Harish Yadav. Development of a Course Recommender System for Students. A Master's Paper for the M.S. in I.S degree. May, 2019. 69 pages. Advisor: Dr. Stephanie W. Haas

Students at the university have an information need to find the courses of their interest. The current university registration portals do not fulfill this information need completely. We have proposed the development of a recommender system which can take a course name and based on the description of that course recommend other courses to students. The recommended course list could help save time and effort for students registering for courses. The proposed system was trained with sample data collected from the course catalog of the University of North Carolina at Chapel Hill. We tested the recommender system with different courses as input and evaluated the resulting recommended courses.

Headings:

Recommender System

Natural Language Processing

Latent Dirichlet Allocation

Topic Modeling

DEVELOPMENT OF A COURSE RECOMMENDER SYSTEM FOR STUDENTS

by
Harish Yadav

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

May 2019

Approved by

_____

Dr. Stephanie W. Haas

Table of Contents

1 Introduction

1.1 Background

Recommendation systems using predictive modeling have been widely implemented in applications from online retail stores to restaurant menu suggestions. Machine learning techniques, for example, clustering and topic modeling, can provide effective ways to identify potential options that might interest the user. It may be beneficial for the user to select directly from the recommendations rather than wasting unnecessary time and energy in browsing through all the available options.

Within course registration portals, there are hundreds of courses in different departments, and for a student to search for a course of interest by browsing is time-consuming and frustrating. The registration portals typically give the flexibility to choose a particular course on the basis of course name, course number (course code), semester (fall/spring/summer) and instructor name. However, the portals generally lack the flexibility to recommend potential courses that a student might be interested in.

1.2 Problem Statement

There is an information need for the students to find courses of their interest. At the same time, current course registration portals do not adequately fulfil this need. So, to address this gap we have built a recommender system which will allow students to enter a

course name and to obtain recommended courses based on the description of the entered course. Consequently, our system will help students in finding courses of interest.

The remainder of the paper is organized as follows. We first review the literature in section 2. Then we discuss our proposed recommender system in section 3. Section 4 details the methodology. Section 5 discusses the results and section 6 provides the conclusion. Finally, we discuss future work in section 7.

2 Literature Review

This section provides an overview of previous research about recommender systems. It introduces and summarizes the methodologies and techniques for different recommender systems which outlines our approach to address the problem statement.

2.1 Recommender systems and their techniques

Different techniques, for example, clustering and topic modeling, can be employed for the purpose of finding items which might fall in the same category due to some of their characteristics. These techniques can serve as an initial step for the similarity functions which help recommender systems to find relevant items for suggestions. Al-Badarenah & Alsakran (2016) proposed a recommendation system that could recommend courses to the students and the expected grades in those courses. The dataset was created with the help of a sample of 2000 graduate students of Electrical Engineering. Their dataset contained records of courses and grades for the students where each record was associated with a unique student ID. The dataset was divided into training and testing sets. They performed K-means clustering on the training dataset to cluster the student data. Once the clustering was performed, they used *n-nearest neighbor* as their similarity function to decide the cluster in which a target student from the test set would belong on the basis of the student's courses and grades. Consequently, the association rules were developed to predict the grades of the target student in a

recommended course. Association rules are a way to find the correlations between the items based on their occurrence together in a transactional dataset (Al-Badarenah & Alsakran, 2016). The association rules were built on the basis of grades. An example of course association rule was [Alg:A] ^ [OS:D] => [Parallel:C], this indicated that if the target student got an A in Algorithms and a D in Operating Systems courses, then the system would recommend the student to take the Parallel Computing course with an expected grade of C. To evaluate the performance, Precision and Recall were calculated with the help of the following formulae –

$$\text{Precision} = \frac{\text{number of recommended courses taken}}{\text{total number of recommended courses}}$$

$$\text{Recall} = \frac{\text{number of recommended courses taken}}{\text{total number of courses taken by students}}$$

However, the proposed system tended to prompt the students to select only those courses where chances of getting better grades were higher.

Similarly, Grewal & Kaur (2016) presented a recommender system that recommended courses on the basis of students' interests and their grades of five courses in high school. The system used students' interests as an added characteristic as opposed to Al-Badarenah & Alsakran (2016) The dataset was created with the help of the university students who volunteered for the study and provided their information through a questionnaire. The information collected was the grades of any five courses from high school and the field of interest like accounting, medicine, etc. Grewal & Kaur (2016) employed K-means clustering algorithm to cluster student data based on students' interests and grades. They used the association rules in a similar fashion as Al-Badarenah & Alsakran (2016) to find links between the data in a cluster in order to recommend a course (Koren, Bell & Volinsky, 2009; Ricci, Rokach, & Shapira, 2011). However, the

evaluation of the system was based on the students' feedback about satisfaction with the recommendations. This method of evaluation did not provide a robust way of assessment, as it did not take into account the bias in the sample population. The bias was created since the responses of students who volunteered for the study may not be representative of the responses of the entire population of student in the university.

The common techniques discussed above like clustering and association rules used in recommender system present a quick way to get the recommendations for users. There have been attempts to improve the quality of recommender systems by deploying added filtration before recommending to the user. Aher & Lobo (2012) proposed a recommender system where the preprocessing time of dataset was reduced. A sample of students was taken from university Moodle Database. The dataset contained student ID and the courses taken by students. The unprocessed dataset contained missing values for some of the records of students. K-Means clustering was performed on the unprocessed student data to form clusters on the basis of courses taken by the students. Then the ADTree Classification algorithm was applied to the clusters to get specific clusters which could be utilized for creating association rules. These clusters did not contain data with missing values and therefore were best fit for association rule creation. Hence, Aher & Lobo (2012) asserted that applying association rule algorithms on the classified dataset could help to get rules with the increased strength of association.

2.2 Recommender systems using different approaches

2.2.1 Content-based recommender system

Sarwar, Karypis, Konstan, & Riedl (2001) proposed generation of recommendations from Content-based algorithms. They created a dataset by randomly selecting enough users to gather 100000 ratings of the movies from MovieLens[1] (a web-based research recommender system). The dataset was divided into train and test sets They used Content-based recommendation algorithm which focused on computing similarity between the items and then selecting the most similar items. The similarity was calculated using Correlation-based similarity, Cosine-based similarity, and Adjusted Cosine similarity. Once the similarity was calculated the recommendations were generated using Regression model and Weighted sum on the most similar items for a user. The results of Content-based algorithms were compared to the basic k-nearest neighbor approach which is the widely accepted technique used in the recommender systems. Their experiments showed that Content-based recommender systems provided better sensitivity than k-nearest neighbor based recommender systems.

2.2.2 Context-aware recommender system

The Content-based recommender systems use simple models. Adomavicius & Tuzhilin (2011) proposed a Context-based recommender system which was built based on the knowledge of partial contextual user preferences. The context here took into account the factors like in what situation the product was being used. For example, a user may be willing to buy a phone because of its better camera quality while another user

[1] https://movielens.org/

might give preference to its sound quality. A user's choice would differ in a different context and hence recommending choices without considering the context might not prove beneficial for the users. The context independent approach of Content-based recommender systems loses its prediction ability.

Adomavicius & Tuzhilin (2011) discussed the use of contextual prefiltering where the selection of only relevant features was performed. The selection was based on the importance of features of the product in determining user's choice. Then with the help of those features, the traditional recommendation system was utilized for the prediction. However, in contextual post filtering, only the users and items were considered as input in the recommendation process to predict the rating at first, and the context was ignored. The usage of two features, users and items, is referred to as two-dimensional (2D) recommendation model. The recommendations were finally adjusted by applying the contextual information. Thus, contextual based filtering not only considered the reasons due to which the user liked the item, but also the information about the context in which the item was used.

2.2.3 Context-aware recommender system in a multi-dimensional model

Adomavicius, Sankaranarayanan, Sen, & Tuzhilin (2005) talked about a multi-dimensional model to provide recommendations by considering different dimensions including context. Traditional recommender systems like Content-based recommender systems take only two dimensions into account, users and items. Initially, the ratings are implicitly inferred by the system or explicitly provided by the users. Once the initial ratings are specified the rating function R is calculated for a new incoming item which has not been rated yet by the users.

Adomavicius et al. (2005) explained that the multi-dimensional model provided recommendations over several other dimensions like time, place, and so on along with user and item. For example, for movie recommendations user, time and place could be considered as the different dimensions where place represented the places where the movie can be seen like a movie theatre, home TV, VCR and DVD. Time represented when the movie can be seen as a time of day, day of the week, month and year. User represented for whom the movie was recommended which was defined by user attributes (Age, Occupation, Address etc.)

The author considered different dimensions as different sides of a cube in 3D space in the model. A traditional recommender system would provide recommendations of movies to a user, on the other hand, the multi-dimensional recommender system would recommend movies to the user considering the time and place as well.

2.2.4 Knowledge-based recommender system

Another approach of finding the recommendations is employing Knowledge-based recommender system. Knowledge-based recommender system utilizes knowledge about users and products to pursue a Knowledge-based approach for generating a recommendation by reasoning about which products meet the user's requirements. For example, if a user enters the movie name in recommender system as "The Verdict," a courtroom drama starring Paul Newman, the system would recommend a handful of other movies that are similar. However, if the user wants something suspenseful as well, the system allows user to add features. The knowledge of this feature will help the system to find a movie with courtroom drama and mystery. Burke (2000) suggested that Knowledge-based recommender system could avoid a lot of drawbacks in other

recommender systems as it did not need a large dataset, also there was no dependence on an initial base user set. Burke (2000) favored knowledge-based recommender system over Content-based. However, there has not been much development in the area of Knowledge-based recommender systems.

2.3 Recommendation system using topic modeling and association rule mining

Topic modeling has been widely used for the purpose of recommender system development. The section below describes two such uses of topic modeling, one to check the behavior of Android application against its description mentioned in the Google Play Store and other is to identify the legitimate access to the files in hospital management systems.

2.3.1 Topic modeling (Latent Dirichlet Allocation LDA)

With the growing importance of recommendations in E-Commerce, recommender system has also found its place in recommending the Android applications. It has been a concern for developers to check whether an application behaves as it was advertised or not. There are many Android applications which do not provide the exact functions that they claim to provide. To date, the malicious behavior of any application has been detected by comparing the application code with respect to a predefined malicious code. Gorla, Tavecchia, Gross, & Zeller (2014) suggested that this technique would not be able to handle the new attacks as it was hard to define in advance whether some program behavior would be beneficial or malicious. Thus, Gorla et al. (2014) developed an approach to check the behavior of the application against its implementation known as

CHABADA (CHecking App Behavior Against Descriptions of Apps) approach. They collected the description of more than 25,500 Android applications downloaded from the Google Play Store to create a dataset.

CHABADA approach used LDA to perform topic modeling using descriptions of all the applications. Then the topics were categorized under different categories like "weather", "map" etc. After that, they clustered applications by related topics. For instance, apps related to "navigation" and "travel" shared several topics in their description, so they formed one cluster. Finally, they identified the APIs used by the applications in each cluster and tried to check if any of the application using the APIs which were not expected as per the cluster they belonged. For example, if an application belonging to a cluster related to "navigation" and "travel" but it was using the APIs to fetch personal account information then it was flagged as malicious application due to this behavior. Gorla et al. (2014) were able to flag 56% percent of known malware using their dataset. Pandita, Xiao, Yang, Enck, & Xie (2013) also tried to find the risk associated with an Android application using application description but their approach required manual annotation of the sentences in the description which might be using sensitive APIs whereas CHABADA approach did not require this annotation.

Gupta, Hanson, Gunter, Frank, Liebovitz, & Malin (2013) also used topic modeling to detect insider threat in hospitals. They collected user access and audit logs for four months derived from Cerner Powerchart EMR system at Northwestern Memorial Hospital (NMH). Their dataset contained 4.9 million accesses made by 7932 users to 14606 patients. They proposed the use of Random Topic Access (RTA) model. In this model, the topic modeling was performed on patient logs with the help of LDA. The

topics helped to create the categories of diagnosis, medications, procedures and locations/services. They further made use of these categories to classify users on the basis of their access to the records in any of those categories. If a user's access was category specific, for example only to location/services, with no random access to another category, then the access was not considered a threat to the hospital system. However, if the access was class agnostic then it was considered to be a threat to the system. The Area Under curve from Receiver Operating Curve for all the users was calculated to decide the efficiency of the model. The maximum Area Under Curve obtained was 0.8 for all the users.

