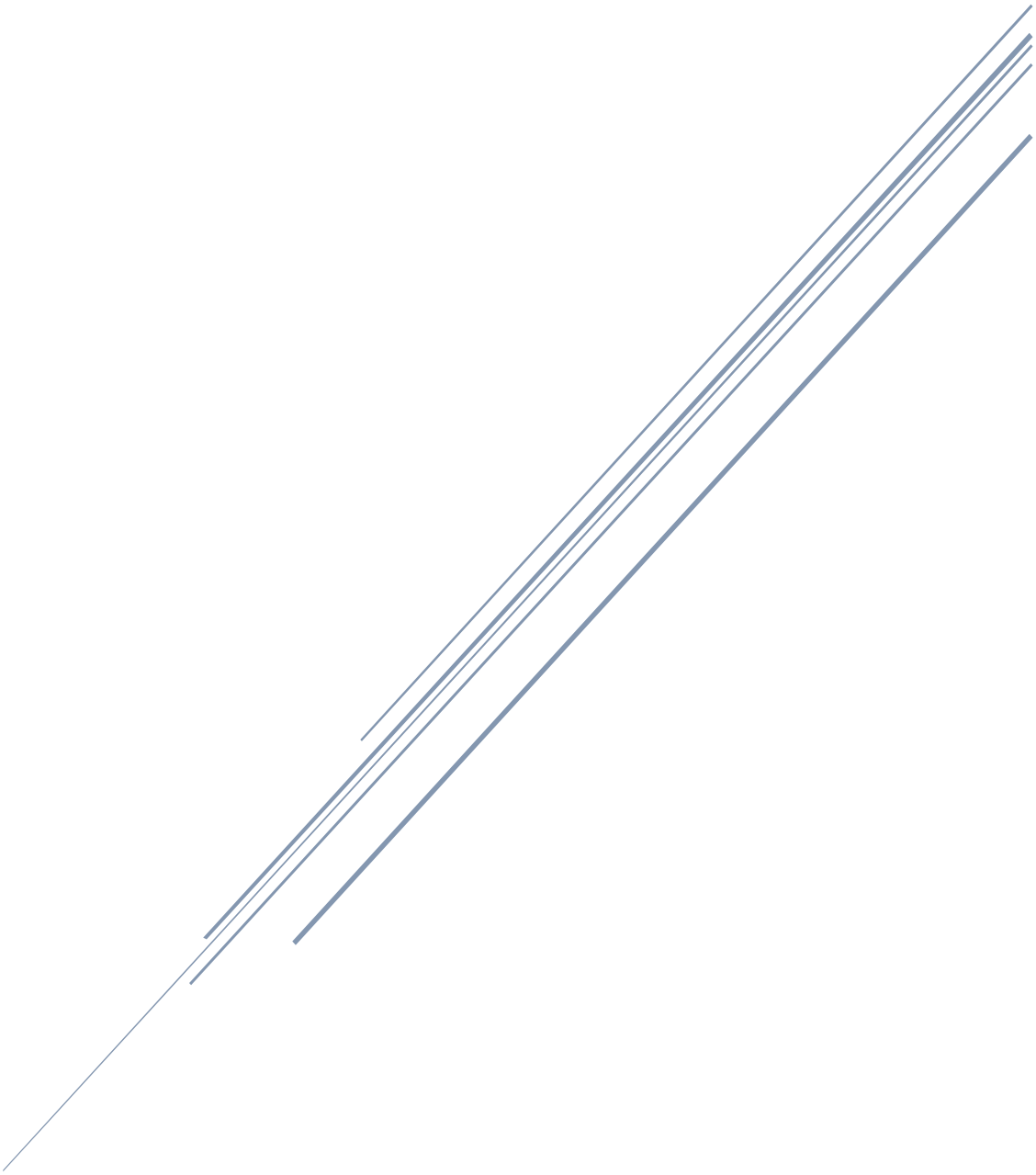
NLP RECOMMENDATION SYSTEMS

Assessment 2



College of Science and Engineering
Data Science Master Class 1

Contents

Introduction

Recommender systems are a set of algorithms that suggest a particular product, service or item to a person or entity based on what the system knows about the person and/or the products in its repository (Ricci et al., 2011). Such systems can be divided into two main types: content-based and collaborative-based (Banik, 2018). Content-based recommenders use the attributes of products and services as a basis for comparison and make recommendations whereas collaborative-based recommendations use information about product-user interactions to produce their recommendations. As a result, one of the biggest difficulties with collaborative-based systems is that they tend to suffer from the "cold-start problem". That is, they are unable to generate new recommendations if there is insufficient data regarding past interactions with users and/or items (Aggarwal, 2016). Other types of recommender systems include knowledge-based systems, demographic-based systems and hybrid systems that combine aspects of one or more of the above.

In Europe, existing and historic reading material from different universities is located in a central repository that can be used by academics, librarians and publishers to inform them as to (for example) what items are being used, where they are being used, and how popular they are. However, in Australia no such central repository exists and academics and publishers must perform their own research to obtain that sort of information.

The aim of this research, therefore, was to develop two recommender systems that could be used by academics as a starting point for research into new subject material for their courses. In other words, the recommender systems would be able to suggest new material for a given course based on item or course similarity. The second half of the research involved an assessment of the quality of each recommender using training and test sets derived from the underlying dataset and a comparative analysis.

