Recommender System Project

By Anthony Lighterness

Table of Contents

Abstract

The current study investigated the use of two classes of recommender systems (RS's), including collaborative filtering (CF) and content-based (CB), to produce recommendations of university courses and reading titles, respectively. A dataset containing university courses, associated reading titles and other data items was processed using natural language processing techniques such as stop words removal, stemming, lemmatization, and n-gram dependency grammar. Topic modelling using Latent Dirichlet Allocation (LDA) was also implemented to enhance user profiles for unique courses. The models were quantitatively assessed using the descriptive statistics of recommendations' cosine similarity scores above a 0.4 threshold. In sum, all six RS's successfully produced reasonable recommendations given 21 input queries. No significant differences were observed between outputs using metadata enrichment nor LDA topic modelling. The study recommends future work prioritise the metadata enrichment process.

Introduction

Recommender systems (RS's) have become a prioritised area of research in the last decade. Their implementation in various domains, including knowledge management, search engines, and e-health has led to a substantial increase in digital libraries for information retrieval systems (Champiri et al., 2015). Two of the most widely deployed RS's include content-based (CB) and collaborative-filtering (CF) (Adomavicius & Tuzhilin, 2005). CF approaches typically recommend products to users based on a profile that is similar to other user profiles (Champiri et al., 2015). In contrast, CB methods acts as an information retrieval system as they compute the similarities of input products with those stored in a corpus (Champiri et al., 2015). Despite their wide use, RS's are greatly dependent on the preparation of text data into numeric data. Many processes exist to facilitate this under the field of natural language processing (NLP).

Alan Turing first published work showing the possibilities of NLP and AI in 1950 (Turing, 1950). Since then, NLP has been used in statistical models to analyse semantics and grammar in large corpora. A number of common and important tasks of NLP used to prepare texts for recommender systems include lemmatisation, tokenisation, stemming, dependency grammar, and topic modelling.

Text is arguably the most unstructured form of data. Text data is therefore usually processed through a cleaning pipeline to prepare it for analysis and machine learning. Seen below in Figure 1, noise removal is the first step. Noise can be user-defined and context-dependent however, it usually includes case, punctuation, stop words, or other entities deemed irrelevant such as weblinks, hashtags or webpage tags.

Following the noise removal, the text data is normalised using three key methods. These include tokenisation, lemmatisation, and stemming. Tokenisation separates dense texts into identifiable sentences, and again into separate words or tokens. These are subsequently reduced into root forms, or stems, by removing possible suffixes that indicate variants of the same word, such as "*drive*", "*driver*", "*drives*", and "*driving*". In this example, the stemmed form would remove suffixes, reducing the token to "*driv*". Interestingly, the term "*environmental*" and "*environment*" are stemmed to "*environment*" and "*enviro*", respectively. This shows a potential limitation in the stemming process. Lastly, lemmatisation removes inflectional endings only and returns the base dictionary form of a word, i.e. a lemma. The lemma of the aforementioned example would instead be, "*drive*". However, while the stem of the word "*better*" would remain "*better*", its lemma would be "*good*".

The third step of the cleaning text pipeline is standardisation. Text data or documents may contain words or phrases which are not present or defined in standard lexical dictionaries, such as acronyms, colloquialisms, or slang terminology. As such, these texts may be unrecognised by text-driven models. Data scientists may choose to use regular expressions to manually update the underlying dictionary or corpus used to support a text-driven model. As such, following the cleaning and pre-processing pipeline, the text data is ready for feature extraction and model development.

Figure 1: The process of text cleaning and preparation for NLP.

Feature engineering on text data is a complex, somewhat subjective, yet critical step in preparing datasets for NLP. Techniques include syntactic parsing, such as detecting dependency grammar, entity parsing, like topic modelling and n-grams, and quantifying statistical features, such as term frequency-inverse document frequency (TF-IDF) (Kim et al., 2017). These features are compiled into corpora of texts that are vectorised, clustered, and ingested into NLP models such as RS's. The current investigation sought to analyse two classes of RS's, including CB and CF. To test the hypothesis that metadata enrichment and LDA topic modelling improve the quality of recommendations of these models, the aforementioned NLP techniques were applied which resulted in six RS models. These were analysed and compared both quantitatively and qualitatively.