2.3.2 Association rule mining

Association rule mining is one of the popular data mining techniques which focuses on finding the association rules in a dataset (Shaheen & Shahbaz, 2017). Association rule, in its basic form, is described as" if an event X occurs it is more likely for event Y to occur". It is written as $X \rightarrow Y$ where X is the antecedent and Y is the consequent (Shaheen & Shahbaz, 2017). Association rule mining is the underlying technique used in many recommender systems to find similar items. It uses Support and Confidence to create the rules. Support is an indication of how frequently the itemset appears in the dataset and Confidence is an indication of how often the rule has been found to be true. Apriori algorithm which is a conventional association rule mining technique finds general trends in a dataset by searching the antecedent (X) and the consequent (Y). The conventional association rule mining algorithms have been modified to provide better association rules. Lin, Alvarez, & Ruiz (2002) have proposed that for targeted user recommendations the conventional association rule mining techniques

required a modified version to find patterns in a dataset. Shardanand & Maes (1995) and Resnick, Iacovou, Suchak, Bergstrom, & Riedl (1994) have also proposed variants of association rule mining techniques that used the Pearson correlation between two user profiles to predict the rating of an item. They predicted the rating for an item I by a user U on the basis of the rating given by a user U' for the same item I. One of the drawbacks of conventional rule mining is that it uses the entire dataset which can lead to a lot of time consumption with a large number of rules. Lin et al. (2002) proposed to mine rules only for one target user at a time. They proposed an algorithm which selected a certain number of rules for each, thereby reducing the overhead of using the entire dataset. The certain number of rules were selected by adjusting the minimum support for each user. This saved a significant amount of processing time.

2.4 Recommender systems using additional controls

Course recommender systems in the majority of the literature discussed so far do not take account of any constraints while generating recommendations. There was no consideration of curriculum requirements before the courses were recommended. For example, completion of certain prerequisite courses for higher level Math courses is one of the constraints which course recommender systems should take into account. Parameswaran, Venetis, & Garcia-Molina (2011) proposed a recommender system that took the prerequisite constraint into account. They collected data using the transcripts of 558 undergraduate students at Stanford University who graduated in Fall 2008 with majors in biology, computer science, economics, electrical engineering, or human biology. The prerequisites for the courses in these majors were collected separately to

account for the constraints required in taking a course. They applied the Ford-Fulkerson algorithm, one of the max-flow algorithms, on the dataset. The algorithm could help select a path on a network of nodes where nodes can represent anything like cities, airports etc. The selected path ensures that all the constraints are met in the network to reach from source to destination node. Parameswaran et al. (2011) used the algorithm to find a combination of courses which satisfied all the constraints. Once the combination was retrieved, traditional recommender system techniques (Sarwar et al., 2001; Adomavicius & Tuzhilin, 2005; Burke, 2000; Pazzani & Billsus, 2007; Burke, 2002) were used to find the best recommendation.

Likewise, Feng, Cao, Wang, & Qian (2017) proposed a recommender system which could recommend movies. They discussed that movie recommendation should not be based on movie ratings only. They suggested the use of text-based information available about the movies and their ratings to generate recommendations. They talked about extracting the hidden features like the adventure picturized in the movie or some other characteristic from the text which can make the user to like the movie. The use of hidden features of movies enabled their recommender system to recommend movies with no user ratings as well.

2.5 Evaluation metrics for recommender systems

It has been a very challenging task to identify the best algorithm for a given purpose since the researchers do not agree on the attributes to be measured and the metrics to be considered for each attribute. As mentioned by Herlocker, Konstan,

Terveen, & Riedl (2004) evaluating recommender system and their algorithms was a very tough task due to the different reasons mentioned below,

Firstly, many collaborative filtering algorithms have been designed specifically for datasets where there were many more users than items (e.g., the MovieLens data set has 65,000 users and 5,000 movies). Such algorithms may be entirely inappropriate in a domain where there were many more items than users (e.g., a research paper recommender with thousands of users but tens or hundreds of thousands of articles to recommend).

Secondly, the goals for evaluation performance may differ. Much early evaluation work mainly focused on the accuracy of the recommender systems in prediction but in today's era, the evaluation relies on other properties which have much more effect on user's satisfaction. Serendipity was one of the properties which played a great role in the recommendation system. In a recommender serendipity is the experience of receiving an unexpected and fortuitous recommendation. Also, the system should take into account if recommendations have been previously explored by the user. However, if the system did not take this into account, then the accuracy of the system was not affected.

Finally, another challenge in the evaluation was about the decision of a combination of measures to be used in a comparative evaluation. Herlocker et al. (2004) noticed that evaluation of the newest algorithm for a recommender system was based on reducing Mean Absolute Error. The reduced Mean Absolute Error assured that the algorithm performed better on the test dataset. Although the algorithm may be better in performance, but the author found out that nearly all the algorithms, when tuned to their

optimum, produced almost similar results. So, reduced Mean Absolute Error may not increase the usefulness of the recommender system.

All the recommender systems get evaluated on rather common evaluation techniques like accuracy, precision, recall, or Mean Absolute Error. Current accuracy metrics, such as Mean Absolute Error (Herlocker, Konstan, Borchers, & Riedl, 1999), measure recommender algorithm performance by comparing the algorithm's prediction against a user's rating of an item. In essence, the recommender system is rewarded for recommending courses a user has previously explored. This prevents the recommender system to provide recommendations which are fruitful. McNee, Riedl, & Konstan (2006) discussed that the common metric of evaluation of recommender systems as accuracy could hurt the quality of recommender system even if the accuracy of the system was very high. Such a recommender system providing the same recommendation, again and again, would achieve higher accuracy even though practically they were not a quality recommendation. Rashid et al. (2002) also expressed concerns about the same suggestion popping up again and again in a recommender system with the example of Star Trek (once a user rated one Star Trek movie, she would only receive recommendations for more Star Trek movies). Hence, McNee et al. (2006) proposed some aspects which could be considered to find out if the recommendations provided by a recommender system could potentially be relevant to the user.

Serendipity was proposed as one aspect. A recommender system with serendipitous recommendations could be very useful for a user but may not score higher on the scale of accuracy.

User's experience and the expectation of finding user specific recommendations was another aspect. Many recommender systems with high accuracy are employed in a diverse setting without taking into account as to which background and culture the recommender system was originally developed. For example, the Portuguese version of a research paper recommender system being converted to the English version without considering subtle nuances of the language can disappoint users (Torres, McNee, Abel, Konstan, & Riedl, 2004).

Finally, the recommender system providing recommended items which are most similar to the searched item can achieve higher accuracy, but would not produce novel or helpful recommendations. So, in the case of course recommender system if it provides the recommended courses which user already expects then the system has no utility.

3 The Proposed Recommender System

Course recommender systems can be seen as a tool for students. Generally, the registration process starts before the end of a semester when students are busy preparing for exams and other deliverables. A student who has many deliverables during the end of the semester is required to do both, register for the mandatory courses and explore the courses of personal interest. In this situation, students may not be able to provide ample time for proper research about the courses they plan to take in the next semester. We developed a recommender system in this paper which may help ease the pain points of students and universities. With the help of the recommender system, the students would be able to enter any course in which they have some interest and the system would recommend a few courses which the student might like to enroll in. These recommendations will benefit students as they will have a shorter list for the purpose of course selection.

Our recommender system would provide students a search box to enter a course title. The system would make use of the course title to map its description and department name for finding similar courses which student might like to register. We used three datasets, Dataset 1, Dataset 2 and Combined Dataset, containing courses from different departments of the University of North Carolina at Chapel Hill. We trained our recommender system on each of the datasets. We used the description of course, course name and department abbreviation to perform topic modeling using LDA. Our

approach was similar to Gorla et al. (2014) who used the description of the Play Store applications for topic modeling. With the help of topic modeling, we determined the similarity between the course from the search box and courses in the dataset. The similarity helped in finding the courses to be recommended. We followed the Content-based strategy to find similar courses for recommendations as used by Sarwar et al. (2001) in their research.

4 Methodology

To develop the proposed recommender system, we created three datasets. We prepared Dataset 1 containing courses mainly related to science, technology, and mathematics while Dataset 2 contained courses mainly related to arts. We also developed a Combined Dataset which was a union of Dataset 1 and 2. This type of dataset creation strategy helped us to train our model on homogeneous course datasets (Dataset 1 and 2) and on the combined courses dataset. The disparate courses in the Combined Dataset could play a role in recommending courses which are not always obvious or expected.

In order to test the recommender system trained on Dataset 1, it was fed with a query course (a query course is the input course provided to the recommender system). We fed 10 different query courses to the recommender system trained on Dataset 1. Similarly, the recommender system trained on Dataset 2 was fed with 10 different queries. Finally, the recommender system trained on the Combined Dataset was fed with the same queries used in the case of Dataset 1 and 2, making a total of 20 queries.

4.1 Dataset Creation and Preprocessing

The data was scraped from the course catalog of the University of North Carolina at Chapel Hill[2]. The catalog provides the following information:

- Department abbreviation

[2] http://catalog.unc.edu/courses/

- Course number
- Course name
- Number of credits
- Course description
- Requisites
- Grading status

Figure 1 shows a sample course from the course catalog website:

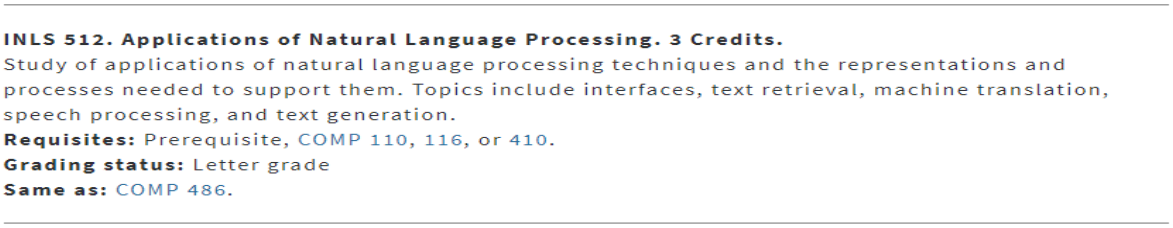

Figure 1: Course information in the course catalog

The modeling was performed with the help of three fields described above: department abbreviation, course name and course description. These three fields played a major role in course distinction. The course description was most crucial as it elucidated different characteristics of the course, the department abbreviation helped in distinguishing the department of the course. The course name helped in distinguishing from other courses as all the courses have distinct names. We used the combination of department name and course number as the unique ID for each course.

All the information other than department abbreviation, course name and course description were not considered to be a discriminating factor because there were some fields which had the identical values for the majority of the courses and therefore were not scraped from the catalog page. Among the fields mentioned above, grading status field was dropped to scrape because it was not a distinctive feature for the courses. All the courses had same grading status with the exception of a few, for instance almost all the graduate

courses in the Department of School of Information and Library Science had grading status as “Letter Grade”. Similarly, the number of credits ranged from 0 to 4 for all the courses and did not play a role in distinguishing what a course was about, so the number of credits was also dropped to scrape. Furthermore, Course Number and Requisites also did not play a role in discriminating one course from another, on the contrary, they had the potential to create ambiguity in topic modeling since different departments tend to have same course numbers for different courses. Hence course number and requisites were also not scraped. We used BeautifulSoup[3], a Python library which can read HTML files and search relevant content in the file using *find_all* functionality. BeautifulSoup helped to create a dataset in a CSV format out of the web page from the course catalog. All Python codes have been presented in a GitHub repository[4]

Below is the detailed description of three datasets developed using the three fields (department abbreviation, course name and course description).

4.1.1 Dataset 1

The first dataset was created using three departments: School of Information and Library Science (SILS), Statistics and Operations Research (STOR) and Computer Science (COMP). These three departments were chosen to create the first dataset to train the model because they had courses which were of common interest with respect to the career paths. The content of the courses also overlapped to some extent in these departments. So, one can assume that the students searching for courses in one of these departments could potentially find the courses useful in the other two departments. Hence, these departments

[3] https://pypi.org/project/beautifulsoup4/
[4] https://github.com/harishyadav909/Course-Recommender-System

would serve as a good choice to be included in Dataset 1. The dataset contained 343 courses scraped from the three departments. We excluded the courses with no description and the courses pertaining to thesis or master's paper. In order to test the recommender system trained on Dataset 1, we arbitrarily chose query courses, 3 from STOR, 3 from COMP and 4 from INLS.

4.1.2 Dataset 2

The second dataset was developed using three departments, distinct from the departments used in Dataset 1. The departments were History (HIST), Geography (GEOG) and English (ENGL). The courses in these departments fall under the arts category. The same information, i.e., department abbreviation, course name and course description were scraped for these departments. This dataset contained 775 courses. The courses pertaining to thesis or master's paper and the courses with no description were eliminated. In order to test the recommender system trained on Dataset 2, we arbitrarily chose query courses, 3 from GEOG, 3 from ENGL and 4 from HIST.