Data and Methods

Initial Data Cleaning

The dataset supplied as a starting point for this research consisted of 68,530 rows and 19 columns, including the university id, course name, title of the reading, relevant page

numbers and a number of other pieces of descriptive metadata (as per Table 1 below). However, it quickly became apparent that the dataset had to undergo significant cleaning before it could be used as input into the development of a recommender system. For example, there was inconsistent use of capitalisation and abbreviations across many of the fields, invalid values in the universal identifier fields and a significant amount of misaligned data (that is, data that was sitting in the incorrect field). This initial cleaning was performed using OpenRefine, a tool designed to clean, process and augment messy datasets (OpenRefine, n.d.).

Table 1: Overview of the contents of the supplied dataset

Field	Description	Number of Null Values	Number of Unique Values
ID	University ID	0	4
COURSENAME	Name of course	0	3,657
ITEM_COUNT	Number of items in reading list	9,274	139
TITLE	Major title	46	44,905
RESOURCE_TYPE	Item type	50	6,178
SUBTITLE	Minor title	46	45,611
ISBN10S	Universal identifier	55,048	9,371
ISBN13S	Universal identifier	49,175	9,377
ISSNS	Universal identifier	46,515	12,520
EISSNS	Universal identifier	63,042	2,254
DOI	Digital object identifier	63,666	3,974
EDITION	Publication edition	51,469	1,423
EDITORS	Editor name(s)	1,613	5,580
PUBLISHER	Publisher	35,965	10,158
DATES	Publication date	21,778	808
VOLUME	Volume	56,288	3,158
PAGE_END	Page numbers	54,175	1,374
AUTHORS	Page numbers	53,635	1,582
Column 19	Author name(s)	59,256	5,589

API Data Supplementation

The second stage of the data preparation process involved supplementing the dataset with information from at least one external resource through the use of an API. This would serve a number of purposes; to verify the metadata information in the existing dataset, to correct or update it if necessary, to fill in missing values where possible[1], and add additional useful descriptive metadata such as the item's language, topic or subject area tags and description.

[1] It was acknowledged that it was possible many of the null values were actually valid as the editor, volume and edition fields would not be applicable for all items, and certain identifiers only applied to certain item formats.

The sources chosen for this task were OCLC Classify and Google Books. Trove was considered but ultimately discarded as it only contained information relevant to Australia and would therefore not be useful for any items written or published internationally (Trove, n.d.). Furthermore, two APIs were chosen as OCLC Classify held information about a wider variety of item formats, but a more limited set of metadata, whereas Google Books returned a wider range of metadata, but for a more limited number of item formats.

During the data cleaning phase, effort was concentrated on cleaning the identifier[2] and title fields as these were going to be the primary fields that would be used for the API calls. For example, any values that could not be coerced into a valid representation for any of the identifier fields were disregarded and then the entire set of identifier fields were consolidated into a single field, multi-valued cells were split into separate rows (which temporarily increased the size of the dataset to 77,139 rows) and an extra field created to classify each identifier.

The API calls were done in stages by filtering the dataset on the identifier classification field and the throttle delay reduced to 500 seconds to maximise the speed of each call without being blocked. A secondary call was also made to each API using the title field for any items for which the previous calls using the identifier field had been unsuccessful.

For results containing multiple works, the first work in the list was chosen as it was impossible to manually verify which result (if any) was correct given the size of the dataset. The title fields from each of the API calls were then compared to the initial title field to determine the amount of descriptive metadata that could be confidently supplemented without risking a change to the reading. As a result, 24,688 rows were classified fully verified and used all of the metadata returned from the API calls, and a further 26,069 rows had metadata supplemented by one of the API calls (in total approximately 65.8 per cent of the dataset).

[2] ISBN10S, ISBN13S, ISSNS, EISSNS and DOI

The dataset then underwent a second cleaning to discard unwanted fields, re-join rows belong to same record and remove duplicate rows before it was exported as a csv file. Table 2 below presents an overview.

Table 2: Overview of the contents of the cleansed and supplemented dataset

Field	Description	Number of Null Values	Number of Unique Values
ID	University ID	0	4
COURSE_NAME	Name of course	0	3,452
TITLE_STATUS	API result verification classification	0	9
TITLE	Title	45	43,402
FORMAT	Resource type	49	54
AUTHOR	Author	32,722	19,162
PUB_DATE	Publication date	15,101	4,318
PUBLISHER	Publisher	34,136	6,706
IDENTIFIER	Universal identifier/DOI	21,678	28,673
TOPIC	Topic(s)	32,838	7,351
LANGUAGE	Language	40,744	14
DESCRIPTION	Description/summary	45,432	11,690
PAGES	Relevant pages	47,650	7,798
AVG_RATING	Average rating	58,047	9
NUM_RATINGS	Number of ratings	58,047	100

Ontology Development

From the third stage onwards, all the analysis was performed with python[3]. This particular stage required the creation of a labelled dataset and was achieved through the application of a k-means clustering algorithm on the course names field. K-means is an unsupervised clustering algorithm that attempts to group similar data points together by minimising the sum of the distances between the points and their respective cluster centroid and thereby reveal underlying patterns in the data (Scikit-learn: Machine learning in python, n.d.-a):

$$\sum_{i=0}^{n} \min_{\mu_j \epsilon C} \left(\left\| x_i - \mu_j \right\|^2 \right) \quad (1)$$

In order to be able to use the algorithm, the course names had to first be transformed into a feature set that it would recognise as valid input. This was achieved in two steps. The first step involved extracting a list of all unique course names, pre-processing them by removing punctuation and stop words, tokenising each name into individual words and then applying a part-of-speech tagger to remove all non-noun words. The second step involved using a

[3] Copies of the python scripts for this and the following sections are in Appendices F through I, and the OpenRefine operation history for the first two stages is in Appendix E.

term frequency-inverse document frequency (TF-IDF) vectoriser on the pre-processed course names to obtain a weighted matrix of tokenised names.

