MELBOURNE BUSINESS SCHOOL
MASTERS OF BUSINESS ANALYTICS
MODULE 3
LECTURERS: PROF TREVOR COHN & DR JEY HAN LAU

Text & Web Analytics Syndicate Project

Due dates: 10 Aug (proposal), 07 Sept (essay), 21 Sept (application)

1 INTRODUCTION

In groups (syndicates), the students will identify a text analysis problem of interest. This project will explore real-world application of Natural Language Processing or Information Retrieval methods to address a text analysis task.

There will be two primary components to this project:

1. **Syndicate essay on a Text Analysis task.** The students will prepare a report that identifies a textual data source for analysis, defines the text analysis task to be solved, proposes an approach to solving the problem, and discusses how the performance of that approach to the task will be evaluated. You will need to motivate your approach with respect to the research literature where relevant, including a clear presentation of the methods used and critical analysis of the strengths and weaknesses of existing approaches.
2. **Text Analysis application development.** In this portion of the project, students will build on the investigation done for the essay by implementing a solution to the problem for the selected data set, and evaluating the implementation against an appropriate baseline. You will report your findings in a written summary.

You are free to use external software and programming code in any language, but please keep to software that we will be able to run, e.g., limit yourself to open source projects or standard software (e.g., Python, SAS) installed on University or MBS machines. The same goes for any data you need: we would encourage you to use openly available resources, such as corpora bundled with `nltk` or on-line data sets, such as might be available via Kaggle (`https://www.kaggle.com/`) or data.world (`https://data.world/`). Ensure that you attribute your sources carefully, so we know which components are yours and form part of the assessment.

A list of project ideas is given below; but you are also free to choose another project of interest to you that meets the goal of exploring NLP and/or IR to address a text analysis objective. The project ideas below are just ideas, which will need to be elaborated as part of your project. If you are unclear about what is implied in any of the ideas, please ask for a fuller explanation.

2 REPORT TOPICS

You may choose one of the topics below, or another idea of your own which you may discuss with the lecturer or tutors to get feedback as to suitability. There are many evaluation resources available, and shared tasks or evaluations for both natural language processing and information retrieval topics. The topics below have evaluation data associated with them; not all tasks you might imagine will lend themselves so clearly to formal evaluation. In that case, you will need to consider either how you can acquire evaluation data, or how you would otherwise assess the effectiveness of your approach?

Please start by reading the lecture materials and recommended reading. Many of the topics below also have research papers associated with them describing the tasks (and some approaches for addressing the tasks). From this you should be able to find many more papers using Google Scholar (e.g., to see which papers cite a given paper, other papers by the same authors, or to search anew). Most papers will be easily accessed online, however you may need to use the library's search facility to access certain papers.

1. Movie review sentiment analysis `http://ai.stanford.edu/~amaas/data/sentiment/` or `http://www.cs.cornell.edu/people/pabo/movie-review-data/`
2. Amazon product reviews `http://www.cs.jhu.edu/~mdredze/datasets/sentiment/index2.html`
3. Processing the Enron email corpus to understand who is emailing who, and/or about what (e.g., what category does a message fall into).
   `http://bailando.sims.berkeley.edu/enron_email.html`
   (also at: `https://data.world/brianray/enron-email-dataset`) (See also Chapter 7 of the reference book *Mining the Social Web*).
4. Analysing TED Talks
   `https://data.world/owentemple/text-and-content-features-of-most-persuasive-ted-talks`
5. There are a number of opinion-related corpora that are available, including those at `http://mpqa.cs.pitt.edu/corpora/` such as the Product Debate data. These could be interesting to work with; please look over the papers that describe them before defining a specific task.
6. Recovery of missing information about word casing and punctuation in English text.
   `http://www.alta.asn.au/events/sharedtask2013/description.html`
7. Information Retrieval tasks. Please review the many tasks listed at `http://trec.nist.gov/tracks.html`; some data sets will be difficult to get but others will not. You may attempt to build an Information Retrieval system, or make use of an open source system and explore extensions, e.g., to query processing.
8. The NLTK book is distributed with a substantial number of corpora and linguistic resources. See `http://www.nltk.org/nltk_data/` for data for word sense disambiguation, parsing, named entity recognition, etc.

Here are some examples from prior year syndicate projects:

1. Extracting customer needs from TripAdvisor hotel reviews
2. Spam filtering
3. Sentiment analysis of restaurant reviews
4. Music genre classification via lyrics analysis
5. Product range sentiment analysis on Amazon
6. Analysis of AirBnB reviews (e.g. as available at `http://insideairbnb.com/get-the-data.html`)

Please choose carefully as some of the above topics will be easier than others. In particular, the availability of off the shelf tools, such as taggers and parsers (e.g., in Stanford CoreNLP), may influence your choice.