4.1.3 Combined Dataset

The third dataset was a union of Dataset 1 and 2. It contained 1118 courses. We aimed to train a model which could perform well on this heterogenous dataset as this would provide an opportunity for students to find courses out of serendipity. To test the recommender system trained on Combined Dataset, we used the same query courses as in the case of Datasets 1 and 2. This helped us to analyze the output when the same query courses were fed as input to the recommender system trained on the Combined Dataset. Figure 2 shows the schematic diagram of the three datasets

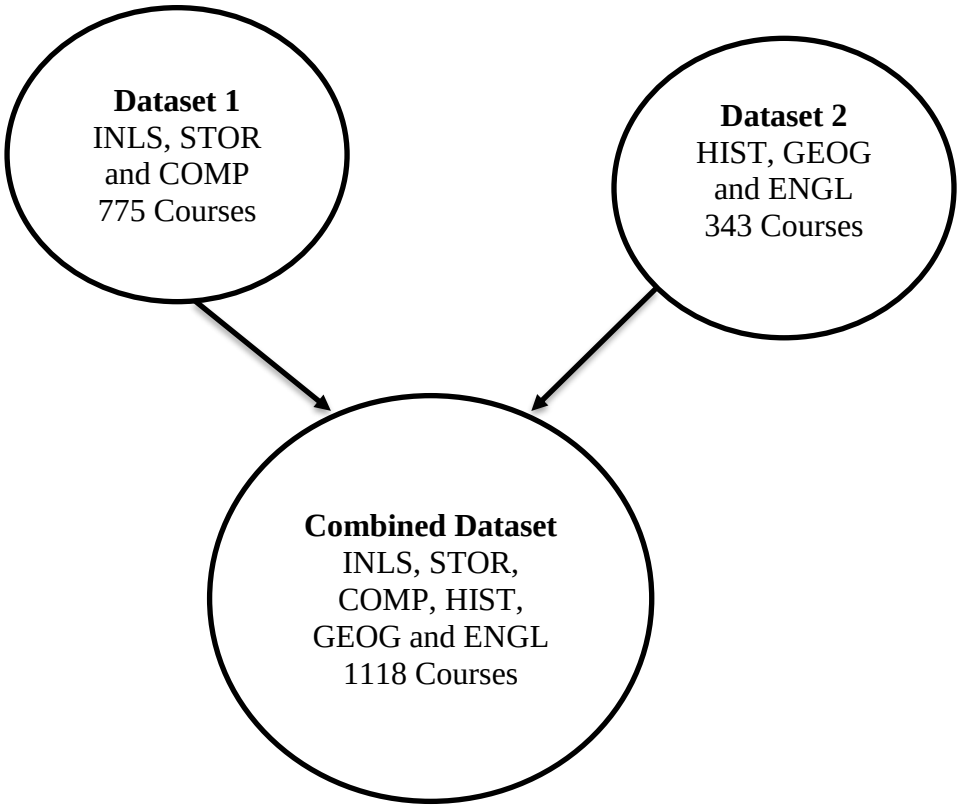

Figure 2: Schematic diagram of the three datasets

Once the three fields were scraped in a CSV format for all the datasets, the preprocessing of the dataset was performed. The preprocessing involved the following steps:

4.2 Preprocessing

a) **Tokenization:** Tokenization is one of the initial steps in Natural Language Processing. Before any real text processing is to be done, the text needs to be segmented into linguistic units such as words, punctuation, numbers, alpha-

numerics, etc. This process is called tokenization[5]. There are different Python packages available for tokenization. We used Gensim[6] for tokenization.

b) **Lemmatization:** Then lemmatization or stemming were the available options to convert the inflected words like “walking”, “walked”, “walks” to the base form as “walk”. Since the morphological analysis performed in stemming is less accurate than lemmatization (Manning, Raghavan, & Schütze, 2008), so lemmatization was performed on the datasets. We used Python package Spacy[7] for the same.

c) **Stopwords removal:** The resulting datasets from lemmatization were employed for stopword removal. Then the datasets were processed to remove the unwanted text in the description of the courses. For example, the description contained a lot of hyperlinks which were not required in topic modeling. All of these characters and stop words were removed using NLTK[8].

d) **Vectorization:** In order to ensure convenient data handling, one of the pre-processes which typically needs to be performed for recommender system development is vectorization. We used Genism, an efficient open-source Python toolkit for vector space modeling and topic modeling. Gensim can be used to convert the datasets into vector space representation which essentially is the number of occurrences of the feature terms in each document. It can be seen as a bag of words representation used to create the word-frequency tuples for each document.

[5] https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en
[6] https://tedboy.github.io/nlps/gensim_tutorial/tutorial.html
[7] https://spacy.io/api/doc
[8] https://www.nltk.org/

4.3 Model for the Recommender system

The description of each course along with its name and department abbreviation was used to find out the similar courses for recommendations. Below are a few terms which we used in our model of the recommender system:

- **Document:** One course with its description, name and department abbreviation was referred to as one document.
- **Query Document:** The name, description and department abbreviation of the course provided as an input to the recommender system through the search box was referred to as Query Document.
- **Recommended Documents:** The course description, name and department abbreviation of the courses provided by the model which were most similar to the Query Document were referred to as recommended documents.

The vector representation of the datasets generated after preprocessing contained the word-frequency tuple for each document. In our model words represented the features and their frequency was the weight of the feature. This vector representation of datasets was provided as an input to the LDA algorithm. As a result, the algorithm provided topics. The number of topics to be produced was provided as an input to the algorithm. We selected 10 as the number of topics after considering a number of options and finding the number with the best coherence, i.e., no large overlaps between the topics. We experimented with different values from 5 to 25 and empirically decided that 10 should be the number of topics in the model. This helped in getting the topics from LDA having sufficient inter-topic distance and least overlaps. The topics provided by the algorithm contained features and their respective share in the topic. The share of the features in every topic could be

considered as the fingerprint of that topic, therefore, each topic was a distinct topic with respect to the dataset. Figure 3 shows the distribution of topics over different features.

	access	activity	address	advanced	aesthetic	africa	african	age	algorithm	america
Topic0	8.676861	10.612485	0.100035	0.100006	0.100008	0.100000	0.100004	1.118563	0.100007	0.100003
Topic1	0.100053	0.100016	0.100023	4.719429	0.100000	0.100001	0.100005	0.100015	10.303619	0.100001
Topic2	1.736289	0.155685	8.405301	0.100027	0.100000	0.100003	0.100000	0.100000	5.698283	0.100000
Topic3	0.100003	0.100010	1.257832	0.100004	8.177846	0.100004	0.100009	0.456659	0.100000	1.504206
Topic4	0.100000	0.100032	0.100001	2.901106	3.021997	0.100002	2.482046	0.100015	0.100005	0.100001
Topic5	1.886782	0.100031	0.100016	29.353979	0.100001	0.100002	0.100005	0.100024	10.298084	0.100004
Topic6	0.100007	7.531734	3.670610	1.425440	0.100000	0.100000	0.100000	0.100000	0.100003	0.100003
Topic7	0.100000	0.100004	0.100000	0.100000	0.100017	16.099967	2.695621	0.100057	0.100000	7.273526
Topic8	0.100004	0.100002	2.066161	0.100008	0.100121	0.100010	26.153185	0.100010	0.100000	24.522247
Topic9	0.100001	0.100001	0.100022	0.100001	0.100009	0.100011	1.069124	8.724656	0.100000	0.100009

Figure 3: Topic distribution among different features (not all features are shown)

In Figure 3 it can be seen that each topic is represented by a list of features which are associated with a numeric value. The value is the share of that feature in the topic. If the value is larger for a feature in a topic then the feature is important for the topic. For example, if we consider topic 0, the features like “access” and “activity” have higher values, approximately 8.6 and 10.6 respectively, therefore these two features are important for topic 0. Also, if we consider the topic 8, we could see that the features like “africa” and “america” have higher values approximately 26.1 and 24.5 respectively which represents that these features are important for topic 8. So, it can be inferred that the features which are important in one topic may not be important in other topics.

The output of LDA is shown in Figure 4. The figure shows 10 topics in the form of bubbles on the left-hand side drawn in a two-dimensional space depicting the inter topic distance. The right-hand side shows the top 30 distinctive/relevant features overall which are major representatives of 10 topics. The length of the bars associated with features represents their overall frequency in the dataset. Figure 5 shows the top 30 distinctive/relevant features for topic 3. The red portion of the bars associated with the

features shows their frequency within the selected topic as compared to the overall frequency.