Methods

The current investigation sought to implement two types of RS's, including CB and CF methods. To do this, a dataset containing university courses with associated reading lists was processed to develop unique profiles for users, i.e. course names, and products, i.e. reading titles. The CBRS was implemented to recommend new reading titles given an input field of study that matched the profile of previous course names. Then, the CFRS was developed to recommend relevant course names based on input subject areas. Both RS's were implemented in parallel using a downsized dataset that was enriched with metadata fetched using the GoogleBooks API on OpenRefine. Berbatova (2009) reported on the efficacy of using this protocol. Lastly, both CBRS's was again implemented in another duplicate to assess the impact of including LDA for topic modelling to label course names with potentially relevant field of

study. In total, six RS's were implemented and compared by quantifying the distributions of recommendations' cosine similarity scores given 21 inputs of search queries.

Firstly, a duplicate dataset was created, downsized, pre-processed, and enriched with metadata using OpenRefine. A random sample of 1,500 unique course names were extracted from each of four universities to generate a sum of 6,000 rows including all original columns. Following general cleaning, the Google Books API ("*https://www.googleapis.com/books/v1/volumes?q=\"+value.replace(\" \",\"%20\")")* was used to retrieve metadata items including a summary text snippet, authors, category, and description. The original dataset was cleaned and prepared for pre-processing by renaming the columns and replacing *RESOURCE_TYPE* values to more relevant substitutes. Then, for both datasets, a bag of words (BOW) was constructed by concatenating data items such as title, subtitle, resource type, and author for unique course names, and replacing title with course name for unique reading titles. These formed the basis of the user and product profiles, respectively and is consistent with the protocol reported by Glauber and Loula (2019).

The BOW text for each respective dataset was pre-processed in a user-defined function. It converted the text to lower case, removed all NaN values, special characters, numbers and stop words, and then was tokenized. The stop words were extended to include extraneous words such as, "the", "and", "of", and "in". Then, the pre-processing function used the NLTK SnowBallStemmer to stem tokens as it is reportedly preferable to the PorterStemmer. Lastly, tokens were lemmatised using the WordNetLemmatizer(). The course name centred data frames were then processed for bigram and trigram sequences to develop a dictionary for LDA.

LDA modelling was performed to extract and classify fields of study for each unique course name. To do this, a dictionary containing processed tokens and n-grams were applied to the doc2bow function, creating a corpus for ingestion into the LDA model. Then, the LDA model was constructed to extract 20 key topics, setting the alpha and eta parameters to 0.01. This was based on the a-priori beliefs on each topic and word's probability to ensure that the most representative topics for each course name were extracted. Topics were extracted through manual qualitative assessment of the LDA results. This was achieved by analysing the contributions of word tokens to each topic, manually defining each topic, and then extracting the dominant topic clusters assigned to each course name. The topic distributions over the body of course names were visualised. In particular, an inter-topic distance map, using multidimensional scaling, was produced using pyLDAvis to confirm the relevant field of study of each topic based on the contributions of constituting word tokens (Sievert & Shirley, 2014).

The literature recommends topic extraction at a lambda value of 0.7 to yield highest possible accuracy (Sievert & Shirley, 2014).

The CB and CF RS's were built using a user-defined function. In it, the constructed BOW for each data frame is transformed into a sparse TF-IDF matrix. Given a string input, which is applied to the aforementioned pre-processing function and also vectorised into a TF-IDF, the cosine similarity is calculated of the BOW matrix and the input string. Then, thirty recommendations, with associated cosine similarity scores, are produced following a five-number statistic which describes the recommendations produced that reached a cosine similarity threshold of equal to or greater than 0.4. As such, these are described in terms of frequency count, the minimum, maximum, median, and interquartile range. The use of these descriptive statistics operates under the assumption that recommendations of similar cosine similarity scores are deemed qualitatively similar as well. For example, "*environmental law*" and "*psychological law*" may yield similar scores for the input "*law*". The cosine similarity metric was chosen due to its wide use particularly for CBRS's (Agarwal & Chauhan, 2017). The six RS models, summarised in Table 1 below, were therefore quantitatively assessed by analysing the five-number descriptive statistics as well as the qualitative value of some example output recommendations.