The k-Means algorithm requires the number of clusters to be pre-specified, and this was chosen based on the number of narrow fields in the Australian Standard Classification of Education. The results were assessed using within-cluster-sum-of-squares and the silhouette coefficient to help determine an optimal number of clusters.

The clusters were labelled according to the top three words listed and then assigned back to the dataset as each course name's field of education.

Content-Based Recommender System

In content-based recommender systems, the descriptive attributes of items are used as the basis for recommendations. For this dataset, the users were the different university courses and the products that were going to be compared were the individual readings. Therefore, the feature set for the content-based recommender was made up of the contents of the title, author, publisher, format, language and topic fields.

The first requirement was to convert the above-named feature set into a keyword-based vector-space representation. This was achieved using an almost identical process to that used in the ontology creation; that is, the selected features were converted to lower-case, punctuation, numbers, single letters, additional white space and stop words were removed, word tokenisation was applied, the words were lemmatised and stemmed and finally combined back together into a single string before being fed into a TF-IDF vectoriser.

Given the size of the resulting feature matrix and the limitations of the hardware on which this system was being developed, it was then necessary to apply singular value decomposition (SVD) to the matrix to reduce its dimensionality. The number of latent features was selected based on a compromise between the cumulative amount of variance explained and hardware memory limitations.

The latent feature matrix was then used to build individual user profile against all the titles in the training set using cosine similarity, which computes the L2-normalised dot product of two vectors (Scikit-learn: Machine learning in python, n.d.-b):

$$k(x, y) = \frac{xy^T}{\|x\|\|y\|} \quad (2)$$

The result was an array of similarity scores that ranged between zero and one (where one indicated an exact match and zero indicated no similarity) that was then joined back to the training set to produce a set of title recommendations for that user. Due to the computation time required, this was repeated for a random sample of 900 users and the overall performance evaluated on the same set of users on a separate test set.

Collaborative-Based Recommender System

In contrast, collaborative-based recommender systems harness the power of user ratings and/or information about user-item interactions to make recommendations. Given the sparsity of the pre-existing ratings data, it was necessary to construct a set of synthetic ratings based on the popularity of each title for each university. That is, the number of times each reading was used was converted to a scale out of five so these additional ratings could be integrated in with the pre-existing data. Furthermore, in order to maximise the effectiveness of the algorithm is discovering similarity between items and/or users, the titles and course names were pre-processed into bag-of-words feature strings as they had been for the content-based recommender.

A ratings matrix of courses against titles was constructed and missing values filled with zeros to indicate no rating for that particular user-item combination. Similar to the approach taken for the content-based recommender, this user-item matrix was then subjected to a SVD transformation and the cosine similarity of all users against the reference user calculated.

(Recall that SVD is a form of matrix factorisation in which a matrix (of dimensions $m \times n$ and rank k) is decomposed down into three sub-matrices that approximates the full vector space with a smaller set of latent features according to the following equation:

$$M \approx U_{k'} \Sigma_{k'} I_{k'}^{T} \quad (3)$$

where the user and item factor matrices are constrained to be mutually orthogonal (Aggarwal, 2016; Hadian, 2019)).

The title recommendations were then extracted based on the similarity scores of the individual courses.

The process was repeated for all users in the training set and the trained SVD and user-product matrix applied to a separate test set to evaluate the performance of the recommender on an unseen set of data.

Quality and Performance Assessment

A robust assessment of the quality and performance of each recommender system required a test on a set of unseen data points. Therefore, prior to the training of each model, the dataset was split into a set of training and test datasets that were stratified by course id (course name plus university id). It must be noted that different train-test splits were applied for each recommender due to the different data input requirements and computational complexity of the respective algorithms.

The performance of each recommender system was evaluated using precision@k, recall@k and F1@K, which are widely used metrics in information retrieval scenarios, as well as the average similarity of items in the top k recommendations.

1. **Precision** is defined as the number of recommended items that the user has previously interacted with (in this case the number of recommended readings already used by a particular course), divided by the total number of recommended items:

$$Precision@k = \frac{Number\ of\ items\ @k\ the\ user\ already\ uses\ in\ their\ course}{Number\ of\ recommended\ items\ (k)} \qquad (4)$$

2. **Recall** is defined as the number of recommended items already used by a user in their course, divided by the total number of items used in that course:

$$Recall@k = \frac{Number\ of\ items\ @k\ the\ user\ already\ uses\ in\ their\ course}{Total\ number\ of\ items\ used\ in\ their\ course} \qquad (5)$$

3. **F1** combines precision and recall into a single measure. As with the previous two metrics, values range from zero to one, with zero indicating a poor score and one indicating a perfect score.

$$F1@k = \frac{precision@k \times recall@k}{precision@k + recall@k} \qquad (6)$$

These metrics were calculated for each user in the respective datasets on the top 20 recommendations and the mean results across the entire training and test sets are summarised in Table 3 below. Note that these were the order-unaware versions; that is, the calculations were made on the basis of the presence or absence of a particular reading irrespective of where it was listed in the set of recommended items):

Table 3 - Mean performance metric scores (per cent) at k = 20

	Content recommender	**Collaborative recommender**
	n = 900	n = 1,188
Training data		
Precision	2.88	0.14
Recall	8.09	0.23
F1	9.73	5.27
Similarity score	46.28	60.91
Number of 'correct' recommendations	0.58	0.03
Test data		
Precision	2.01	0.11
Recall	19.98	0.73
F1	10.69	7.97
Similarity score	38.68	41.87
Number of 'correct' recommendations	0.40	0.02

Whilst the precision, recall and F1 metrics indicated that the quality of the recommendations was very poor, visual inspection of the recommended titles for a selection of users indicated that these were potentially misleading and should not be relied on as the sole source of information for any conclusion regarding recommendation quality. This is particularly pertinent given that the aim of the recommender systems was to suggest new readings that could be used to supplement pre-existing course readings lists and not repeat items already used.

