BioNLP 2004 dataset*** of abstracts from MEDLINE, a database containing journal articles from fields including medicine and pharmacy. The data was collected by searching for the terms 'human', 'blood cells' and 'transcription factors', and then annotated with five entity types: DNA, protein, cell type, cell line, RNA. More information can be found in the paper: https://aclanthology.org/W04-1213.pdf .

We provide a cache of the data and code for loading the data in 'data_loader_demo' in our Github repository, https://github.com/uob-TextAnalytics/intro-labs-public. This script downloaded the data from HuggingFace, where you can also find more information about the dataset: https://huggingface.co/datasets/tner/bionlp2004 .

The data is presented in this paper:

Nigel Collier, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Jin-Dong Kim. 2004. Introduction to the Bio-entity Recognition Task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78, Geneva, Switzerland. COLING.

2.1. Design and implement a method for tagging the five types of named entities in the BioNLP 2004 dataset. Refer to the labs, lecture materials and textbook to identify a suitable method. In your report:

- Explain how your chosen method works and its main strengths and limitations;
- Briefly explain how entity spans are encoded as tags for each token in a text;
- Briefly describe your software implementation;
- Detail the features you have chosen, why you chose them, and hypothesise how your choice will affect your results.

(10 marks)

2.2. Evaluate your method, then interpret and discuss your results. Include the following points:

- Explain your choice of performance metrics and their limitations;
- Describe the testing procedure (e.g., how you used each split of the dataset);
- Show your results using suitable plots and/or tables;
- How could you improve the method or experimental process? Consider the errors your method makes.

(9 marks)

Suggested length of report for task 2: 2 pages.

Task 3 (50%)

3.1. Use Tableau to create plots that enable the user to explore two datasets relating to child health which are available on Blackboard or can be downloaded from the Global Health Observatory (by age https://apps.who.int/gho/data/node.main.nHE-1559AGE?lang=en and wealth quintile: https://apps.who.int/gho/data/node.main.nHE-1559?lang=en). Note that each figure comes with a confidence interval based on the fact that sampling was used to gather the data. For the purposes of this task, you may just use the main figure if you prefer.

You should enable the user to answer these questions:

- In which countries has child malnutrition improved over the period and in which countries has malnutrition got worse?
- Is there a link between wealth and child malnutrition?
- Show the values on a world map with information on both 0-1 years and 2-5 years appropriately presented.

In about two pages, write a 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. (40 marks: 30 marks for the visualization; 10 marks for the description and justification). Suggested length of report c. 3 pages.

3.2. 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 4th May will be dedicated to this activity, so you will need a complete visualization by then. You will also need to upload your Tableau packaged workbook to a special submission point on Blackboard. Your report on this should cover the testing framework used (with explanation for its design), an analysis of the results, and recommendations for changes/improvements in the visualization. This should be no more than 1.5 pages. (10 marks).

Implementation

Text Analytics: The lab notebooks provide useful example Python 3 code. 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. You can use the citation style 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 that this often contains errors.

Submission

- Deadline for report + code: 13:00 (GMT+1) on 24th May.
- Deadline for Tableau workbook solution for task 3: 17:00 on 3rd May.

[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/.

- 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 text analytics 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.
3. A **packaged** Tableau workbook (use this link to find out more) with filename **<student_number>.twbx** containing your solution to Task 3. This enables us to run the workbook in Tableau reliably. This should be submitted by 17:00 on 3rd May to a special submission point.

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.

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 break any rules. 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.

Do not copy text directly from your sources – always rewrite in your own words and provide a citation.

Work independently -- do not share your code or reports with others.

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.

Extensions and Extenuating Circumstances

If the completion of your assignment has been significantly disrupted by serious health conditions, personal problems, periods of quarantine, or other serious issues, you can apply for consideration in accordance with the normal university policy and processes. Students should refer to the guidance and complete the application forms as soon as possible when the problem occurs. Please see the guidance below and discuss with your personal tutor for more advice:

https://www.bristol.ac.uk/students/support/academic-advice/assessment-support/request-a-coursework-extension/

https://www.bristol.ac.uk/students/support/academic-advice/assessment-support/extenuating-circumstances/