Assessment 1: NLP Recommendation Engine

Length: 2000 words (+/-30%)

Overview

Natural language processing (NLP) is commonly used to build recommendation engines. This assignment involves building reading recommendation engines for teachers in Australian schools based on textbooks lists sourced from public websites.

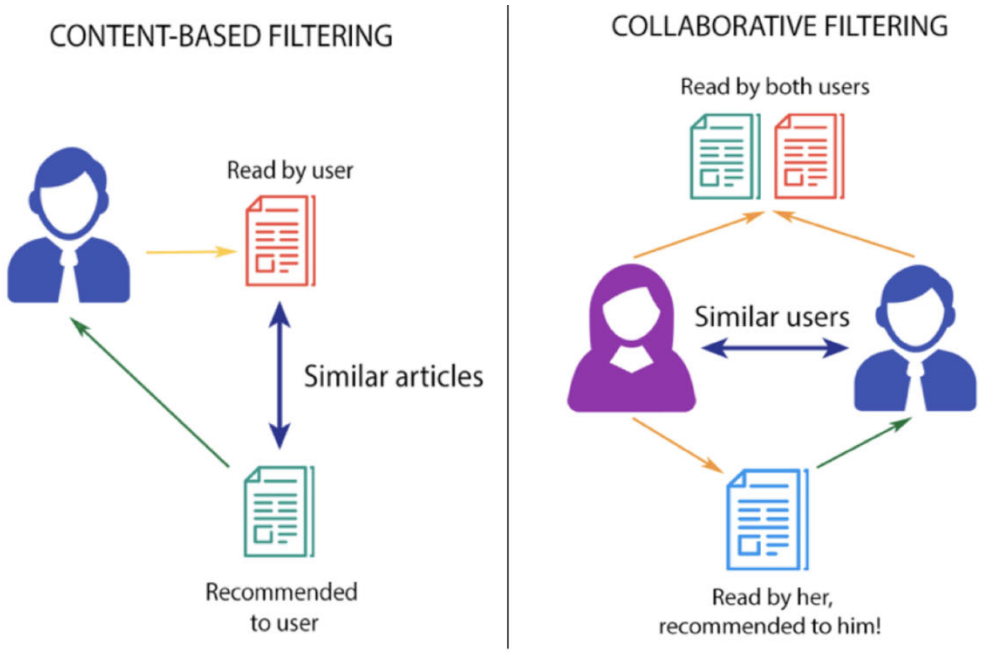

Figure 1 source: https://www.analyticsvidhya.com/blog/2020/11/create-your-own-movie-movie-recommendation-system/

The two main types of NLP reading recommendations models that are generally applicable to the reading list recommendation problem are:

Content-based filters — use item metadata (description, rating, products features, reviews, tags, genres) to find items like those the user has enjoyed in the past.

Collaborative filtering — Collaborative filtering systems analyse users' interactions with the items (e.g. through ratings, likes or clicks) to create the recommendations.

Learning outcomes

Understand and apply new data science skills, knowledge, and techniques to solve problems in data science using natural language processing (NLP).

Work-based skills

The ability to automatically build text datasets and map text hierarchies using NLP is valuable back-office automation opportunity saving time and increasing accuracy for organisations.

Background

Currently schools typically rely on the same publisher for textbooks. When the curriculum changes, then a school typically uses the same publisher for an updated textbook. Changing publishers requires considerable investment by school staff. One difficulty in changing textbooks is keeping appraised on different publishers updated textbooks which may be suitable for a school subject and year level.

In Europe, it is possible to use existing and past reading list information from different University's to inform academics, librarians and publishers about comparative reading that might be applicable to any selected topic. This type of reading recommendation system could be transferred to school textbooks so that a school could be informed of alternative textbooks that are suitable. Such a school textbook recommender would assist teachers choose new textbooks.

Data

The data to be used as the initial starting point for the Tasks is given in the Assessment 1 folder on Learn JCU. A summary of the variables is given in Table 1. The provided data is insufficient to develop NLP recommenders or provide assessments of NLP recommender quality.

Table 1 Data table dictionary

Field ID	Description
School_ID	Unique number for school identification
State	Australian state the school operates in
Year	The year level the textbook is used for
Subject	ASSUMMED subject area the book is used for. MAYBE inaccurate!
ISBN	ISBN of the textbook

Task

1. Generating text-based data for a NLP recommender using the provided data.
    a. **Data Generation:** detail**, d**iscuss and demonstrate:
        i. The use of API's to collect text data
    b. **Data wrangling and Exploratory text data analysis**: detail and discuss
        i. Corpus data wrangling for the intended recommender
        ii. Descriptive statistics of both the sample and the corpus
        iii. Visualisation and interpretation of sample and corpus distributions

2. Develop a **content based** NLP recommendation engine derived on the supplied data to Recommend existing textbooks to similar schools | year level |subject
    a. **Machine Learning for the NLP Recommender System:** detail and discuss
        i. Feature Normalisation appropriate to the intended ML algorithm
        ii. Application of ML techniques for the recommender
        iii. Evaluation of the recommender using quality metrics

The report must show comprehensive thought of your decisions, clearly communicate your ideas, and linked to NLP and machine learning theory/applications with appropriate references.