MODEL	MODEL TYPE	DATASET	SAMPLE ROWS	PROFILE	BOW INPUT	OUTPUTS
1	CB	Original	68,530	Titles	Courses + subtitles, resource type	Titles
2	CB	Reduced + Metadata	6,000	Titles	Courses + metadata	Titles
3	CF	Original	68,530	Courses	Titles + subtitles + resource type	Courses
4	CF	Reduced + Metadata	6,000	Courses	Titles + metadata	Courses
5	CF	Original + LDA Topics	68,530	Courses	Titles + subtitles, resource type + LDA topics	Courses
6	CF	Reduced + Metadata + LDA Topics	6,000	Courses	Titles + metadata + LDA topics	Courses

Table 1: Summary of each recommender system model implemented and analysed.

Results and Discussion

Overall, all six RS's were able to produce some relevant recommendations for at least several of 21 input test queries. Generally, the recommendations produced using the original dataset in its full length were more relevant and fruitful. This is mostly due to the significant loss of information sustained in downsizing the dataset for metadata enrichment. The results also

indicate an overall improvement in recommendation outputs for some input queries when the model is enhanced with LDA topic modelling. This is somewhat indicative of a successful protocol for labelling course names for building profiles.

The results of LDA modelling showed reasonable separation of courses into separate clusters. Appendix 1 below shows the distribution of topics across all individual unique course names for both the original and the reduced, metadata-enriched datasets. This is also observed in the dimensionally reduced scatterplot of clusters, seen in Appendix 2. Despite this, the LDA-labelled corpora yielded very little improvement in terms of quality of recommendation outputs. This can be seen when comparing the outputs of models 3 and 5, and models 4 and 6, where the latter pair were built using the downsized, metadata-enriched corpora. Comparing these two pairs of models, no significant improvements are seen in terms of the frequency count, median, IQR, or maximum values of cosine similarities above 0.4.

Since very little to no changes are observed quantitatively, it can be assumed that the topic labels, e.g. '*sociology'* or '*business'*, were absolved by pre-existing terms of the same stems or lemmas. In other words, the labelling protocol did not assign topics to courses incorrectly on a large scale. The qualitative assessment of sample outputs is consistent with this conclusion as irrelevant topics did not appear to be prioritised.