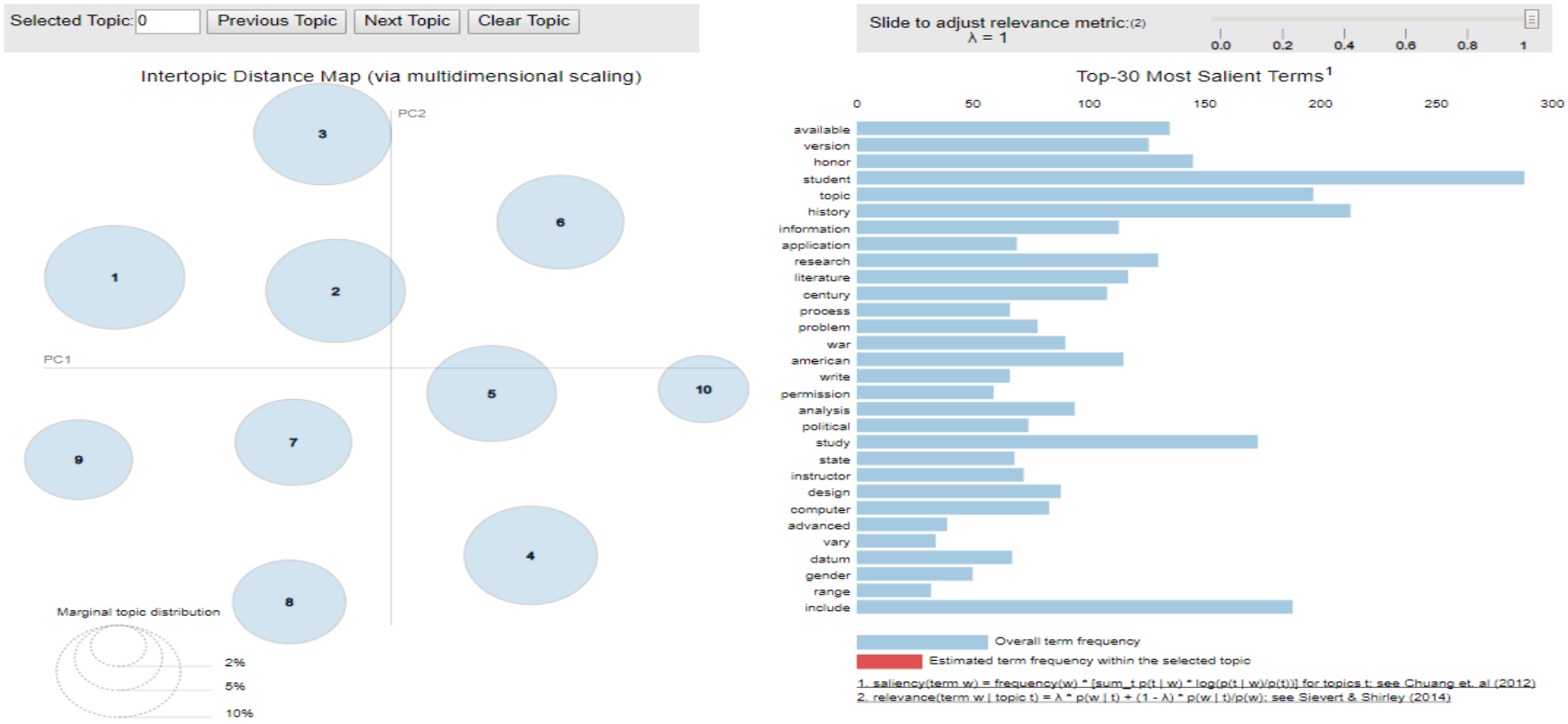

Figure 4: Showing the output with different topics as a result of LDA

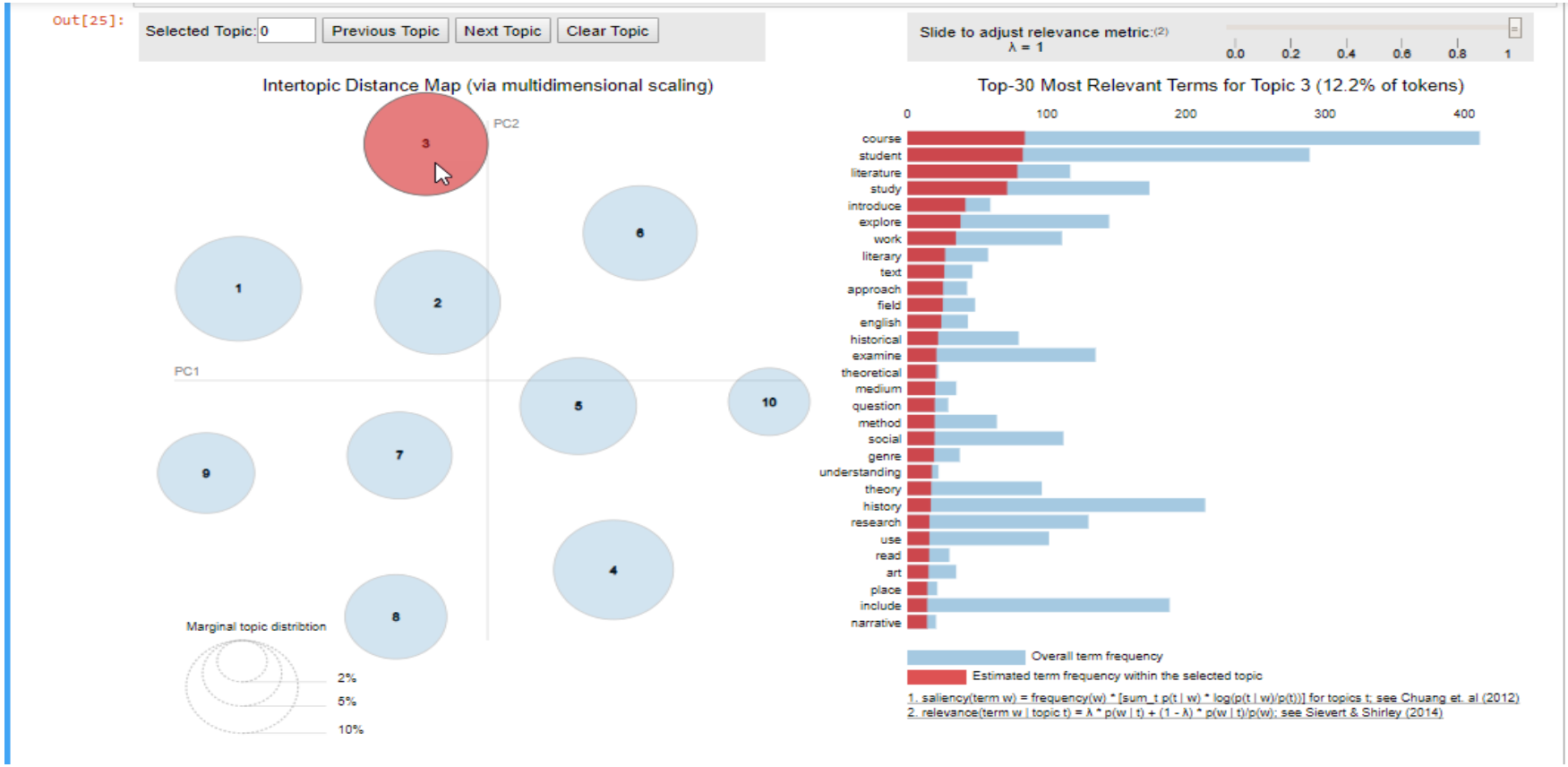

Figure 5: Showing the most relevant/distinctive words for topic 3

Because the LDA algorithm provided topics and the features with their respective shares, and each document was a collection of features, we were able to represent each document as a combination of different topics in the form of vector representation. The same vectorization was applied to the Query Document so that the similarity of the Query Document with the documents in the dataset can be calculated. Figure 6 shows the vector representation of a sample Query Document.

Sample Query Document:

0.00555873 **(T1)** + 0.00555681 **(T2)** + 0.00555758 **(T3)** + 0.00555654 **(T4)** + 0.12258361 **(T5)** + 0.12722173 **(T6)** + 0.00555698 **(T7)** + 0.09109146 **(T8)** + 0.00555678 **(T9)** + 0.62575977 **(T10)**

Figure 6: Vector representation of sample Query Document

We used the Euclidean distance to find the similarity as it is efficient when we need to find out the distance between two points in an n-dimensional space. The Query Document and the documents in the dataset were easily represented in an n-dimensional space through their vectors. The Euclidean distance between two points in an n-dimensional space is given as follows,

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}$$

where p and q are two points in an n-dimensional space.

The documents in the dataset were arranged in increasing order of distance from the Query Document. The recommended documents were selected based on their distances.

We developed a simple User Interface for the recommender system. The interface could be used by the students to input the query course. The query course would undergo preprocessing steps before the recommender system can find the similarity of the query course from the documents in the dataset. The recommender system would use the model and generate a ranked list of recommended courses.

5 Results and Discussion

5.1 Dataset 1 (STOR, INLS and COMP)

The recommender system obtained after training on Dataset 1 was tested using query courses. There were 10 query courses which were fed as an input one by one to the recommender system. For each query course, we recorded 12 recommended courses. Out of 12 recommended courses, the first or first two courses were the same as the query course depending on the case whether the query course was listed in one or two departments.

The graph in Figure 7 shows the relationship between the distances of recommended courses and their corresponding query course. In Figure 7 the horizontal axis shows ranked recommended courses and the vertical axis shows the distance of recommended courses from the query course.

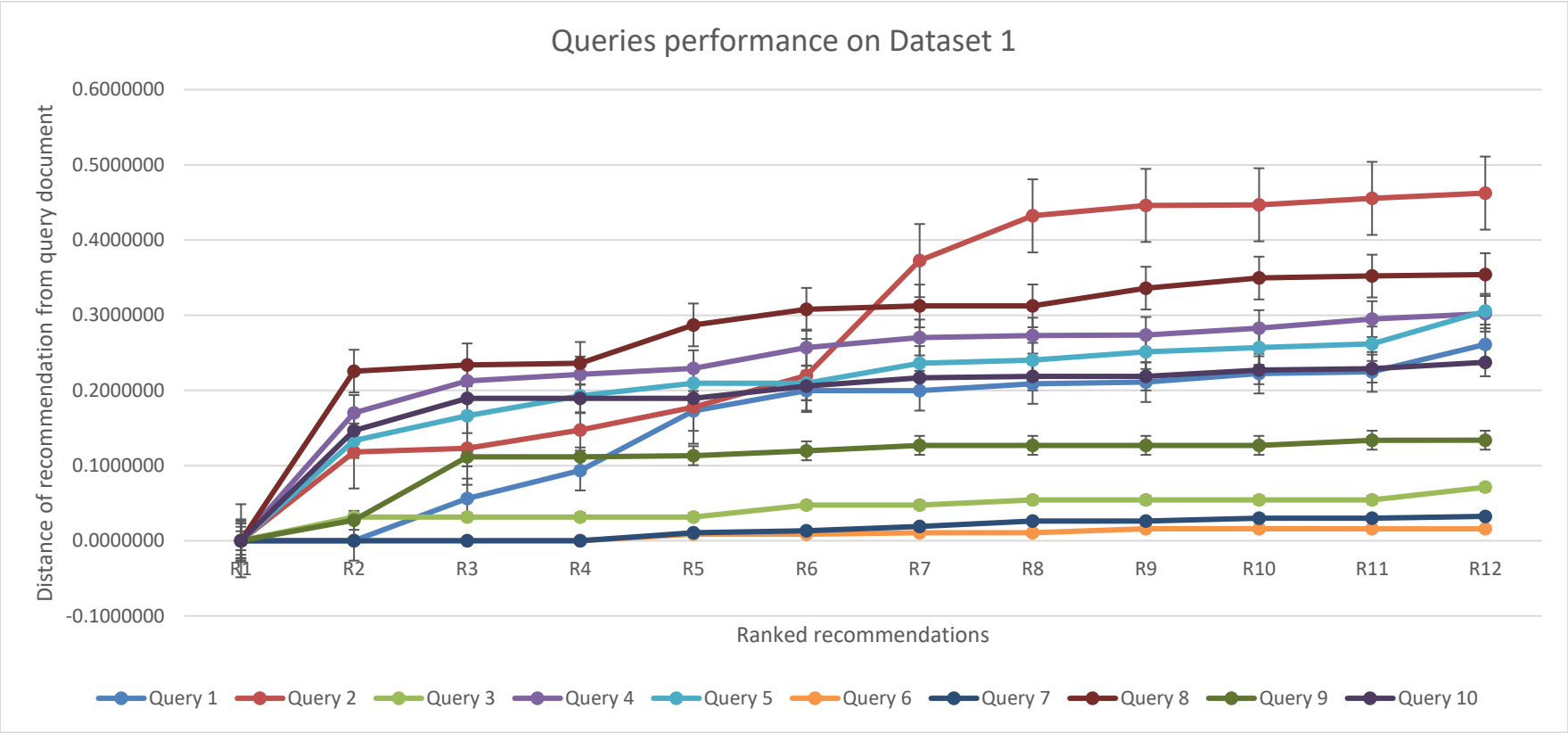

Figure 7: Queries performance on Dataset 1

The distances of recommended courses for each query course is represented with a different color. The graph also shows that after R6 the recommended courses have approximately the same distance from the query course and the recommended courses before R6 have varying distance from the query course. We defined the point at R6 as the articulation point. If we look, for instance, the recommended courses of query course 4 then after the recommended course R6 the distance between query course and recommended courses R7 to R12 is nearly constant whereas the recommended courses before R6, i.e., R1 to R5 have greater variations in distance from query course. Therefore, R6 was referred as the articulation point. We looked at the topic distribution of R1 to R12 for the query course 4. We observed that the proximity of R6 to R12 for all the topics is approximately the same whereas the proximity of R1 to R5 varies substantially for all the topics. Also, R6 to R5 were at a higher distance from the query course in comparison to R1 to R5 (Figure 7). Thus, R6 to R12 would not provide any relevant recommendations. So, we decided that only the first five courses would be presented to the user as recommendations. Figure 8 shows the proximity of R1 through R12 to topic 10 (T10) for query course 4.

Topic 10 distribution for query course 4 for R1 through R5
R1 → 0.62575977 (T10)
R2 → 0.68182716 (T10)
R3 → 0.60776517 (T10)
R4 → 0.77499174 (T10)
R5 → 0.55831455 (T10)

Topic 10 distribution for query course 4 for R6 through R12
R6 → 0.81997527 (T10)
R7 → 0.79292473 (T10)
R8 → 0.78260177 (T10)
R9 → 0.77976421 (T10)
R10→ 0.8499555 (T10)
R11→ 0.76591458 (T10)
R12→ 0.87140836 (T10)

Figure 8: Topic 10 distribution for R1 through R12 of query course 4

In order to analyze the cause for the high distance of the recommended courses, we looked at the query course 4 in Figure 9 and inspected its R12 (12^{th} recommended course). The topic distribution of the query course 4 and its R12 helped to illustrate how far the descriptions of the courses were with respect to each topic (T1 to T10). Figure 9 shows the course description and topic distribution of query course 4 and its R12. In Figure 9, T1 to T10 denote the 10 topics generated by the LDA algorithm and the number preceding each topic denotes the proximity of the topic to the course, i.e., smaller the number strongly the course is related to the topic. It can be seen that R12 was further away with respect to all the topics as compared to the query course because the words in the description of R12 were not in the feature set except the three words in bold. This drifted R12 away from all the topics and hence ended up being farthest from the query course with respect to the distance.

Course description: query course 4

"An introduction to information visualization through reading current literature and studying exemplars. The course reviews information visualization techniques provides a framework for identifying the need for information visualization and emphasizes interactive electronic visualizations that use freely available tools. Students will construct several visualizations. No programming skills are required."

R12:

"Explores the evolution implications and complications of **social** media in **multiple** spheres of life including sociality community politics power and inequality education and **information** from theoretical and empirical perspectives."