You will need to include figures, tables, and code sections in your report. Figures, tables and code sections do not contribute to the word count. Code sections must be presented as text, screen capture of code will not be graded. All figures, tables and code must have a caption and be referenced in your discussions using the appropriate captions. Figures, tables, and code sections that are not referenced in discussions do not contribute to grading. Appendices are not graded.

NOTE: The **code sections** in the report must be accomplished using **Python**. Any calculations, visualisations, results and so on produced using software other than Python (e.g. Excel, Tableau, RStudio etc.) are **not** accepted and therefore will not be assessed.

Data Augmentation

The provided data will require supplementation of the dataset from at least one external resource (API).

Some API's that may be used are:

https://developers.google.com/books

Trove API: https://trove.nla.gov.au/about/create-something/using-api/api-technical-guide#examples [note: you need to register to get an API key]

Other API's are allowed.

Assessment submission guidelines

If you use MS Word or any other program, save your work as a PDF for submission.

Your submission for Assessment 1 should be uploaded to LearnJCU as one file.

Your work must meet the following requirements:

- Saved in the following format A1_NLP_Reccomender_firstname_lastname (PDF format)
- Length: 2000 words (+/-30%)
- 12pt font size with 1.5 spacing
- APA referencing style applied.

Upload all submission files in one go. You can upload as many times as you want, but only the last submission is graded.

Marking criteria: MA5851 Assignment 1

Criteria	High Distinction / Distinction: Sophisticated/Exceeds Expectations (75-100%)	Credit /Pass: Above/Meets Expectations (50-74%)	Fail: Unsatisfactory / Below Expectations (0-49%)
Data Generation **20%**	Identifies and discusses and demonstrates: • Multiple API's, or alternative data streams, used to collect text data • Multiple fields from APIs used to augment data • Code to interact with API is utilises sample structure to minimise API calls used to gather the data • Code to interact with API is utilises performance considerations to gather the data Discussions are in a complex data science related situation, drawing upon relevant theory from a wide range of credible sources; eliciting insightful knowledge linking to broader relationships and, bring in originality of perspective	Identifies and discusses and demonstrates: • Single API to collect text data • Single field from API used to augment data • Code to interact with API is sufficient to gather the data Discussions are in a routine data science related situation, drawing upon relevant theory	• API not used collect text data • Code to interact with API is incoherent and/or typically breaches API limitations Partially identifies and/or explains some key issues in a superficial data science related situation
Data wrangling and EDA **30%**	Identifies and discusses: • Corpus data wrangling methods that begin to feature engineer towards the intended NLP tasks • Visualisation and interpretation of sample distribution • Visualisation and interpretation of corpus • Descriptive statistics of both the sample and the corpus • Corpus limitations • Sampling biases • Generation of an appropriate training and test sets with reference to any sample distributions, biases and or data limitations Discussions involve integrating sampling and linguistic theory to elicit insight to the application of NLP data wrangling to improve machine learning tasks. Discussions also draw upon relevant theory from a wide range of credible sources.	Identifies and discusses: • Standard corpus data wrangling methodologies appropriate for basic NLP tasks • Visualisation and interpretation of sample distribution • Visualisation and interpretation of corpus, • Descriptive statistics of both the sample and the corpus Discussions are generally descriptive and identify most key criteria.	• Corpus data wrangling not appropriate to the intended recommender(s) Partially identifies and/or explains some key issues in a superficial data science related situation

Criteria	High Distinction / Distinction: Sophisticated/Exceeds Expectations (75-100%)	Credit /Pass: Above/Meets Expectations (50-74%)	Fail: Unsatisfactory / Below Expectations (0-49%)
NLP Recommender System **40% of total grade**	Discussion including: • Feature Normalisation appropriate to the intended ML algorithm • ML techniques appropriately applied with hyperparameters identified, discussed with some parameters optimised using appropriate optimisation methodologies • Quality metrics calculated and interpreted with reference to some/all: sample distributions, data distributions/limitations, applied ML algorithm limitations/properties • Example use demonstrated, including examples pertaining to data/ML limitations Discussions involve integrating sampling and linguistic theory to elicit insight to the application of NLP machine learning tasks. Discussions also draw upon relevant theory from a wide range of credible sources.	Discussion including: • Feature Normalisation appropriate to the intended ML algorithm • ML techniques appropriately applied. • Quality metrics calculated and interpreted for both the recommender. Discussions are generally descriptive and identify key criteria.	• Feature Normalisation inappropriate to the intended ML or not evident • ML technique inappropriately applied • Quality metrics calculated but not interpreted or quality metrics discussion seriously flawed or irrelevant Partially identifies and/or explains some key issues in a superficial data science related situation
Report structure **10% of total grade**	• Sequencing of sections logical and coherent. No out of sequence material or discussions. • Output results, code, figures appear in the sections where initially discussed • Grammar and spelling errors are rare • Internal cross referencing always used • External referencing style appropriate	• Sequencing of sections logical and coherent. Some out of sequencing of content. • Output results, code, figures appear in the sections where initially discussed • Grammar and spelling contain some errors • Internal cross referencing sometimes used • External referencing style appropriate	• Sequencing of sections routinely illogical and/or incoherent, frequent out of sequencing of content. • Output results, code, figures routinely do not appear in the sections where initially discussed • Grammar and spelling contain frequent errors • Internal cross referencing rarely/not used • External referencing style inappropriate