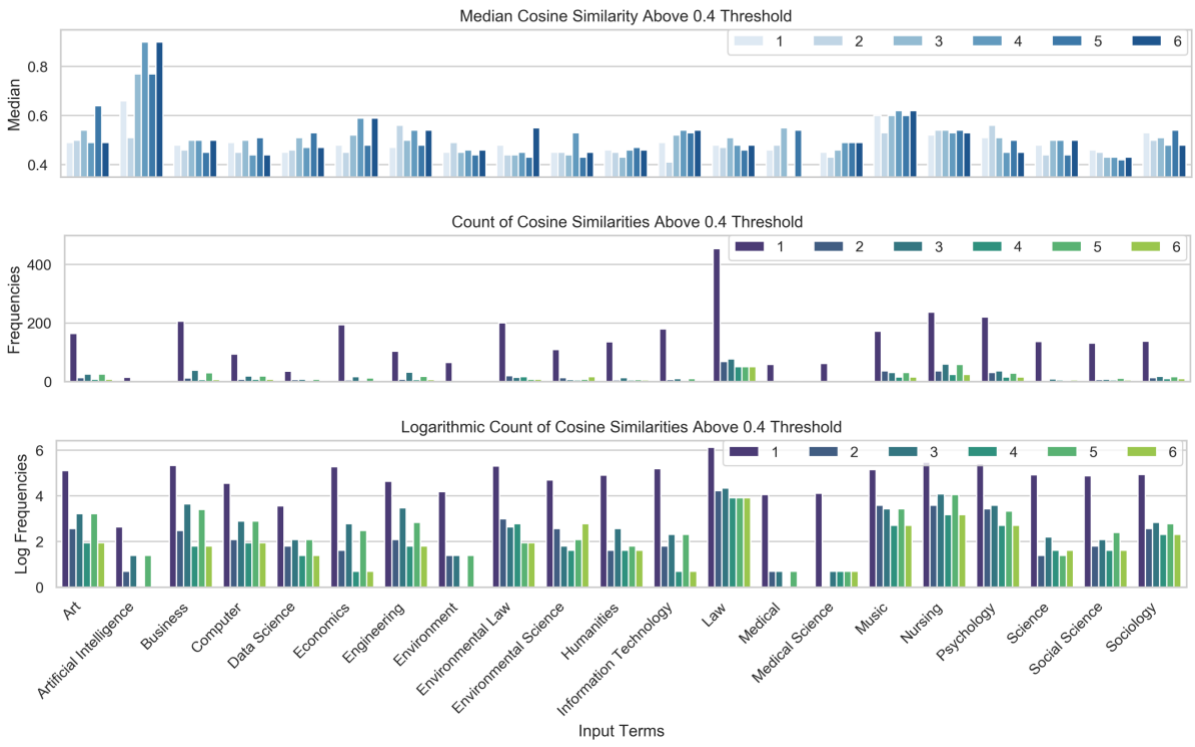


Figure 2: The median and frequency count of recommendations produced above a cosine similarity of 0.4 for given input queries for all six models.

Most strikingly, model 1 achieved the highest frequency count of recommendations with cosine similarities above 0.4 for all test input queries. As seen in Figure 2 below, which shows the median cosine similarity score and the frequency and log frequency counts of scores above 0.4, model 1 outperformed in nearly every test input. This was not surprising as the BOW corpus ingested by model 1 contains the highest volume of text compared to all models. This is because reading title profiles for model 1 and 2 were constructed using the course names, subtitles, which are mostly similar to title names, and the resource type. For the latter, this was extended by including other metadata such as summaries, descriptions, authors, and categories.

Varying input queries impacted the quality of recommendations differently for each model. Interestingly, the quantity and quality of recommendations improved comparing the inputs of "*environment"* versus "*environmental science*' and "*environmental law*". Yet, no significant difference was observed between *"medical"* and "*medical science*". This could be due to the stemming process as "*environment"* becomes "*environ"* whereas "*environmental"* is stemmed to "*environment*". Table 2 below shows example outputs of all six models for three given test inputs.