Topic distribution: query course 4

[0.00555873 **(T1)**, 0.00555681 **(T2)**, 0.00555758 **(T3)**, 0.00555654 **(T4)**, 0.12258361 **(T5)**, 0.12722173 **(T6)**, 0.00555698 **(T7)**, 0.09109146 **(T8)**, 0.00555678 **(T9)**, 0.62575977 **(T10)**]

R12

[0.0142863 **(T1)**, 0.01428744 **(T2)**, 0.01428689 **(T3)**, 0.01429411 **(T4)**, 0.01428571 **(T5)**, 0.01429148 **(T6)**, 0.01428572 **(T7)**, 0.01428603 **(T8)**, 0.01428797 **(T9)**, 0.87140836 **(T10)**]

Figure 9: Description and Topic distribution of query course 4 and its R12 (12^{th} recommended course)

We also found from Figure 7 that all the recommended courses for query course 3, 6 and 7 were nearly at the same distance. In order to analyze this behavior of these query courses we looked at the description of the query course 3 and for its recommended courses R2, R3, R4 and R12. We examined the words of R2, R3, R4 and R12. We tried to find these words in the top 30 distinctive words of all 10 topics and determined which topic was influencing the most. We found that the description of query course 3 and its R2, R3, R4 and R12 contained at least 7-8 words which belonged to the top 30 distinctive/relevant words in topic 1. This made the courses to be in close proximity to topic 1. We also found that besides those 7-8 words no other words from the description of query course 3 and its R2, R3, R4 and R12 were present in the feature set with the exception of "statistical" "develop" and "structure". Therefore, the proximity of these courses to the topics other than topic 1 was decided by the contribution of these 7-8 words. This made the topic distribution of query course and recommended courses to be nearly the same, which in turn resulted in producing all the recommended courses having approximately zero distance from the query course. Figure 10 shows the top 30 distinctive words of topic 1 and Figure 11 shows the description of query course 3 and its R2, R3, R4 and R12. It can be noted that in Figure 11 the italicized words are those 7-8 words of the description of query course 3 and its R2, R3, R4 and R12 which are present in the top 30 distinctive/relevant words of topic 1 in Figure 10. The exception words "statistical" "develop" and "structure" are in bold in Figure 11 which are not present in the list of top 30 distinctive words for topic 1 and hence are not related to the topic 1.

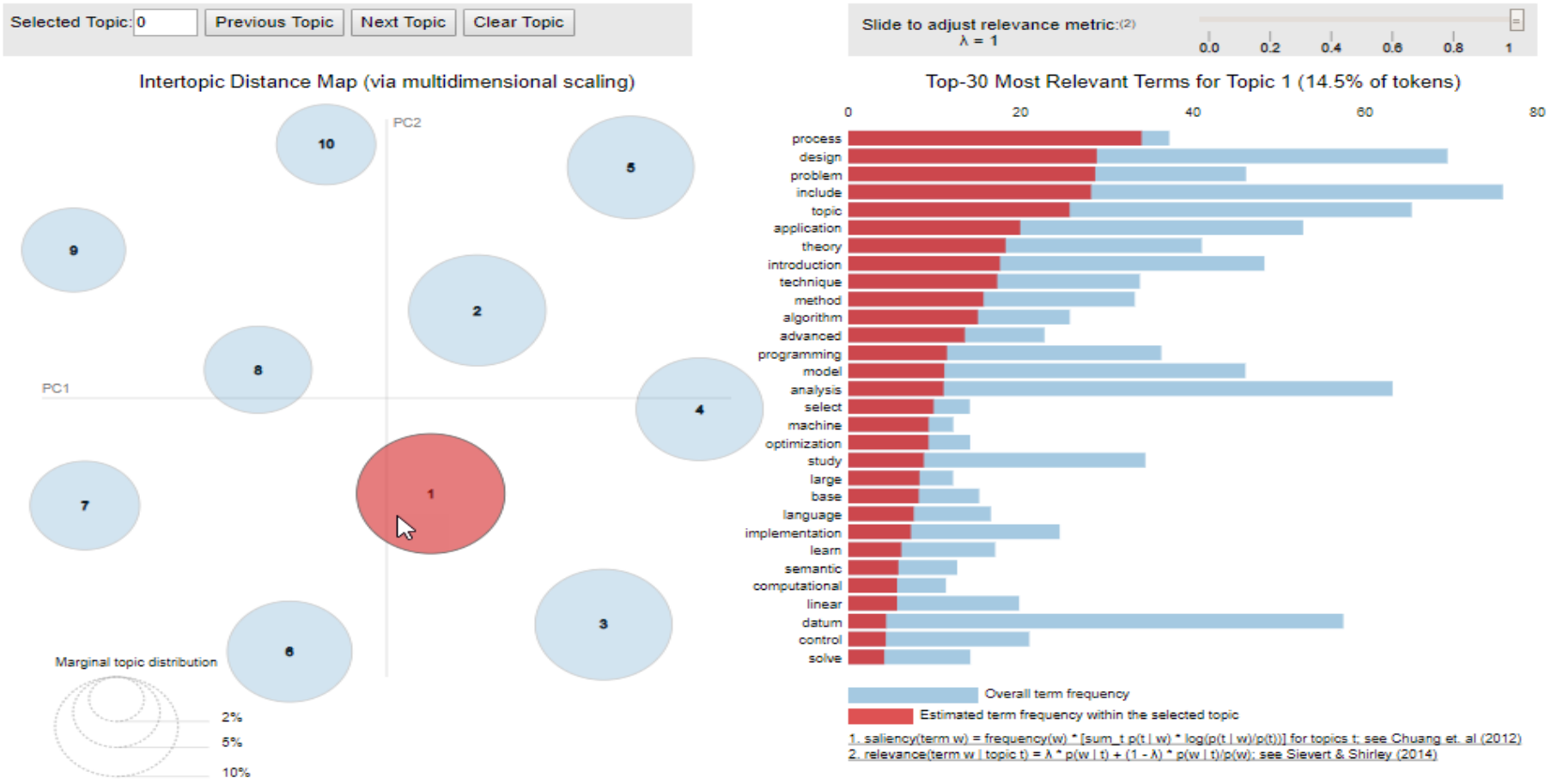

Figure 10: Showing top 30 distinctive/relevant words in topic 1

Query course 3

"*Introduction* to Markov chains Poisson *process* continuous-time Markov chains renewal *theory*. *Applications* to queueing systems inventory and reliability with **emphasis** on systems *modeling design* and *control*."

R2 STOR 113 Decision Models for Business and Economics

"An *introduction* to multivariable quantitative *models* in economics. Mathematical *techniques* for formulating and *solving optimization* and equilibrium *problems* will be **developed** *including* elementary *models* under uncertainty."

R3 STOR 756 Design and Robustness

"*Introduction* to experimental *design including* classical *designs* industrial *designs* optimality and sequential *designs*. *Introduction* to robust **statistical** *methods*; bootstrap cross-validation and resampling."

R4 STOR 712 Mathematical Programming I

"*Advanced topics* from mathematical *programming* such as geometry of *optimization* parametric *analysis* finiteness and convergence proofs and *techniques* for *large*-scale and specially **structured** *problems*."

R12 INLS 512 Applications of Natural Language Processing

"*Study* of *applications* of natural language *processing techniques* and the representations and *processes* needed to support them. *Topics include* interfaces text retrieval *machine* translation speech *processing* and text generation."

Figure 11: Description of query course 3 and its R2, R3, R4 and R12

5.2 Dataset 2 (HIST, GEOG and ENGL)

The recommender system obtained after training on Dataset 2 was also tested using query courses. The same method was followed as in Dataset 1 for testing Dataset 2. We used the same method as in Dataset 1 to plot the data points in Dataset 2. Figure 12 shows queries performance on Dataset 2.

The distances of recommended courses for each query course is represented with a different color. The articulation point can be seen at R6 in Figure 12. Like the Dataset 1, in Dataset 2 as well, the courses after the articulation point did not prove to be relevant recommendations. Hence, we decided to take only the first five recommended courses to present as recommendations.

It can be seen in Figure 12 that unlike Dataset 1, the Dataset 2 had only one query course for which the recommended courses were nearly at the same distance from the query course. The query course 9 showed this kind of behavior. We analyzed the query course 9 and its recommended courses R2, R3, R4 and R12. First, we examined the words of R2, R3 and R12. We matched these words with the top 30 distinctive/relevant words of all the 10 topics of the recommender system. We found on an average 9 words in query course 9 and its R2, R3, R4 and R12 were present in all the 10 topics. This made the courses to be in close proximity of all the topics in approximately similar ratio. Hence the distance of all the recommended courses from query course was similar which made all the recommended courses to lie approximately on a flat line in the graph of Figure 12.

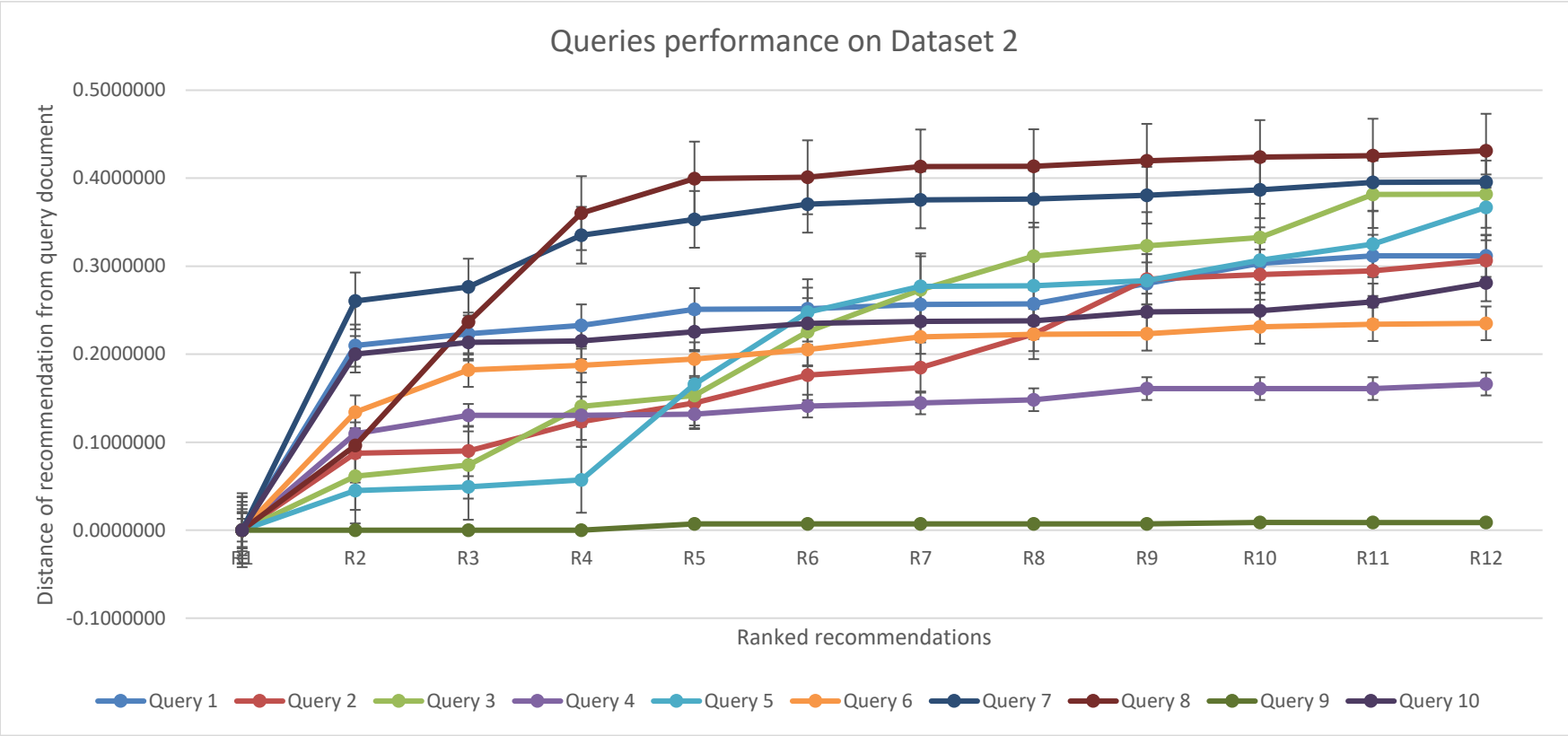

Figure 12: Queries performance on Dataset 2

Figure 13 shows the description of query course 9 and its R2, R3, R4 and R12. The italicized words in Figure 13 are the words which are present in all of the ten topics of the recommender system.

Description

Query course 9:

Required preparation a *working* knowledge of *Old English*. The translation and interpretation of *Old English poetry including works* such as The Wanderer The Seafarer Deor The Dream of the Rood and Beowulf.

R2: Hist 714 Introductory Colloquium in the History of Latin America since 1810

Directed readings on Latin *American history* in the National *Period*; *required* for students entering the *field*.

R3: ENGL 100 Basic Writing

Required for incoming *students* with SAT I *Writing* scores of 460 or lower. Provides frequent *practice* in *writing* from *short* paragraphs to longer papers *focusing* on *analysis* and argument. *Workshop* format.

R4: HIST 726 Introductory Colloquium in United States History to 1788

Directed readings on *American history* from the precolonial *period* through the *American* Revolution; *required* for students entering the *field*.

