Assessment 2: NLP Recommendation Engine

Name: NLP Recommendation Engine
Type: Written document and code submission
Issued: 8:00 PM AEST Monday of Week 1
Due: 11:59 PM AEST Sunday of Week 4
Weight: 40%
Length: 3000 words (+/-10%)

Overview

Natural language processing (NLP) is commonly used to build recommendation engines. This assignment involves building reading recommendation engines for students in Australian higher education based on a sample of subject reading lists sourced from public sites.

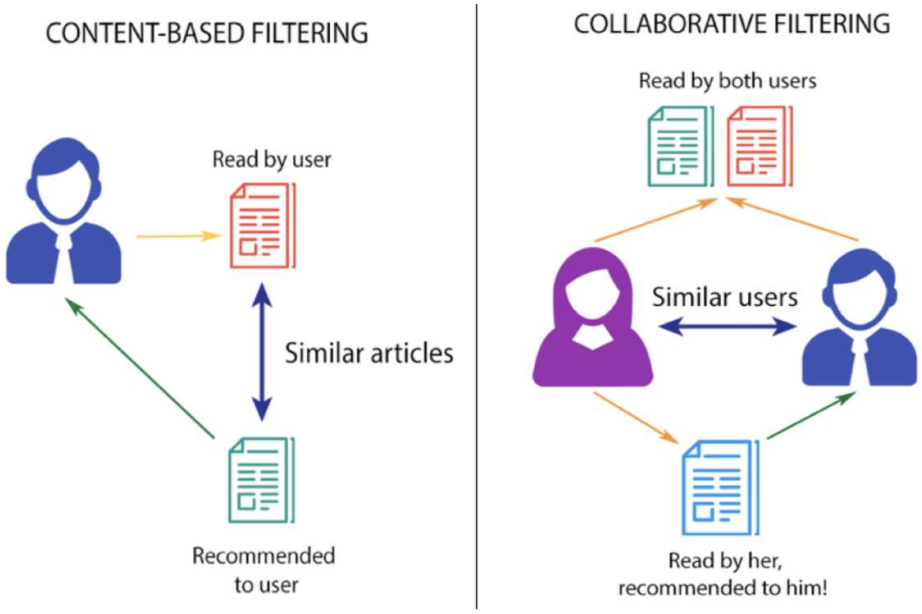

Figure 1 source: https://www.analyticsvidhya.com/blog/2020/11/create-your-own-movie-movie-recommendation-system/

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 labelled datasets and map text hierarchies using NLP is valuable back-office automation opportunity saving time and increasing accuracy for organisations.

Background

Currently University lectures rely on librarians to do background research on relevant, compliant, and current reading material suitable for course teaching units.

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 is not possible yet in Australia as no up to date central repository exists for reading list material.

The ability to recommend reading material down to a book and page level and at a journal level would be a useful product for academics, publishers and agencies supporting higher education.

For education providers a recommendation engine will :

- Be simpler and more intuitive to use
- Reduce research time in material selection
- Resolve current process deficiencies
- Improve data availability.

For students it will:

- Make material more relevant and contemporary
- Improve engagement through usage monitoring
- Help students complete their studies.

For publishers and software companies it will allow:

- Insights based on which of their titles are used, where and how much.
- More focused selling of course eBooks to institutions for their student
- Better Integration with learning management systems

Tasks

1. Develop two NLP recommendation engines derived on the University's reading list material (supplied data) and apply the NLP recommenders to one of the following options:
    a. Recommend existing course material to similar subjects, or
    b. Recommend new reading material to existing subjects, or
    c. Provide a complete reading list of existing readings for a new subject.

2. Determine the quality of both NLP recommenders from Task 1 using test and training sets derived from the supplied data.

3. Compare the two NLP recommenders.

For each of the three tasks, a written report is to be provided. The report must show comprehensive thought of your decisions, clearly communicate your ideas, and linked to NLP theory/applications with appropriate references.

Data

The data to be used as the initial starting point for the Tasks is given in the Assessment 2 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 element Dictionaries

Field ID	Description
ID	University ID
COURSENAME	Name of Course
ITEM_COUNT	Number of items in reading list
TITLE	Major Title (book, journal)
RESOURCE_TYPE	Book Journals
SUBTITLE	Minor Title (article)
ISBN10S	Universal Identifiers
ISBN13S	Universal Identifiers
ISSNS	Universal Identifiers
EISSNS	Universal Identifiers
DOI	Digital Object Identifier
EDITION	Edition of Publication
EDITORS	Names of Editors
PUBLISHER	Publisher
DATES	Publication Date
VOLUME	
PAGE_END	Pages selected
AUTHORS	Authors

The provided data will require:

- Creation of a labelled dataset from the source data (ontology), and
- Supplementation of the dataset from at least one external resource (API)

OpenRefine or Python are be used to supplement the dataset with external resources. Some API's that may be applied are:

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]

OCLC WorldCat: http://classify.oclc.org/classify2/ClassifyDemo?search-standnum-txt=9780199022274&startRec=0 & http://classify.oclc.org/classify2/Classify?isbn=9781863955799&detail=true

Other API's are allowed.