Comparative Analysis and Model Choice

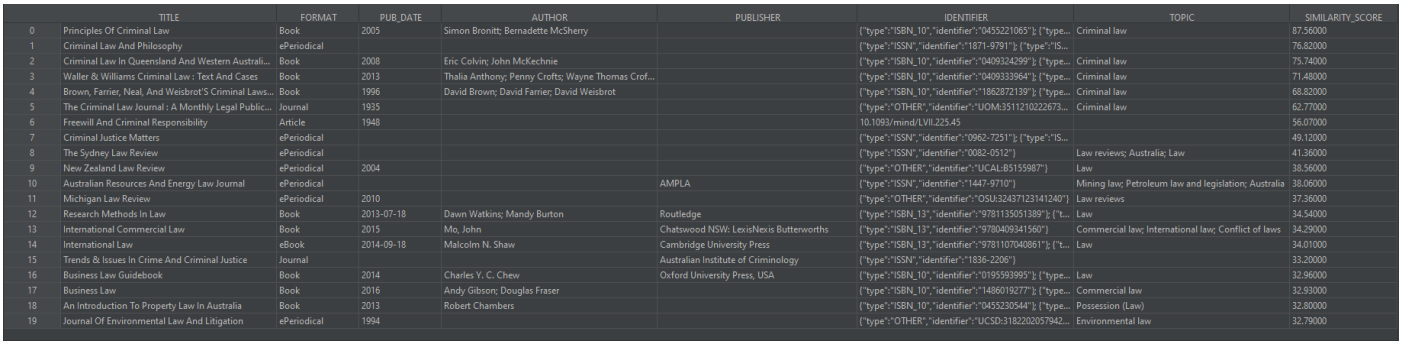
As was observed in the previous section, the quality and performance of the two recommender systems developed for this research project was very similar. However, it was observed that the average similarity scores of the top 20 recommendations produced by the collaborative-based model were higher than those produced by the content-based model, whereas the content-based model performed better according to the average precision, recall and F1 metrics.

In terms of the computational complexity and pre-processing requirements of each model, the collaborative-based recommender was quicker to train (1.3 hours compared to approximately 10.8 hours for a similar sized dataset) and required less data pre-processing prior to model training. That being said, the collaborative recommender required the construction of a set of synthetic ratings based on within-university item popularity due to the due to the lack of pre-existing ratings data (a common problem for collaborative-based systems as highlighted earlier).

A comparison of the training and test set recommendations for new reading material for the same user are presented in Figure 1 (content-based recommender) and Figure 2 below (collaborative-based recommender). In the first figure, the suggestions appear to be reasonable and relevant in for both datasets. This was not the case in the second figure. The results appeared to be reasonable for the training data, but it was far more difficult to see the connection between the input and the collaborative system's recommendations in the test set. Furthermore, this appeared to be quite common as the author was unable to find an example after several attempts in which the collaborative system produced a set of expected results.

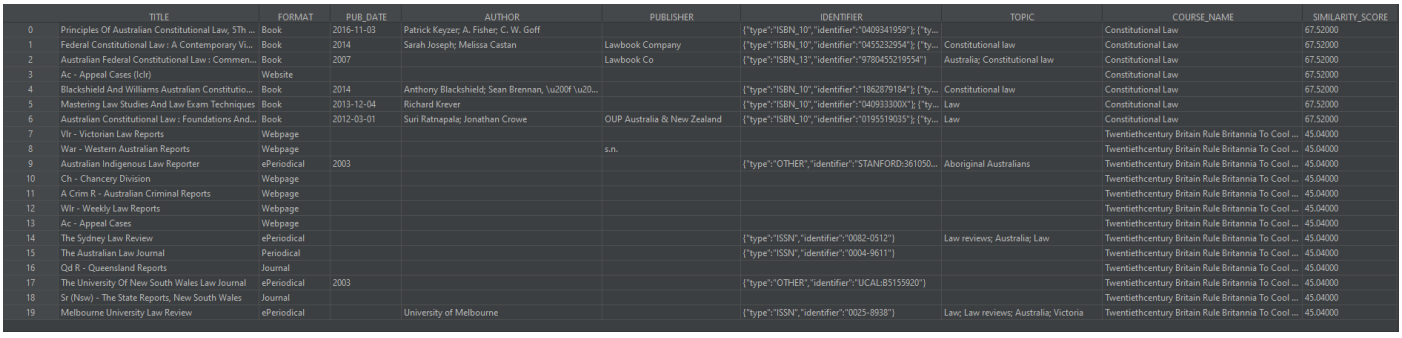
Figure 1 – Examples of the recommendations from the content-based system with the input (“Criminal Law”)