R12: HIST 534 The African Diaspora

A comparative *examination* of the movements experiences and contributions of *Africans* and people of *African* descent from the *period* of the Atlantic slave trade to the *present*.

Figure 13: Description of query course 9 and its R2, R3, R4 and R12

5.3 Combined Dataset (STOR, INLS, COMP, HIST, GEOG and ENGL):

The third dataset, a combination of Dataset 1 and Dataset 2, was used to train the recommender system. To study the behavior when query courses of Dataset 1 and 2 were provided as input to the recommender system trained on the Combined Dataset, we used the same query courses of Dataset 1 and 2. So, we analyzed a total of 20 query courses on this recommender system. Similar to the Dataset 1 and 2, twelve recommended courses were recorded for each query course. Out of 12 courses the first or first two courses matched with the query course depending on the case whether the query course was listed in one or two departments.

Similar to the case in Dataset 1 and 2 we plotted the graph in Figure 14 for queries performance on the Combined Dataset. The distances of recommended courses for each query course is represented with a different color. In Figure 14 the articulation point is at R6.

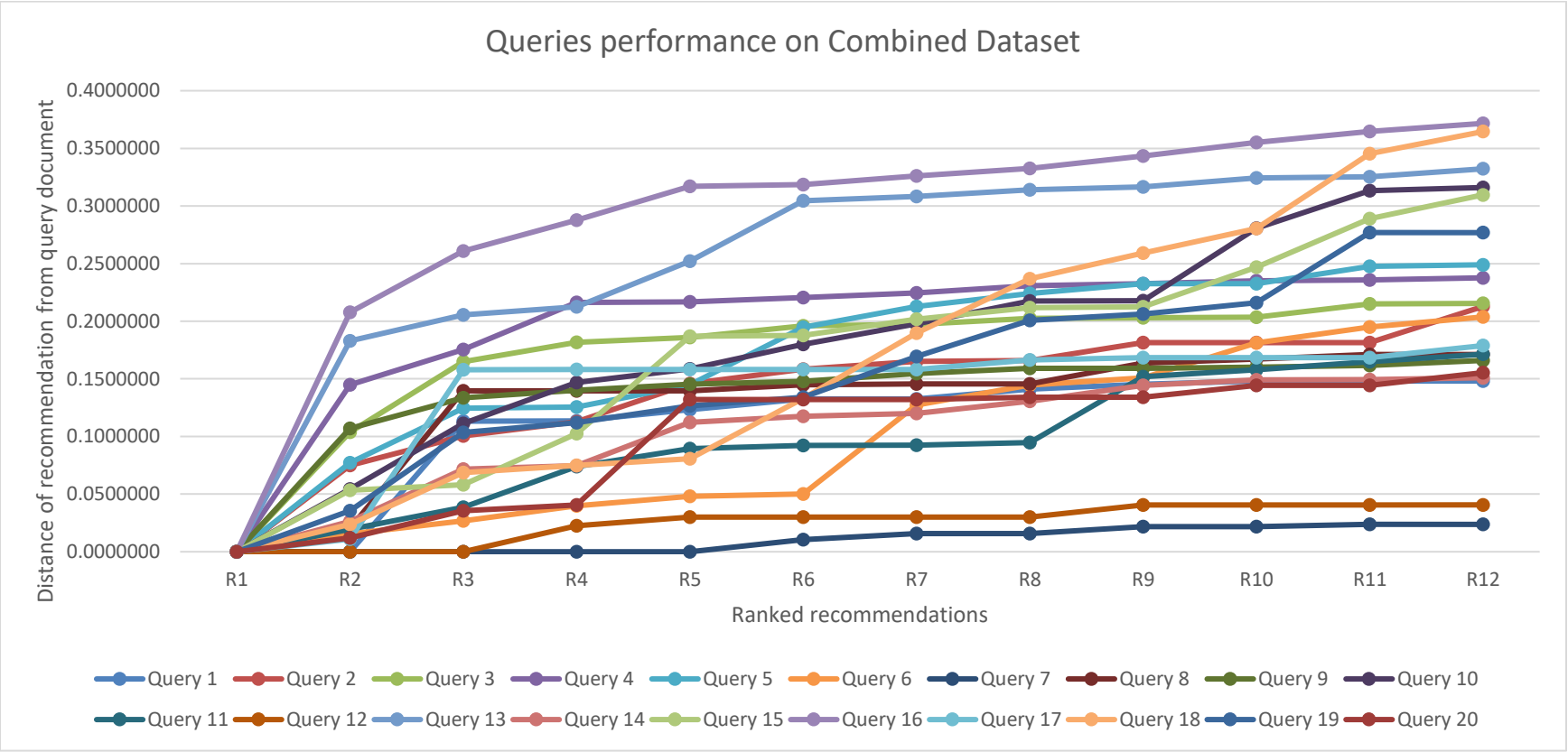

Figure 14: Queries performance on Combined Dataset

We compared the courses before and after articulation point and found that the quality of recommended courses after articulation point was getting degraded. For example, if the query course was from technology or Mathematics field then after the articulation point there were many courses from the arts field. So, we decided to take only the first five recommended courses to present as recommendations. Figure 15 shows the recommended courses for INLS 578 Protocols and Network Management. The courses from arts are in bold from R6.

Query course - INLS 578 Protocols and Network Management		
R1	INLS 578	Protocols and Network Management (same as query course)
R2	INLS 624	Policy-Based Data Management
R3	INLS 586	Project Management
R4	INLS 752	Digital Preservation and Access
R5	ENGL 307	Studies in Fiction and Poetry: Stylistics
R6	INLS 765	Information Technology Foundations for Managing Digital Collections
R7	**HIST 493**	**Internship in History**
R8	INLS 465	Understanding Information Technology for Managing Digital Collections
R9	**GEOG 704**	**Communicating Geography**
R10	**ENGL 317**	**Networked Composition**
R11	**ENGL 706**	**Rhetorical Theory and Practice**
R12	INLS 750	Introduction to Digital Curation

Figure 15: Recommended courses for INLS 578 Protocols and Network Management

For each of the 20 query courses we compared their recommended courses obtained in the case of Dataset 1 or 2 with that of the recommended courses obtained in the case of Combined Dataset. We found that for 5 query courses in the Dataset 1, their recommended courses matched with the recommended courses generated in the case of Combined Dataset. However, no recommended courses for the query courses from Dataset 2 were matched with the recommended courses generated in the case of Combined Dataset. In Figure 16 the horizontal axis shows the query courses and the vertical axis shows the count of matched courses for all the query courses. It can be seen in Figure 16 that there is a great variation in the count of matched courses for the query courses 1 to 10 which are from the Dataset 1 and the count of matched courses for the query courses 11 to 20 is zero.

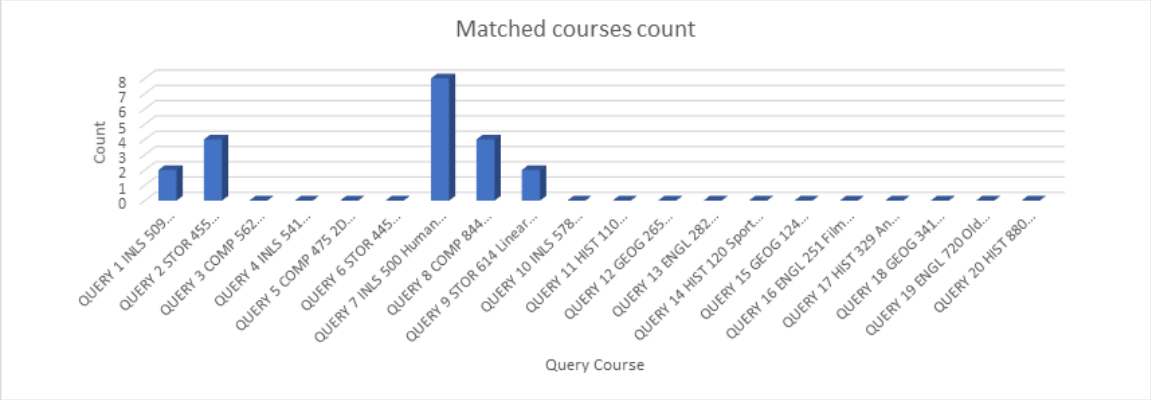

Figure 16: Matched course count for the queries between combined and individual datasets

We analyzed the query courses which had zero match and which had a considerable number of matched recommended courses. To analyze the query courses when they had zero match for the recommended courses, we looked at the topics developed in the case of Dataset 2 and Combined Dataset. We looked at the top distinguishing/relevant words in topic 5 (chosen arbitrarily) developed in case of Dataset 2 and the contribution of those

words in that topic. We looked for the same words in topic 5 developed in case of Combined Dataset. We found that the words which were distinguishing/relevant words of topic 5 in Dataset 2 were no longer the distinguishing/relevant words of topic 5 in the Combined Dataset. This reflected that the topics developed in the Combined Dataset had different distinguishing words as compared to the Dataset 2. This was possible since the contribution of a word in a topic depended on the overall word list of the dataset. Therefore, the difference in the words associated with the topics and their contribution resulted in different course recommendations.

Figure 17 shows the top distinguishing/relevant words of topic 5 in Dataset 2 and Combined Dataset. The number to the right of each word shows their share in the topic 5.

Topic 5: Dataset 2

23.42232462 write, 7.287268113 art, 18.46869932 contemporary, 7.884062302 cultural, 7.746595757 environment, 19.74341978 explore, 20.15437652 field, 18.94350835 focus, 17.64226138 geographic, 18.47048471 geography

Topic 5: Combined Dataset

12.32652259 write, 6.841830192 art, 0.100001191 contemporary, 1.293170758 cultural, 0.100017514 environment, 8.274840325 explore, 0.100019252 field, 0.100031085 focus, 0.100003769 geographic, 2.858061006 geography

Figure 17: Contribution of top distinguishing/relevant words of topic 5 in Dataset 2 and their contribution in Combined Dataset

To analyze the query courses with considerable matched recommended courses from Combined Dataset we chose query course 7, i.e., INLS 500 Human Information Interactions, which had the highest number of matched recommended courses from the

Combined Dataset. There were 8 matched recommended courses for query 7. Figure 18 contains the description of query course 7 and GEOG 123 Cultural Geography, one of the recommended courses.

query course 7: INLS 500 Human Information Interactions

Description:

"The behavioral and cognitive activities of those who interact with information with emphasis on the role of information mediators. How information needs are recognized and resolved; use and dissemination of information"

Recommended course:

GEOG 123 Cultural Geography

Description:

"How population environment and human culture is expressed in technology and organization interact over space and time"

Figure 18: Description of query course 7 and recommended course GEOG 123

Figure 19 shows the recommended courses for query course 7 in case of Dataset 1 and Combined Dataset. The matched recommended courses are in bold.

Combined Dataset		**Dataset 1**	
INLS 500	Human Information Interactions	INLS 500	Human Information Interactions
INLS 202	Retrieval and Organizing Systems	**INLS 754**	**Access**
INLS 151	Retrieving and Analyzing Information	INLS 703	Science Information
GEOG 123	Cultural Geography	STOR 390	Special Topics in Statistics and Operations Research
GEOG 440	Earth Surface Processes	**INLS 704**	**Humanities Information**
STOR 642	Stochastic Models in Operations Research II	**INLS 748**	**Health Sciences Environment**
INLS 754	**Access**	INLS 501	Information Resources and Services
INLS 704	**Humanities Information**	INLS 705	Health Sciences Information
INLS 748	**Health Sciences Environment**	**INLS 702**	**Social Science Information**
INLS 702	**Social Science Information**	STOR 893	Special Topics
GEOG 442	River Processes	INLS 660	Social Media and Society: A Theoretical and Empirical Overview
COMP 455	Models of Languages and Computation	INLS 701	Information Retrieval Search Strategies

Figure 19: A list of recommended courses for query course 7 on Combined Dataset and Dataset 1

We found that the recommended courses for query course 7 in Combined Dataset included GEOG 123 “Cultural Geography”. Looking at the description of GEOG 123 “Cultural Geography” in Figure 19 it can be seen that this course talks about the technological aspect for population environment and human culture. So, the course might be of interest for the students from the technology field entering the “Human Information Interaction” as a query course. This is a good recommendation which a student from technology field might not have considered to register.