Another suitable resource is the Australian Standard Classification of Education (ASCED). It is a statistical classification for use in the collection and analysis of data on educational activity and attainment. ASCED comprises two classifications: Level of Education and Field of Education. https://www.abs.gov.au/AUSSTATS/Abs@.Nsf/Latestproducts/3AB4B1E1404D7C91CA256AAF001FCA54?opendocument

Both the level and field of education components can be used from here: https://heimshelp.dese.gov.au/resources/field-of-education-types and relevant data can be used to map existing courses of the dataset to respective field of study and thereby a meta-data can be added to the input dataset.

Assessment submission guidelines

Submission method options: Pdf submission to LearnJCU with supporting Python file (.py or .ipynb) and data file input/output from mapping using NLP.

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

Your submission for Assessment 2 should be uploaded to LearnJCU as two (2) separate files:

1. Your work must meet the following requirements:
    - Saved in the following format A2_NLP_Reccomender_firstname_lastname (PDF format)
    - Length: 3000 words (+/-10%)
    - 12pt font size with 1.5 spacing
    - APA referencing style applied.

2. Supplied source data file(s), Python Notebook (.ipynb), Python scripts (.py) or OpenRefine GREL code (text file) with the information about the version of Python/OpenRefine that you have used and any associated package used.

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

IMPORTANT NOTE

The **entire project** must be accomplished using **Python and/with OpenRefine**. Any calculations, visualisations, results and so on produced using software other than Python or OpenRefine (e.g. Excel, Tableau, etc.) are **not** accepted and therefore will not be assessed. The code itself must be prepared using **Python either as a script in notebook form**. Noncompliance will result in your work being considered as **not delivered**.

Marking criteria: MA 5851 Assignment 2 NLP Recommendation Engine (30% of total grade)

Criteria task 1 (Total: 80 marks)	Exemplary (100% marks)	Good (75% marks)	Satisfactory (50% marks)	Limited (25% marks)	Poor (12.5% marks)	Incomplete (0% marks)
Proposes strategies or partial solutions (15 marks)	The solution provided is correct and solves the problem effectively.	Exhibits aspects of columns to the left and right.	The solution provided is mostly correct and partially solves the problem.	Exhibits aspects of columns to the left and right.	The solution provided is incorrect and does not solve the problem.	No solution or Python notebook or output provided.
Effective and correct use of functions in the solution (15 marks)	The functions identified are appropriate and correctly used.	Exhibits aspects of columns to the left and right.	The functions identified are mostly appropriate and correctly used.	Exhibits aspects of columns to the left and right.	The functions identified are inappropriate or incorrectly used.	No solution or Python notebook or output provided.
Formal quantitative assessment of results and output (20 marks)	The write up of results and output solves the problem, and is consistent with the explanation of the algorithm used.	Exhibits aspects of columns to the left and right.	The write up of results and output mostly solves the problem, and is consistent with the explanation of the algorithm used.	Exhibits aspects of columns to the left and right.	The write up of results and output does not solve the problem, and is inconsistent with the explanation of the algorithm used.	No solution or Python notebook or output provided.
Effective and correct use of text mining package in the solution (15 marks)	The text mining elements used are appropriate and correctly used.	Exhibits aspects of columns to the left and right.	The text mining elements used are mostly appropriate and correctly used.	Exhibits aspects of columns to the left and right.	The text mining elements used are mostly inappropriate or incorrectly used.	No solution or Python notebook or output provided.
Effective and correct use of dictionaries and vocabularies and loops in the solution (15 marks)	Labelled datasets, dictionaries and vocabularies are appropriate and correctly used.	Exhibits aspects of columns to the left and right.	Labelled datasets, dictionaries and vocabularies are mostly appropriate and correctly used.	Exhibits aspects of columns to the left and right.	Labelled datasets, dictionaries and vocabularies are inappropriate or incorrectly used.	No solution or Python notebook or output provided.

CONTINUED NEXT PAGE

Criteria Task 2 and 3 (Total: 20 marks)	Exemplary (100% marks)	Good (75% marks)	Satisfactory (50% marks)	Limited (25% marks)	Poor (12.5% marks)	Incomplete (0% marks)
At least two NLP approaches contrasted with final model selection based on model performance **(10 marks)**	The write up includes at least two approaches, is clear and concise with a justifiable recommendation of final model selection, with appropriate metadata enhancements and outputs	Exhibits aspects of columns to the left and right.	The write up includes at least two approaches, is reasonably clear and concise with a justifiable recommendation of final model selection with some appropriate metadata enhancements and outputs	Exhibits aspects of columns to the left and right.	The write up includes at least two approaches, is not very clear or concise with unclear reasoning with regards to the recommendation of a final model selection. No appropriate metadata enhancements and outputs	The write up does not include a recommendation of a final model selection and no levels of the hierarchy mapped to the training data.
Formal quantitative assessment of results and output **(10 marks)**	The write up of results and output solves the problem and is consistent with the explanation of the algorithm used.	Exhibits aspects of columns to the left and right.	The write up of results and output mostly solves the problem, and is consistent with the explanation of the algorithm used.	Exhibits aspects of columns to the left and right.	The write up of results and output does not solve the problem, and is inconsistent with the explanation of the algorithm used.	The write up does not include a recommendation of a final model selection and no levels of the hierarchy mapped to the training data.