	TITLE	FORMAT	PUB_DATE	AUTHOR	PUBLISHER	IDENTIFIER	TOPIC	SIMILARITY_SCORE
0	The Criminal Law Review	Journal	2000			{"type":"OTHER","identifier":"STANFORD:36105062...	Criminal law	94.57000
1	Criminal Law In Australia	Book	2014-12-11	Lorraine Finlay; Tyrone Kirchengast		{"type":"ISBN_10","identifier":"0409338729"}; {"type...	Criminal law	92.51000
2	Principles Of Criminal Law	Book	2005	Simon Bronitt; Bernadette McSherry		{"type":"ISBN_10","identifier":"0455221065"}; {"type...	Criminal law	87.62000
3	Criminal Law And Philosophy	ePeriodical				{"type":"ISSN","identifier":"1871-9791"}; {"type":"IS...		76.82000
4	Criminal Law In Queensland And Western Australia...	Book	1998	Eric Colvin; Suzie Linden-Laufer; Leanne Bunney		{"type":"ISBN_10","identifier":"040931384X"}; {"type...	Criminal law	76.35000
5	Criminal Law And Procedure In New South Wales	Book	2009	Robert Alexander Hayes; Michael Eburn		{"type":"ISBN_10","identifier":"0409325562"}; {"type...	Criminal law	73.44000
6	Criminalisation And Criminal Responsibility In Aust...	Book	2015-05-20	Thomas Crofts; Arlie Loughnan	Oxford University Press, USA	{"type":"ISBN_10","identifier":"0195597567"}; {"type...	Criminal law	70.53000
7	Brown, Farrier, Neal And Weisbrot'S Criminal Laws ...	Book	2011	David Brown; David Farrier; Sandra Egger; Luke Mc...		{"type":"ISBN_10","identifier":"186287817X"}; {"type...	Criminal law	69.40000
8	Practitioner'S Guide To Criminal Law	Book						68.63000
9	Brown, Farrier, Neal, And Weisbrot'S Criminal Laws...	Book	1996	David Brown; David Farrier; David Weisbrot		{"type":"ISBN_10","identifier":"1862872139"}; {"type...	Criminal law	68.08000
10	Criminal Law Guidebook : New South Wales, Victor...	Book	2016-10-13	John Anderson	Oxford University Press, USA	{"type":"ISBN_10","identifier":"0195593987"}; {"type...	Criminal law	65.11000
11	The Criminal Law Journal : A Monthly Legal Public...	Journal	1935			{"type":"OTHER","identifier":"UOM:35112102226 73...	Criminal law	62.71000
12	Current Issues In Criminal Justice	ePeriodical				{"type":"ISSN","identifier":"1034-5329"}	Australia; Criminology; Criminals; Criminal justice, ...	62.00000
13	Freewill And Criminal Responsibility	Article	1948			10.1093/mind/LVII.225.45		56.04000
14	Crimes Act 1958	Webpage	2003	Victoria		{"type":"ISBN_10","identifier":"1741321204"}; {"type...	Criminal law	55.95000
15	Trends And Issues In Crime And Criminal Justice	ePeriodical				{"type":"ISSN","identifier":"0817-8542"}	Australia; Crime; Criminal justice, Administration o...	55.70000
16	A Crim R - Australian Criminal Reports	Webpage						53.12000
17	Journal Of Contemporary Criminal Justice	ePeriodical				{"type":"ISSN","identifier":"1043-9862"}; {"type":"IS...	Criminal justice, Administration of; United States	51.42000
18	Waller & Williams Criminal Law : Text And Cases	Book	2016	Crofts, Thomas (Wayne Thomas)	Gray, Stephen, 1...	Chatswood, N.S.W.: LexisNexis Butterworths	{"type":"ISBN_13","identifier":"9780409343885"}	Australia; Criminal law	49.29000
19	Criminal Justice Matters	ePeriodical				{"type":"ISSN","identifier":"0962-7251"}; {"type":"IS...		48.94000