3 PROJECT SUBTASKS

3.1 PROPOSAL OUTLINE [DUE: TUESDAY 10 AUGUST, 11:59PM] [5% OF SUBJECT MARK]

Provide a brief sketch of the proposed topic for the Syndicate project for discussion and review. The students will work in groups identify a focus area for their project and begin to define the task to be addressed.

3.2 SYNDICATE ESSAY [DUE: TUESDAY 07 SEPTEMBER, 08:59AM] [20% OF SUBJECT MARK]

The students will prepare a report that identifies a textual data source for analysis, defines the text analysis task to be solved, proposes an approach to solving the problem, and discusses how the performance of that approach to the task will be evaluated.

Target length of essay: 2000 words (4-5 pages)

Suggested structure: Due to the open-ended nature of the project, many essay structures are possible; don't feel that you need to be limited to the structure here! However, a typical report might include the following sections:

1. **Introduction:** A high-level statement of the text analysis problem you are interested in solving. What is the broader context and significance of that task?
2. **Background:** A summary of previous approaches to the task, or similar/related tasks, and an analysis of the pros/cons of those approaches.
3. **Task specification**: A description of your approach to analysing your text data.
    a) **Data:** Proposed data source to be analysed
    b) **Methods:** Proposed approach to the analysis
    c) **Experimental Design:** Proposed approach to assessing the performance of your system

Assessment criteria: The projects will be judged on creativity in defining the problem to be investigated, the appropriateness of the proposed approach to the problem, including thoroughness in considering and justifying your design decisions, and the rigorousness of the proposed evaluation methodology.

This essay may address a "hypothetical" or "bigger" task than you will eventually be able to address in the Application portion of the project – in that case, be sure to discuss what the reasons are for not being able to tackle the bigger task. What can you achieve in this module, and where does it fit in addressing the bigger problem you would like to eventually solve?

- Originality and inventiveness of ideas proposed
- Quality and correctness of writing, including clear motivation
- Adequacy and quality of literature citations
- Thoroughness and appropriateness of proposed solution
- Rigorousness of proposed evaluation methodology

3.3 Application [Due: Tuesday 21 September, 08:59am] [20% of subject mark]

The students will implement a solution to the problem for the selected data set, and evaluate it. In addition, the students will prepare a report summarising the results of the implementation and evaluation.

Target length of report: 1500 words (3-4 pages)

Suggested structure: This report may have the more standard "IMRAD" structure of a research paper. IMRAD stands for *Introduction, Methods, Results, and Discussion.*

The Methods section should summarise what you actually did; what was the task? the data set? the approach you implemented (what does your solution do and what is the structure of your solution? what tools did you use? what was the experimental design that you used?). The objective is to describe your methods in enough detail that someone else could re-implement your solution on the same data set, and achieve the same results.

The Results section should summarise the numerical or analytical results on the data.

The Discussion section should interpret those results. How well does your solution solve the problem? How do you know? What problems did you run into (error analysis)? Do you have some insight into what would be needed to improve the solution, if you had more time to work on the problem?

Assessment criteria: The projects will be assessed on the appropriateness of the methods used, and quality of your write-up, including your testing of the system and reporting of results.

- Quality of presentation (including maths, figures, tables)
- Quality and correctness of methods description
- Correctness and thoroughness of experimental work
- Quality of insights derived from the experiment

For you to do well you do not need to report a positive result. It is perfectly adequate to report negative results (of your own, or synthesising others' results from the literature). What matters is your presentation, analysis and insights you have gained. You will not be penalised if your system performs poorly, providing your initial design decisions weren't obviously unjustifiable, and you have made reasonable attempts to analyse why it failed, and to examine how the system might be improved.

Please aim to be concise in your report, overlong or vague reports will be penalised. You will not have space to discuss the many technical aspects of your software implementation (if applicable), instead you should focus on the scientific questions and the knowledge you have gained. Remember that the report is the unit being assessed, not your code.

Submission:
Submission will entail two parts:

1. Your written report, as a single Portable Document Format (PDF) file.
2. A zip or tar archive containing your code, scripts and data files, as well as a README file that explains how to run your submission.

Submission will be via Canvas.

Late submissions will be docked marks at a rate of 10% of the project mark per business day, and no submissions will be accepted more than 5 business days (i.e., one week) late.