Model	Input Queries and Outputs		
	"Law"	*"Environmental Law"*	*"Environmental Science"*
1	TITLE / Score An introduction to law 0.761148 Australian Law Journal 0.718097 Foundations of Australian law 0.697946 LAW5152 Taxation Law now LAW4704 0.692664 Law in context 0.677261	TITLE / Score Journal of Environmental Law 0.935708 Environmental and Planning Law Journal 0.870454 An introduction to law 0.841652 Environmental law in Australia 0.826813 Environmental law 0.811525	TITLE / Score Environmental Science Policy 0.742990 Environmental Sciences 0.676657 Science 0.669158 Environmental Science Technology 0.605496 Introduction to environmental engineering and ... 0.604878
2	TITLE / Score Flinders Law Journal 0.726408 The Law Of Work 0.663653 Introduction To Business Law 0.657343 Law In Context 0.657239 Australian Journal Of Labour Law 0.627361	TITLE / Score Public Participation In Environmental Impact A... 0.640018 Standards, Legitimacy And The Law – The New En... 0.619888 The European Experience: Half-time Environment... 0.563589 The Environmental Law Handbook : Planning And ... 0.502875 Environmental And Planning Law Journal 0.496889	TITLE / Score Essential Environmental Science: Methods & Tec... 0.666380 The Journal Of Environment & Development 0.605670 Environmental Psychology An Introduction 0.530268 Environmental Management 0.525213 Public Participation In Environmental Impact A... 0.518615
3	Coursename / Score LEGANTO TEST 0.711783 SELF AND SOCIETY 0.670283 Introduction to Sociology (OUA) 0.628827 INTRODUCTION TO SOCIOLOGY 0.627716 DIVERSITY, SOCIAL JUSTICE AND LEARNING 0.611836	Coursename / Score Environmental Policy and Law 0.670770 Climate Change, Sustainability and Environmen... 0.615260 WATER AND WASTEWATER TREATMENT 0.536987 FOUNDATIONS OF TORTS 0.505860 ENTERPRISE LAW 0.482817	Coursename / Score WATER AND WASTEWATER TREATMENT 0.571552 Environmental Management 0.461034 Environmental Policy and Law 0.459089 Environment and Society 0.420371 ADVANCED STATISTICAL HYDROLOGY 0.412173
4	Coursename / Score Introduction To The Business Law Of Papua New ... 0.712567 Law, Justice And Social Policy 0.690994 Mining And Natural Resources Law 0.685512 Foundations Of Business Law 0.682747 Contract Law 0.669108	Coursename / Score Pollution And Its Control 0.532680 Mining And Natural Resources Law 0.488092 Environmental Policy And Law 0.485905 Introduction To The Business Law Of Papua New ... 0.478171 Law, Justice And Social Policy 0.463695	Coursename / Score Pollution And Its Control 0.621953 Environmental Assessment And Management 0.542426 Environment And Society 0.528143 Behaviour And Environment 0.522984 Evolutionary Ecology 0.500628

5	Coursename / Score FOUNDATIONS OF TORTS 0.704633 Foundations of Business Law 0.682057 ENTERPRISE LAW 0.665407 INDIGENOUS RIGHTS 0.628421 UEH : ENTERPRISE LAW 0.574406	Coursename / Score WATER AND WASTEWATER TREATMENT 0.749894 Environmental Policy and Law 0.606369 Climate Change, Sustainability and Environmen... 0.546159 FOUNDATIONS OF TORTS 0.429344 Foundations of Business Law 0.415588	Coursename / Score WATER AND WASTEWATER TREATMENT 0.611458 Statistical Consulting 0.498893 Environmental Policy and Law 0.438679 Contemporary Public and Environmental Health P... 0.435144 Environmental Management 0.429693
6	Coursename / Score Introduction To The Business Law Of Papua New ... 0.712567 Law, Justice And Social Policy 0.690994 Mining And Natural Resources Law 0.685512 Foundations Of Business Law 0.682747 Contract Law 0.669108	Coursename / Score Pollution And Its Control 0.532680 Mining And Natural Resources Law 0.488092 Environmental Policy And Law 0.485905 Introduction To The Business Law Of Papua New ... 0.478171 Law, Justice And Social Policy 0.463695	Coursename / Score Pollution And Its Control 0.621953 Environmental Assessment And Management 0.542426 Environment And Society 0.528143 Behaviour And Environment 0.522984 Evolutionary Ecology 0.500628

Table 2: Samples of recommendations for all six models given three example inputs.

The current investigation suffers from a number of limitations. Firstly, the metadata-enriched dataset suffered a significant information loss during the process of downsizing moving from 68,530 rows to 6,000. This was necessary to complete the Google Books API fetch on OpenRefine, requiring over ~24 hours on 6,000 records. As such, it was deemed unfeasible to repeat the same protocol on the original dataset in full. To address this limitation, the study recommends that future work aim to pursue alternative approaches to enriching the original dataset in its full length.

Another limitation is that the study assumed that a total number of 20 topics for modelling was appropriate. However, techniques to identify the optimal number of topics, such as measuring the clustering coherence across a range of topic numbers, should be investigated. The current study also lacks rigorous evaluation protocols. Since it is reported that the evaluation stage is of great relevance in the design of RS's, further work needs to address this by exploring alternative offline assessment protocols (Glauber and Loula, 2019). Lastly, future work should explore the use of other similarity metrics for comparisons. However, to maintain brevity, the current study did not investigate Jaccard similarity, pearson correlation, or fuzzy string matching (Agarwal & Chauhan, 2019).