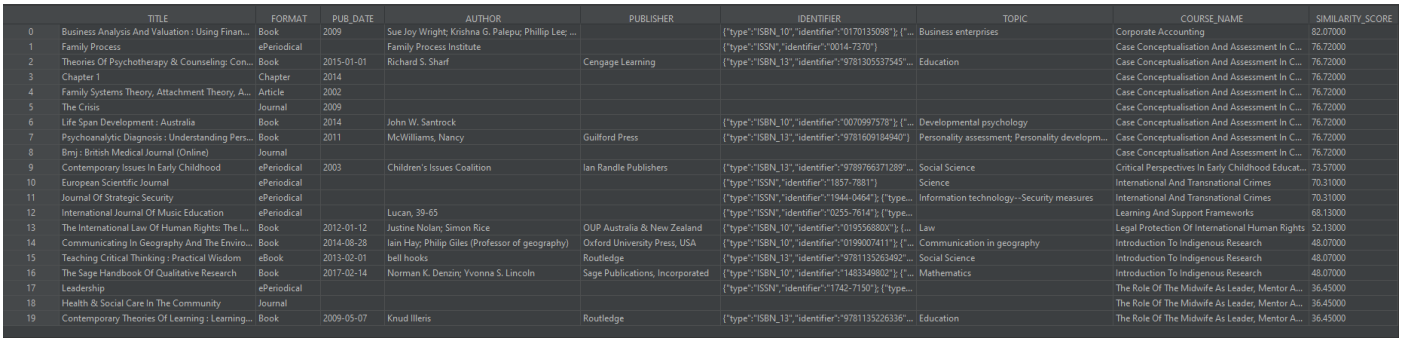


	TITLE	FORMAT	PUB_DATE	AUTHOR	PUBLISHER	IDENTIFIER	TOPIC	SIMILARITY_SCORE
0	Principles Of Criminal Law	Book	2005	Simon Bronitt; Bernadette McSherry		{"type":"ISBN_10","identifier":"0455221065"}; {"type...	Criminal law	87.56000
1	Criminal Law And Philosophy	ePeriodical				{"type":"ISSN","identifier":"1871-9791"}; {"type":"IS...		76.82000
2	Criminal Law In Queensland And Western Australi...	Book	2008	Eric Colvin; John McKechnie		{"type":"ISBN_10","identifier":"0409342499"}; {"type...	Criminal law	75.74000
3	Waller & Williams Criminal Law : Text And Cases	Book	2013	Thalia Anthony; Penny Crofts; Wayne Thomas Crof...		{"type":"ISBN_10","identifier":"0409333964"}; {"type...	Criminal law	71.48000
4	Brown, Farrier, Neal, And Weisbrot'S Criminal Laws...	Book	1996	David Brown; David Farrier; David Weisbrot		{"type":"ISBN_10","identifier":"1862872139"}; {"type...	Criminal law	68.82000
5	The Criminal Law Journal : A Monthly Legal Public...	Journal	1935			{"type":"OTHER","identifier":"UOM:35112102226 73...	Criminal law	62.77000
6	Freewill And Criminal Responsibility	Article	1948			10.1093/mind/LVII.225.45		56.07000
7	Criminal Justice Matters	ePeriodical				{"type":"ISSN","identifier":"0962-7251"}; {"type":"IS...		49.12000
8	The Sydney Law Review	ePeriodical				{"type":"ISSN","identifier":"0082-0512"}	Law reviews; Australia; Law	41.36000
9	New Zealand Law Review	ePeriodical	2004			{"type":"OTHER","identifier":"UCAL:B5155987"}	Law	38.56000
10	Australian Resources And Energy Law Journal	ePeriodical			AMPLA	{"type":"ISSN","identifier":"1447-9710"}	Mining law; Petroleum law and legislation; Australia	38.06000
11	Michigan Law Review	ePeriodical	2010			{"type":"OTHER","identifier":"OSU:324371231414 240"}	Law reviews	37.36000
12	Research Methods In Law	Book	2013-07-18	Dawn Watkins; Mandy Burton	Routledge	{"type":"ISBN_13","identifier":"9781135051389"}; {"t...	Law	34.54000
13	International Commercial Law	Book	2015	Mo, John	Chatswood NSW: LexisNexis Butterworths	{"type":"ISBN_13","identifier":"9780409341560"}	Commercial law; International law; Conflict of laws	34.29000
14	International Law	eBook	2014-09-18	Malcolm N. Shaw	Cambridge University Press	{"type":"ISBN_13","identifier":"9781107040861"}; {"t...	Law	34.01000
15	Trends & Issues In Crime And Criminal Justice	Journal			Australian Institute of Criminology	{"type":"ISSN","identifier":"1836-2206"}		33.20000
16	Business Law Guidebook	Book	2014	Charles Y. C. Chew	Oxford University Press, USA	{"type":"ISBN_10","identifier":"0195593995"}; {"type...	Law	32.96000
17	Business Law	Book	2016	Andy Gibson; Douglas Fraser		{"type":"ISBN_10","identifier":"1486019277"}; {"type...	Commercial law	32.93000
18	An Introduction To Property Law In Australia	Book	2013	Robert Chambers		{"type":"ISBN_10","identifier":"0455230544"}; {"type...	Possession (Law)	32.80000
19	Journal Of Environmental Law And Litigation	ePeriodical	1994			{"type":"OTHER","identifier":"UCSD:31822020579 42...	Environmental law	32.79000

Figure 2 - Examples of the recommendations from the collaborative-based system with the input (“Criminal Law”)



	TITLE	FORMAT	PUB_DATE	AUTHOR	PUBLISHER	IDENTIFIER	TOPIC	COURSE_NAME	SIMILARITY_SCORE
0	Principles Of Australian Constitutional Law, 5Th ...	Book	2016-11-03	Patrick Keyzer; A. Fisher; C. W. Goff		{"type":"ISBN_10","identifier":"0409341959"}; {"ty...		Constitutional Law	67.52000
1	Federal Constitutional Law : A Contemporary Vi...	Book	2014	Sarah Joseph; Melissa Castan	Lawbook Company	{"type":"ISBN_10","identifier":"0455232954"}; {"ty...	Constitutional law	Constitutional Law	67.52000
2	Australian Federal Constitutional Law : Commen...	Book	2007		Lawbook Co	{"type":"ISBN_13","identifier":"9780455219554"}	Australia; Constitutional law	Constitutional Law	67.52000
3	Ac - Appeal Cases (Ichr)	Website						Constitutional Law	67.52000
4	Blackshield And Williams Australian Constitutio...	Book	2014	Anthony Blackshield; Sean Brennan, \u200f \u20...		{"type":"ISBN_10","identifier":"1862879184"}; {"ty...	Constitutional law	Constitutional Law	67.52000
5	Mastering Law Studies And Law Exam Techniques	Book	2013-12-04	Richard Krever		{"type":"ISBN_10","identifier":"040933300X"}; {"ty...	Law	Constitutional Law	67.52000
6	Australian Constitutional Law : Foundations And...	Book	2012-03-01	Suri Ratnapala; Jonathan Crowe	OUP Australia & New Zealand	{"type":"ISBN_10","identifier":"0195519035"}; {"ty...	Law	Constitutional Law	67.52000
7	Vlr - Victorian Law Reports	Webpage						Twentiethcentury Britain Rule Britannia To Cool ...	45.04000
8	War - Western Australian Reports	Webpage			s.n.			Twentiethcentury Britain Rule Britannia To Cool ...	45.04000
9	Australian Indigenous Law Reporter	ePeriodical	2003			{"type":"OTHER","identifier":"STANFORD:361050...	Aboriginal Australians	Twentiethcentury Britain Rule Britannia To Cool ...	45.04000
10	Ch - Chancery Division	Webpage						Twentiethcentury Britain Rule Britannia To Cool ...	45.04000
11	A Crim R - Australian Criminal Reports	Webpage						Twentiethcentury Britain Rule Britannia To Cool ...	45.04000
12	Wlr - Weekly Law Reports	Webpage						Twentiethcentury Britain Rule Britannia To Cool ...	45.04000
13	Ac - Appeal Cases	Webpage						Twentiethcentury Britain Rule Britannia To Cool ...	45.04000
14	The Sydney Law Review	ePeriodical				{"type":"ISSN","identifier":"0082-0512"}	Law reviews; Australia; Law	Twentiethcentury Britain Rule Britannia To Cool ...	45.04000
15	The Australian Law Journal	Periodical				{"type":"ISSN","identifier":"0004-9611"}		Twentiethcentury Britain Rule Britannia To Cool ...	45.04000
16	Qd R - Queensland Reports	Journal						Twentiethcentury Britain Rule Britannia To Cool ...	45.04000
17	The University Of New South Wales Law Journal	ePeriodical	2003			{"type":"OTHER","identifier":"UCAL:B5155920"}		Twentiethcentury Britain Rule Britannia To Cool ...	45.04000
18	Sr (Nsw) - The State Reports, New South Wales	Journal						Twentiethcentury Britain Rule Britannia To Cool ...	45.04000
19	Melbourne University Law Review	ePeriodical		University of Melbourne		{"type":"ISSN","identifier":"0025-8938"}	Law; Law reviews; Australia; Victoria	Twentiethcentury Britain Rule Britannia To Cool ...	45.04000