6 Conclusion and Limitations

In this paper, we presented a model to build a recommendation system to recommend courses to the students. Such a system could be of great help to the students as a shorter list of recommended courses would help students find courses of their interest by investing lesser time in browsing through the whole course list of different departments. Hence, the system can save a lot of time while registering courses. To build the recommender system we used three datasets. The first dataset contained technology courses, the second dataset contained courses from arts field and the third dataset called Combined Dataset was a combination of the first two datasets. We trained the recommender system on each of the three datasets. We analyzed the recommended courses generated through the recommender system trained on Combined Dataset to find whether a varied dataset containing courses from diverse fields can help recommend courses different than the expected courses and provide serendipitous recommendations. We found that after a certain point on the list of recommended courses the quality of recommendations started deteriorating. After this point the courses tend to lose the variation among themselves and were essentially a group of similar courses from a single dataset. These courses were very dissimilar from the query course and hence were removed from the list of recommended courses for presentation. We also found that the topics developed in Combined Dataset were very dissimilar than the topics of first two

datasets. Therefore, we got different recommended courses for the same query course on the Combined Dataset as compared to Dataset 1 or Dataset 2. We demonstrated that the recommender system trained on disparate set of courses in the Combined Dataset could provide useful recommendations which were unexpected. There were some limitations to the system.

- One of the fields used for finding recommended courses was course description. The course description used here was from the university catalog website where the descriptions were not very detailed and were generic a lot of times. To build a robust recommender system we need to train the system on a rich text with sufficient amount of details available.
- The data used for the recommender system was static, but the course details and description tend to get updated very frequently in academic settings. So, the system would need to be trained regularly on the updated data in order to provide the relevant recommendations.
- Our recommender system does not take the student's interest into account which is important in providing recommendations to the students. Our system recommends only on the basis of the description of the course, department name and course name.

7 Future Work

The recommender system developed in this paper can be improved in a lot of ways. Below are some of the ways which could be used to enhance the efficiency of our recommender system.

- The student's resume could provide details about the student's academic background and details which can be used to develop a personalized dashboard for the student. In this way, students will find recommended courses which are best suited for their profile.
- The course syllabus can be used to enrich the text in the training dataset. The detailed description in the syllabus can help describe courses effectively which can be used to generate recommended courses. However, the course syllabus is lengthy and consists of a lot of information which could pose a problem to search for useful keywords and use them in the training dataset.
- Student's review about the courses they have taken could be provided to the recommender system by the students and this information can be used by the recommender system in order to generate the recommended courses.
- The recommender system can be improved by taking into account the courses already taken by the student and excluding them from popping up in the recommended courses. The course history of the student can be used by the recommender system for this purpose.

Reference

Adomavicius, Gediminas, Sankaranarayanan, R., Sen, S., & Tuzhilin, A. (2005). Incorporating contextual information in recommender systems using a multidimensional approach. ACM Transactions on Information Systems, 23(1), 103–145. https://doi.org/10.1145/1055709.1055714

Adomavicius, G., & 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(6), 734–749. https://doi.org/10.1109/TKDE.2005.99

Adomavicius, Gediminas, & Tuzhilin, A. (2015). Context-Aware Recommender Systems. In Recommender Systems Handbook (pp. 191–226). https://doi.org/10.1007/978-1-4899-7637-6_6

Aher, S. B., & Lobo, L.M.R.J. (2012). Combination of Clustering, Classification & Association Rule based Approach for Course Recommender System in E-learning. International Journal of Computer Applications, 39(7), 8–15. https://doi.org/10.5120/4830-7087

Al-Badarenah, A., & Alsakran, J. (2016). An Automated Recommender System for Course Selection. International Journal of Advanced Computer Science and Applications, 7(3), 1166-1175. https://doi.org/10.14569/IJACSA.2016.070323

Burke, R. (2000). Knowledge-Based Recommender Systems. Encyclopedia of Library and Information Systems, 69(Supplement 32), 175-186, 2000. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.6029

Burke, R. (2002). Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction, 12(4), 331–370. https://doi.org/10.1023/A:1021240730564

Feng, S., Cao, J., Wang, J., & Qian, S. (2017). Recommendations Based on Comprehensively Exploiting the Latent Factors Hidden in Items' Ratings and Content. ACM Transactions on Knowledge Discovery from Data, 11(3), 1–27. https://doi.org/10.1145/3003728

Gorla, A., Tavecchia, I., Gross, F., & Zeller, A. (2014). Checking app behavior against app descriptions. Proceedings of the 36th International Conference on Software Engineering - ICSE 2014, 1025–1035. https://doi.org/10.1145/2568225.2568276

Grewal D.S., Kaur K. (2015). Developing an Intelligent Recommendation System for Course Selection by Students for Graduate Courses. Business and Economics Journal, 7(2), 1–9. https://doi.org/10.4172/2151-6219.1000209

Gupta, S., Hanson, C., Gunter, C. A., Frank, M., Liebovitz, D., & Malin, B. (2013). Modeling and detecting anomalous topic access. 2013 IEEE International Conference on Intelligence and Security Informatics, 100–105. https://doi.org/10.1109/ISI.2013.6578795

Herlocker, J. L., Konstan, J. A., Borchers, A., & Riedl, J. (1999). An algorithmic framework for performing collaborative filtering. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR ’99, 230–237. https://doi.org/10.1145/312624.312682

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems, 22(1), 5–53. https://doi.org/10.1145/963770.963772

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix Factorization Techniques for Recommender Systems. Computer, 42(8), 30–37. https://doi.org/10.1109/MC.2009.263

Lin, W., Alvarez, S. A., & Ruiz, C. (2002). Efficient Adaptive-Support Association Rule Mining for Recommender Systems. Data Mining and Knowledge Discovery, 6(1), 83–105. https://doi.org/10.1023/A:1013284820704

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/CBO9780511809071

McNee, S. M., Riedl, J., & Konstan, J. A. (2006). Being accurate is not enough. CHI ’06 Extended Abstracts on Human Factors in Computing Systems - CHI EA ’06, 1097-1101. https://doi.org/10.1145/1125451.1125659

Pandita, R., Xiao, X., Yang, W., Enck, W., & Xie, T. (2013). {WHYPER}: Towards Automating Risk Assessment of Mobile Applications. 22nd {USENIX} Security Symposium ({USENIX} Security 13) (pp. 527–542). https://www.usenix.org/conference/usenixsecurity13/technical-sessions/presentation/pandita

Parameswaran, A., Venetis, P., & Garcia-Molina, H. (2011). Recommendation systems with complex constraints. ACM Transactions on Information Systems, 29(4), 1–33. https://doi.org/10.1145/2037661.2037665

Pazzani, M. J., & Billsus, D. (2007). Content-Based Recommendation Systems. In the Adaptive Web (pp. 325–341). https://doi.org/10.1007/978-3-540-72079-9_10

Rashid, A. M., Albert, I., Cosley, D., Lam, S. K., McNee, S. M., Konstan, J. A., & Riedl, J. (2002). Getting to know you: learning new user preferences in recommender systems. Proceedings of the 7th International Conference on Intelligent User Interfaces - IUI ’02, 127-134. https://doi.org/10.1145/502716.502737

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). GroupLens. Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work - CSCW ’94, 175–186. https://doi.org/10.1145/192844.192905

Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to Recommender Systems Handbook. In Recommender Systems Handbook (pp. 1–35). https://doi.org/10.1007/978-0-387-85820-3_1

Sarwar, B., Karypis, G., Konstan, J., & Reidl, J. (2001). Item-based collaborative filtering recommendation algorithms. Proceedings of the Tenth International Conference on World Wide Web - WWW ’01, 285–295. https://doi.org/10.1145/371920.372071

Shaheen, M., & Shahbaz, M. (2017). An Algorithm of Association Rule Mining for Microbial Energy Prospection. Scientific Reports 7, Article number 46108. https://doi.org/10.1038/srep46108

Shardanand, U., & Maes, P. (1995). Social information filtering. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI ’95, 210–217. https://doi.org/10.1145/223904.223931

Torres, R., McNee, S. M., Abel, M., Konstan, J. A., & Riedl, J. (2004). Enhancing digital libraries with TechLens+. Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries - JCDL ’04, 228-236. https://doi.org/10.1145/996350.996402

Appendix

Recommended courses on Dataset 1. The first course is R1 and the last course is R12

QUERY COURSE 1 INLS 509 - Information Retrieval

[('COMP 487', ' Information Retrieval'), ('INLS 509', ' Information Retrieval'), ('INLS 758', ' International and Cross-Cultural Perspectives for Information Management'), ('INLS 786', ' Marketing of Information Services'), ('COMP 787', ' Visual Perception'), ('INLS 318', ' Human Computer Interaction'), ('INLS 418', ' Human Factors in System Design'), ('INLS 151', ' Retrieving and Analyzing Information'), ('INLS 723', ' Database Systems III: Advanced Databases'), ('INLS 620', ' Web Information Organization'), ('INLS 581', ' Research Methods Overview'), ('INLS 745', ' Instruction for Youth in School and Public Libraries')]

QUERY COURSE 2 STOR 455- Methods of Data Analysis

[('STOR 455', ' Methods of Data Analysis'), ('STOR 754', ' Time Series and Multivariate Analysis'), ('STOR 64', ' First-Year Seminar: A Random Walk down Wall Street'), ('STOR 855', ' Subsampling Techniques'), ('STOR 155', ' Introduction to Data Models and Inference'), ('STOR 665', ' Applied Statistics II'), ('INLS 781', ' Proposal Development'), ('STOR 61', ' First-Year Seminar: Statistics for Environmental Change'), ('INLS 572', ' Web Development I'), ('COMP 723', ' Software Design and Implementation'), ('STOR 66', ' First-Year Seminar: Visualizing Data'), ('INLS 721', ' Cataloging Theory and Practice')]

QUERY COURSE 3 'COMP 562- Introduction to Machine Learning

[('COMP 562', ' Introduction to Machine Learning'), ('COMP 782', ' Motion Planning in Physical and Virtual Worlds'), ('COMP 555', ' Bioalgorithms'), ('STOR 824', ' Computational Methods in Mathematical Programming'), ('STOR 713', ' Mathematical Programming II'), ('COMP 730', ' Operating Systems'), ('COMP 777', ' Optimal Estimation in Image Analysis'), ('COMP 752', ' Mechanized Mathematical Inference'), ('COMP 455', ' Models of Languages and Computation'), ('STOR 822', ' Topics in Discrete Optimization'), ('COMP 822', ' Topics in Discrete Optimization'), ('COMP 825', ' Logic Programming')]

QUERY COURSE 4 INLS 541- Information Visualization

[('INLS 541', ' Information Visualization'), ('STOR 56', ' First-Year Seminar: The Art and Science of Decision Making in War and Peace'), ('INLS 718', ' User Interface Design'), ('INLS 513', ' Resource Selection and Evaluation'), ('INLS 626', ' Introduction to Big Data and NoSQL'), ('STOR 790', ' Operations Research and Systems Analysis Student Seminar'), ('STOR 894', ' Special Topics at SAMSI'), ('INLS 842', ' Seminar in Popular Materials in Libraries'), ('INLS 73', ' First-Year Seminar: Smart Cities'), ('INLS 609', ' Experimental Information Retrieval'), ('INLS 697', ' Information Science Capstone'), ('INLS 660', ' Social Media and Society: A Theoretical and Empirical Overview')]

QUERY COURSE 5 COMP 475- 2D Computer Graphics
[('COMP 475', ' 2D Computer Graphics'), ('COMP 520', ' Compilers'), ('COMP 720', ' Compilers'), ('COMP 735', ' Distributed and Concurrent Algorithms'), ('STOR 853', ' Nonparametric Inference:

Smoothing Methods'), ('COMP 755', ' Machine Learning'), ('COMP 724', ' Programming Languages'), ('STOR 833', ' Time Series Analysis'), ('STOR 734', ' Stochastic Processes'), ('INLS 623', ' Database Systems II: Intermediate Databases'), ('COMP 721', ' Database Management Systems'), ('INLS 382', ' Information Systems Analysis and Design')]

QUERY COURSE 6 STOR 445- Stochastic Modeling

[('STOR 445', ' Stochastic Modeling'), ('STOR 113', ' Decision Models for Business and Economics'), ('STOR 756', ' Design and Robustness'), ('STOR 712', ' Mathematical Programming I'), ('STOR 565', ' Machine Learning'), ('STOR 722', ' Integer Programming'), ('COMP 662', ' Scientific Computation II'), ('COMP 560', ' Artificial Intelligence'), ('COMP 781', ' Robotics'), ('STOR 763', ' Statistical Quality Improvement'), ('COMP 486', ' Applications of Natural Language Processing'), ('INLS 512', ' Applications of Natural Language Processing')]

QUERY COURSE 7 INLS 500- Human Information Interactions

[('INLS 500', ' Human Information Interactions'), ('INLS 754', ' Access'), ('INLS 703', ' Science Information'), ('STOR 390', ' Special Topics in Statistics and Operations Research'), ('INLS 704', ' Humanities Information'), ('INLS 748', ' Health Sciences Environment'), ('INLS 501', ' Information Resources and Services'), ('INLS 705', ' Health Sciences Information'), ('INLS 702', ' Social Science Information'), ('STOR 893', ' Special Topics'), ('INLS 660', ' Social Media and Society: A Theoretical and Empirical Overview'), ('INLS 701', ' Information Retrieval Search Strategies')]