Conclusion

In conclusion, the current study reports the implementation of six RS models with varying features. Profiles were developed for users, i.e. courses, and products, i.e. reading titles, which were then processed using NLP techniques including text cleaning, stemming, lemmatisation, and dependency grammar. Topic modelling using LDA was used to label courses with a field of study. Other efforts including metadata enrichment were used however, due to time constraints was not successfully implemented on the full original dataset. Instead, a downsized dataset was studied.

References

Adomavicius, G., and Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, *17*, pp. 734-749.

Aggarwal, A., and Chauhan, M. (2017). Similarity measures used in recommender systems: A study. *International Journal of Engienering Technology Science and Research, 5(*6), pp. 619-626. Retrieved from https://pdfs.semanticscholar.org/943a/e455fafc3d36ae4ce68f1a60ae4f85623e2a.pdf

Aggarwal, P., Tomar, V., and Kathuria, A. (2017). Comparing content based and collaborative filtering recommender systems. *International Journal of New Technology and Research*, 3(4), pp. 65-67.

Berbatova, M. (2009). Overview of NLP techniques for content-based recommender systems for books. *Proceedings of the Student Research Workshop associated with RANLP-2019*, pp. 55-61. Doi: 10.26615/issn.2603-2821.2019_009

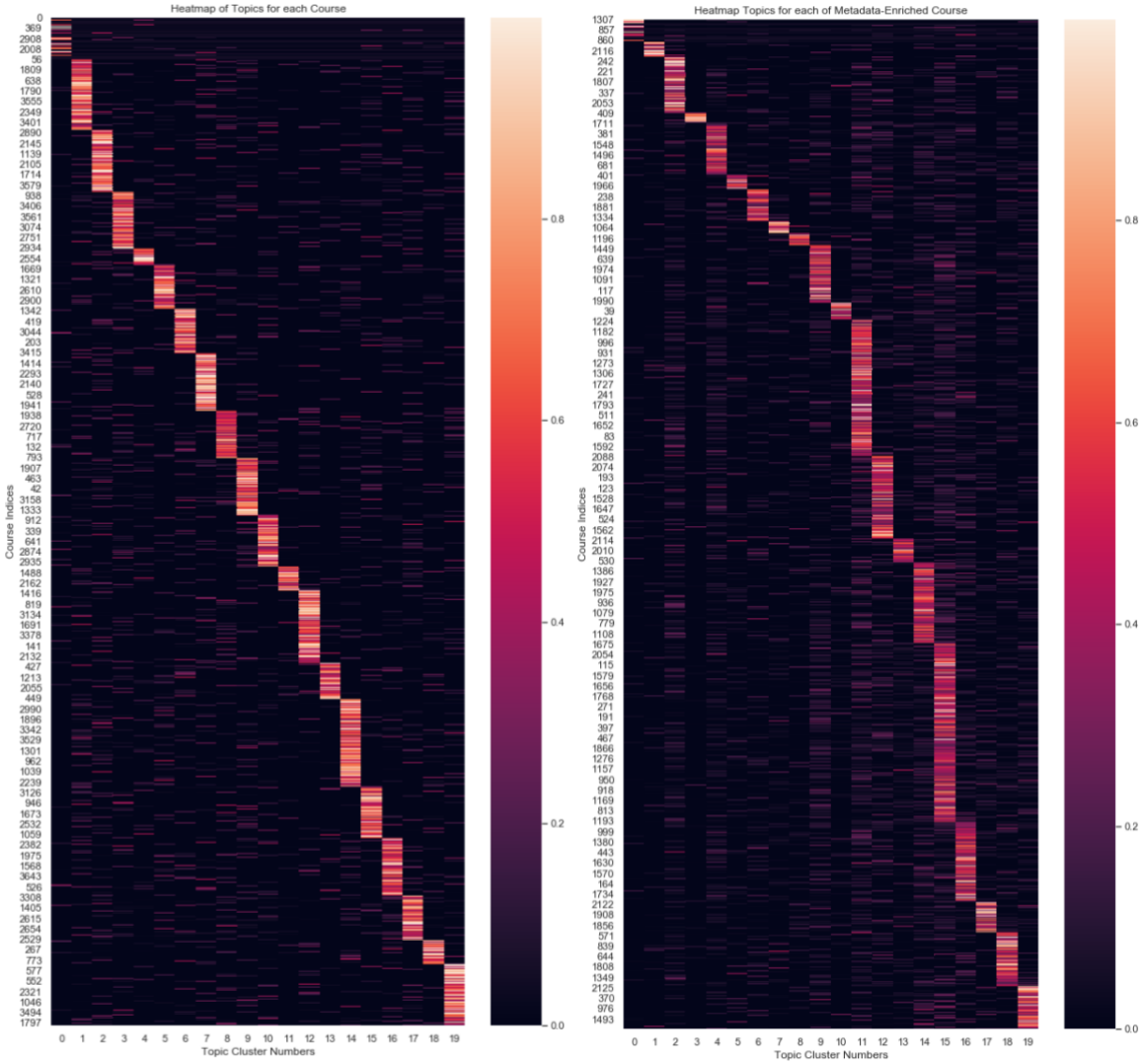
Glauber, R., & Loula, A. (2019). Collaborative filtering vs. content-based filtering: differences and similarities. arXiv: 1912.08932. Retrieved from https://arxiv.org/pdf/1912.08932.pdf