	TITLE	FORMAT	PUB_DATE	AUTHOR	PUBLISHER	IDENTIFIER	TOPIC	COURSE_NAME	SIMILARITY_SCORE
0	Business Analysis And Valuation : Using Finan...	Book	2009	Sue Joy Wright; Krishna G. Palepu; Phillip Lee; ...		{"type":"ISBN_10","identifier":"0170135098"}; {"...	Business enterprises	Corporate Accounting	82.07000
1	Family Process	ePeriodical		Family Process Institute		{"type":"ISSN","identifier":"0014-7370"}		Case Conceptualisation And Assessment In C...	76.72000
2	Theories Of Psychotherapy & Counseling: Con...	Book	2015-01-01	Richard S. Sharf	Cengage Learning	{"type":"ISBN_13","identifier":"9781305537545"...	Education	Case Conceptualisation And Assessment In C...	76.72000
3	Chapter 1	Chapter	2014					Case Conceptualisation And Assessment In C...	76.72000
4	Family Systems Theory, Attachment Theory, A...	Article	2002					Case Conceptualisation And Assessment In C...	76.72000
5	The Crisis	Journal	2009					Case Conceptualisation And Assessment In C...	76.72000
6	Life Span Development : Australia	Book	2014	John W. Santrock		{"type":"ISBN_10","identifier":"0070997578"}; {"...	Developmental psychology	Case Conceptualisation And Assessment In C...	76.72000
7	Psychoanalytic Diagnosis : Understanding Pers...	Book	2011	McWilliams, Nancy	Guilford Press	{"type":"ISBN_13","identifier":"9781609184940"}	Personality assessment; Personality developm...	Case Conceptualisation And Assessment In C...	76.72000
8	Bmj : British Medical Journal (Online)	Journal						Case Conceptualisation And Assessment In C...	76.72000
9	Contemporary Issues In Early Childhood	ePeriodical	2003	Children's Issues Coalition	Ian Randle Publishers	{"type":"ISBN_13","identifier":"9789766371289"...	Social Science	Critical Perspectives In Early Childhood Educat...	73.57000
10	European Scientific Journal	ePeriodical				{"type":"ISSN","identifier":"1857-7881"}	Science	International And Transnational Crimes	70.31000
11	Journal Of Strategic Security	ePeriodical				{"type":"ISSN","identifier":"1944-0464"}; {"type...	Information technology--Security measures	International And Transnational Crimes	70.31000
12	International Journal Of Music Education	ePeriodical		Lucan, 39-65		{"type":"ISSN","identifier":"0255-7614"}; {"type...		Learning And Support Frameworks	68.13000
13	The International Law Of Human Rights: The I...	Book	2012-01-12	Justine Nolan; Simon Rice	OUP Australia & New Zealand	{"type":"ISBN_10","identifier":"0195568800"}; {...	Law	Legal Protection Of International Human Rights	52.13000
14	Communicating In Geography And The Enviro...	Book	2014-08-28	Iain Hay; Philip Giles (Professor of geography)	Oxford University Press, USA	{"type":"ISBN_10","identifier":"0190074117"}; {"...	Communication in geography	Introduction To Indigenous Research	48.07000
15	Teaching Critical Thinking : Practical Wisdom	eBook	2013-02-01	bell hooks	Routledge	{"type":"ISBN_13","identifier":"9781135263492"...	Social Science	Introduction To Indigenous Research	48.07000
16	The Sage Handbook Of Qualitative Research	Book	2017-02-14	Norman K. Denzin; Yvonna S. Lincoln	Sage Publications, Incorporated	{"type":"ISBN_10","identifier":"1483349802"}; {"...	Mathematics	Introduction To Indigenous Research	48.07000
17	Leadership	ePeriodical				{"type":"ISSN","identifier":"1742-7150"}; {"type...		The Role Of The Midwife As Leader, Mentor A...	36.45000
18	Health & Social Care In The Community	Journal						The Role Of The Midwife As Leader, Mentor A...	36.45000
19	Contemporary Theories Of Learning : Learning...	Book	2009-05-07	Knud Illeris	Routledge	{"type":"ISBN_13","identifier":"9781135226336"...	Education	The Role Of The Midwife As Leader, Mentor A...	36.45000