Kim, D., Park, C., Oh, J., and Yu, H. (2017). Deep hybrid recommender systems via exploiting document context and statistics of items. *Information Sciences, 417*, pp. 72-87. Doi: 10.1016/j.ins.2017.06.026

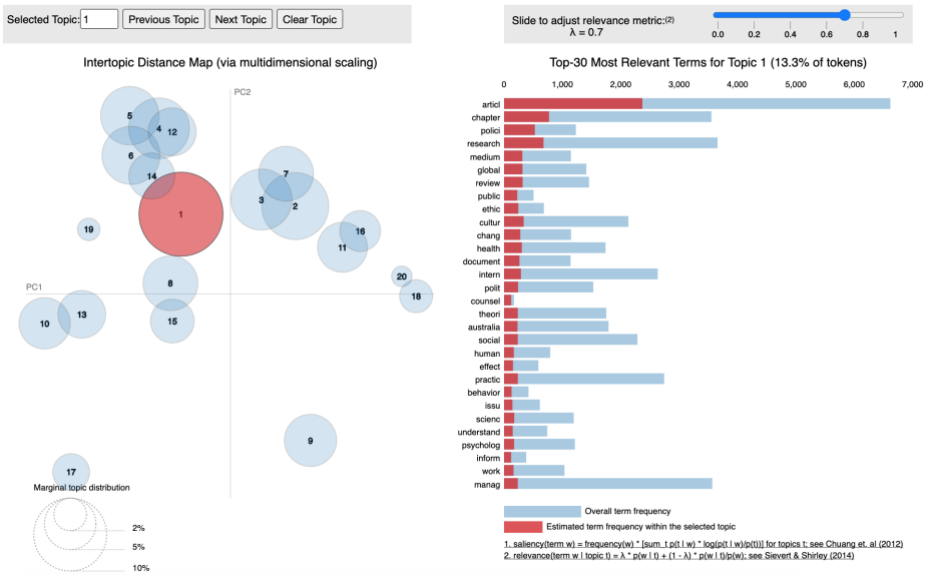
Sievert, C., & Shirley, K.E. (2014). LDAvis: A method for visualising and interpreting topics. Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. pp. 63-70. Doi: 10.3115/v1/W14-3110

Turing, A. (1950). Computing Machinery and Intelligence. Oxford Academic, 236, pp. 433-460. Doi: 10.1093/mind/LIX.236.433

Appendix 1: Topic Distributions



Appendix 2: LDA Topic Modelling



Appendix 3: Quantitative Assessment