Furthermore, both types recommendation systems suffer from their own set of advantages and limitations. The primary advantage of the content-based recommender over the collaborative-based recommender was that the former only required information about the products or items it was trying to compare. However, this also meant that the way in which

the feature matrix was constructed, including what metadata items were included and what text pre-processing steps were used would heavily influence the way in which the system measure item similarity. In contrast, collaborative-based recommender required a far smaller feature set as input – typically only requiring information regarding user preferences. However, collaborative-based recommenders are often challenging to construct given the tendency for users to not provide explicit feedback about their preferences, which means that such systems are often unable to make sensible recommendations about items and/or users on which it has limited to no information.

In aggregate, the evidence suggested that the content-based recommender system was the better system irrespective of the higher computational resources required to train the model over a large dataset. The recommender quality metrics were better across both the training and test sets, the recommendations were more reliably less confusing and likely prove to be more relevant for the user, and would not require generation of a user-item interaction matrix. Furthermore, the feature set used to make predictions in a content-based recommendation system is more flexible in that it could be adjusted to include more relevant features or reduced in size if, for example, it was decided that author names should no longer be used to assist in determining item similarity.

Conclusion

The objective of this project was to produce two different recommender systems that could be used to suggest additional readings for courses in the supplied dataset. Whilst this was achieved, and it was concluded that the content-based system was preferred despite the higher computational complexity involved, the results and observations throughout the project also led to a number of possible areas for further research in the future. These included:

- Experimenting with alternative approaches to construct the field of education ontology as the silhouette scores and elbow method indicated that the results were not completely satisfactory.

- Experimenting with alternative methods to construct the user-product matrix for the collaborative filtering system to improve the recommender quality metrics and test set performance, and
- Experimenting with different combinations of descriptive metadata features to observe the effect on the resulting recommendations.

Total word count: 3,246 words

References

Aggarwal, C. C. (2016). *Recommender systems*. Springer International Publishing. http://www.charuaggarwal.net/Recommender-Systems.pdf

Banik, R. (2018). *Hands-on recommendation systems with python*. Packt Publishing.

Hadian, S. (2019, July 16). Deploying a recommender system for the movie-lens dataset - Part 1. *codecentric Blog: IT knowledge from developers for developers*. https://blog.codecentric.de/en/2019/07/recommender-system-movie-lens-dataset/

OpenRefine. (n.d.). *Welcome!* https://openrefine.org/

Ricci, F., Rokach, L., & Shapira, B. (2011). *Introduction to recommender systems handbook*. Springer. https://www.inf.unibz.it/~ricci/papers/intro-rec-sys-handbook.pdf

Scikit-learn: Machine learning in python. (n.d.-a). *2.3. Clustering*. https://scikit-learn.org/stable/modules/clustering.html#k-means

Scikit-learn: Machine learning in python. (n.d.-b). *6.8. Pairwise metrics, affinities and kernels*. https://scikit-learn.org/stable/modules/metrics.html#cosine-similarity

Trove. (n.d.). *What is Trove*. https://trove.nla.gov.au/about/what-trove

Appendix

Appendix A – Acronym Dictionary

Acronym	Description
API	Application Programming Interface
ASCED	Australian Standard Classification of Education
BoW	Bag of Words
DOI	Digital Object Identifier
ISBN	International Standard Book Number
ISSN	International Standard Serial Number
PoS	Part of Speech
SVD	Singular Value Decomposition
TF-IDF	Term Frequency-Inverse Document Frequency
WCSS	Within Cluster Sum of Squares

Appendix B - Technical Details

Hardware	
Operating System	Windows 10.0 Build 18363
PC	Dell G7 7590 x64-based laptop

Programs	**Version**
OpenRefine	3.4.1
Pycharm Community Edition	2020.3.4
Python	3.8.8
Anaconda Navigator	1.10.0

Python Packages	
en_core_web_sm	2.3.1
matplotlib	3.3.4
nltk	3.5
numpy	1.19.2
os	-
pandas	1.2.3
regex	2021.3.17
scikit-learn (sklearn)	0.24.1
time	-

Appendix C – Algorithm Computational Complexity

The following table indicates the computational complexity of each algorithm, measured by how long it took to produce a recommendation for a single user, for five users, and over the full training and test sets[4].

The numbers in the following table represent the approximate time taken for the author's hardware to loop each recommender system over the pre-defined train and test sets.

	Single user	**Five users**	**All users[5]**
Content recommender			
Training set	32.5 seconds	2.8 minutes	8.3 hours
Test set	1.57 seconds	7.6 seconds	23 minutes
Collaborative recommender			
Training set	3.3 seconds	28.4 seconds	1.3 hours
Test set	0.4 seconds	1.1 seconds	3.6 minutes

[4] Note that the content-based recommender system was only run over a random sample of 900 course ids from the relevant training set (which was repeated in the test set as predictions could not be made for unknown user profiles), because it was estimated that it would have taken approximately 26 hours train the algorithm over all 2,896 course ids in the training set and the author lacked sufficient time to allow the algorithm to run for that long.

[5] As per the previous footnote.