QUERY COURSE 8 COMP 844- Advanced Design of VLSI Systems

[('COMP 844', ' Advanced Design of VLSI Systems'), ('COMP 763', ' Semantics and Program Correctness'), ('INLS 793', ' Health Informatics Practicum'), ('COMP 735', ' Distributed and Concurrent Algorithms'), ('COMP 744', ' VLSI Systems Design'), ('INLS 672', ' Web Development II'), ('STOR 734', ' Stochastic Processes'), ('COMP 721', ' Database Management Systems'), ('COMP 116', ' Introduction to Scientific Programming'), ('COMP 475', ' 2D Computer Graphics'), ('COMP 101', ' Fluency in Information Technology'), ('COMP 991', ' Reading and Research')]

QUERY COURSE 9 STOR 614- Linear Programming

[('STOR 614', ' Linear Programming'), ('COMP 764', ' Monte Carlo Method'), ('STOR 757', ' Bayesian Statistics and Generalized Linear Models'), ('INLS 550', ' History of the Book and Other Information Formats'), ('COMP 776', ' Computer Vision in our 3D World'), ('STOR 635', ' Probability'), ('INLS 755', ' Archival Appraisal'), ('STOR 62', ' First-Year Seminar: Probability and Paradoxes'), ('COMP 180', ' Enabling Technologies'), ('STOR 471', ' Long-Term Actuarial Models'), ('STOR 52', ' First-Year Seminar: Decisions, Decisions, Decisions'), ('STOR 60', ' First-Year Seminar: Statistical Decision-Making Concepts')]

QUERY COURSE 10 INLS 578- Protocols and Network Management

[('INLS 578', ' Protocols and Network Management'), ('STOR 767', ' Advanced Statistical Machine Learning'), ('INLS 89', ' First-Year Seminar: Special Topics'), ('STOR 89', ' First-Year Seminar: Special Topics'), ('INLS 490', ' Selected Topics'), ('INLS 576', ' Distributed Systems and Administration'), ('INLS 890', ' Advanced Special Topics'), ('STOR 744', ' Queueing Networks'), ('INLS 690', ' Intermediate Selected Topics'), ('STOR 842', ' Control of Stochastic Systems in Operations Research'), ('COMP 832', ' Multimedia Networking'), ('COMP 631', ' Computer Networks')]

Recommended courses on Dataset 2. The first course is R1 and the last course is R12

QUERY COURSE 1 HIST 110- Introduction to the Cultures and Histories of Native North America
[('HIST 110', ' Introduction to the Cultures and Histories of Native North America'), ('HIST 84', ' First-Year Seminar: Monsters, Murders, and Mayhem in Microhistorical Analysis: French Case Studies'), ('HIST 85', ' First-Year Seminar: What Concentration Camp Survivors Tell Us'), ('ENGL 319', ' Introduction to Medieval English Literature, excluding Chaucer'), ('HIST 377', ' History of African Americans, 1865 to Present'), ('HIST 562', ' Oral History and Performance'), ('ENGL 67', ' First-Year Seminar: Travel Literature'), ('HIST 722', ' Readings in Contemporary Global History'), ('HIST 576', ' The Ethnohistory of Native American Women'), ('ENGL 125', ' Introduction to Poetry'), ('HIST 120', ' Sport and American History'), ('HIST 571', ' Southern Music')]

QUERY COURSE 2 GEOG 265 - Eastern Asia

[('GEOG 265', ' Eastern Asia'), ('GEOG 435', ' Environmental Politics'), ('GEOG 232', ' Agriculture, Food, and Society'), ('GEOG 212', ' Environmental Conservation and Global Change'), ('GEOG 60', ' First-Year Seminar: Health Care Inequalities'), ('GEOG 370', ' Introduction to Geographic Information'), ('GEOG 110', '' The Blue Planet: An Introduction to Earth's Environmental Systems''), ('GEOG 121', ' Geographies of Globalization'), ('GEOG 55', ' First-Year Seminar: Landscape in Science and Art'), ('GEOG 141', ' Geography for Future Leaders'), ('GEOG 542', ' Neighborhoods and Health'), ('GEOG 253', ' Introduction to Atmospheric Processes')]

QUERY COURSE 3 ENGL 282 - Travel Literature

[('ENGL 282', ' Travel Literature'), ('ENGL 310', ' Fairy Tales'), ('ENGL 281', ' Literature and Media'), ('ENGL 58', ' First-Year Seminar: The Doubled Image: Photography in U'), ('ENGL 326', ' Renaissance Genres'), ('ENGL 441', ' Romantic Literature--Contemporary Issues'), ('ENGL 385', ' Literature and Law'), ('ENGL 355', ' The British Novel from 1870 to World War II'), ('ENGL 466', ' Literary Theory--Contemporary Issues'), ('ENGL 315', ' English in the U'), ('HIST 475', ' Feminist Movements in the United States since 1945'), ('ENGL 270', ' Studies in Asian American Literature')]

QUERY COURSE 4 HIST 120 Sport and American History

[('HIST 120', ' Sport and American History'), ('HIST 584', ' The Promise of Urbanization: American Cities in the 19th and 20th Centuries'), ('HIST 142', ' Latin America under Colonial Rule'), ('ENGL 863', ' Seminar in Postcolonial Literature'), ('ENGL 666', ' Queer Latina/o Photography and Literature'), ('HIST 282', ' China in the World'), ('HIST 585', ' Race, Basketball, and the American Dream'), ('HIST 268', ' War, Revolution, and Culture: Trans-Atlantic Perspectives, 1750-1850'), ('ENGL 271', ' Mixed-Race America: Race in Contemporary American Literature and Culture'), ('HIST 539', ' The Economic History of Southeast Asia'), ('HIST 571', ' Southern Music'), ('HIST 771', ' Topics in Modern European History')]

QUERY COURSE 5 GEOG 124 Gender and Place: Feminist Geographies

[('GEOG 124', ' Gender and Place: Feminist Geographies'), ('GEOG 710', ' Advanced Physical Geography - Biogeoscience'), ('GEOG 419', ' Field Methods in Physical Geography'), ('GEOG 268', ' Geography of Africa'), ('GEOG 63', ' First-Year Seminar: The Problem with Nature and Its Preservation'), ('GEOG 52', ' Political Ecology of Health and Disease'), ('HIST 537', ' Women in the

Middle East'), ('GEOG 225', ' Space, Place, and Difference'), ('HIST 582', ' American Constitutional History since 1876'), ('GEOG 228', ' Urban Geography'), ('GEOG 111', ' Weather and Climate'), ('HIST 460', ' Late Medieval and Reformation Germany')]
QUERY COURSE 6 ENGL 251 Film Performance and Stardom

[('ENGL 251', ' Film Performance and Stardom'), ('ENGL 55', " First-Year Seminar: Reading and Writing Women's Lives"), ('GEOG 692H', ' Honors'), ('ENGL 134H', " First-Year Honors: Women's Lives"), ('GEOG 89', ' First Year Seminar: Special Topics'), ('HIST 863', ' Readings in Urban History'), ('HIST 815', ' Topics in African History'), ('HIST 890', ' Topics in History for Graduates'), ('HIST 234', ' Native American Tribal Studies'), ('ENGL 121', ' British Literature, 19th and Early 20th Century'), ('ENGL 295', ' Undergraduate Research Seminar'), ('ENGL 318', ' Multimedia Composition')]

QUERY COURSE 7 HIST 329 An Introduction to the History of Medicine

[('HIST 329', 'An Introduction to the History of Medicine'), ('HIST 311', ' Ghettos and Shtetls? Urban Life in East European Jewish History'), ('HIST 357', ' The Old South'), ('HIST 510', ' Human Rights in the Modern World'), ('HIST 179H', ' Honors Seminar in American History'), ('HIST 287', " Japan's Modern Revolution"), ('HIST 340', ' Ethics and Business in Africa'), ('HIST 721', ' Readings in European Expansion and Global Interaction, 1400-1800'), ('ENGL 141', ' World Literatures in English'), ('ENGL 71', ' First-Year Seminar: Doctors and Patients'), ('HIST 476', ' Borderlands: Religion and Ethnicity in Modern East Central Europe'), ('HIST 880', ' Readings in the Global History of Capitalism')]

QUERY COURSE 8 GEOG 341 Hydrology, Ecology, and Sustainability of the Humid Tropics

[('GEOG 341', ' Hydrology, Ecology, and Sustainability of the Humid Tropics'), ('HIST 565', ' Civil War and Reconstruction, 1848-1900'), ('HIST 288', ' Japan in the 20th Century'), ('HIST 325', ' Food and History: The Local and Global, the United Kingdom and the United States'), ('HIST 509', ' The World History of Slavery'), ('HIST 460', ' Late Medieval and Reformation Germany'), ('HIST 151', ' European History to 1650'), ('HIST 131', ' Southeast Asia to the Early 19th Century'), ('HIST 347', ' Fascist Challenge in Europe, 1918-1945'), ('ENGL 748', ' Studies in American Poetry'), ('HIST 480', " Russia's 19th Century: Cultural Splendor, Imperial Decay"), ('HIST 364', ' History of American Business')]

QUERY COURSE 9 ENGL 720 Old English Poetry

[('ENGL 720', ' Old English Poetry'), ('HIST 714', ' Introductory Colloquium in the History of Latin America since 1810'), ('ENGL 100', ' Basic Writing'), ('HIST 726', ' Introductory Colloquium in United States History to 1788'), ('ENGL 129', ' Literature and Cultural Diversity'), ('ENGL 300', ' Advanced Expository Writing'), ('ENGL 304', ' Advanced Expository Writing for Business'), ('ENGL 305', ' Advanced Expository Writing for Law'), ('ENGL 301', ' Advanced Expository Writing for the Humanities'), ('HIST 727', ' Introductory Colloquium in United States History, 1788 to 1900'), ('HIST 775', ' Studies in Modern English History'), ('HIST 534', ' The African Diaspora')]

QUERY COURSE 10 HIST 880 Readings in the Global History of Capitalism

[('HIST 880', ' Readings in the Global History of Capitalism'), ('HIST 174H', ' Honors Seminar in African, Asian, and Middle Eastern History'), ('HIST 454', ' The Reformation'), ('ENGL 857', ' Studies in 20th-Century English and American Literature'), ('HIST 312', ' History of France and Algeria'), ('HIST 274', ' History of the Ottoman Empire, 1300-1923'), ('HIST 179H', ' Honors Seminar in

American History'), ('ENGL 141', ' World Literatures in English'), ('HIST 476', ' Borderlands: Religion and Ethnicity in Modern East Central Europe'), ('HIST 64', ' First-Year Seminar: Gorbachev: The Collapse of the Soviet Empire and the Rise of the New Russia'), ('HIST 287', " Japan's Modern Revolution"), ('ENGL 638', ' 19th-Century Women Writers')]

Recommended courses on Combined Dataset. The first course is R1 and the last course is R12

QUERY COURSE 1 INLS 509 - Information Retrieval
[('INLS 509', ' Information Retrieval'), ('COMP 487', ' Information Retrieval'), ('COMP 730', ' Operating Systems'), ('COMP 788', ' Expert Systems'), ('INLS 513', ' Resource Selection and Evaluation'), ('COMP 455', ' Models of Languages and Computation'), ('GEOG 442', ' River Processes'), ('STOR 642', ' Stochastic Models in Operations Research II'), ('GEOG 441', ' Introduction to Watershed Systems'), ('GEOG 440', ' Earth Surface Processes'), ('GEOG 123', ' Cultural Geography'), ('INLS 151', ' Retrieving and Analyzing Information')]

QUERY COURSE 2 'STOR 455'- 'Methods of Data Analysis'
[('STOR 455', ' Methods of Data Analysis'), ('STOR 112', ' Decision Models for Business'), ('GEOG 591', ' Applied Issues in Geographic Information Systems'), ('STOR 654', ' Statistical Theory I'), ('STOR 471', ' Long-Term Actuarial Models'), ('STOR 155', ' Introduction to Data Models and Inference'), ('STOR 641', ' Stochastic Models in Operations Research I'), ('STOR 614', ' Linear Programming'), ('STOR 854', ' Statistical Large Sample Theory'), ('INLS 781', ' Proposal Development'), ('HIST 834', ' The United States in the Middle Period, 1815-1860'), ('STOR 435', ' Introduction to Probability')]

QUERY COURSE 3 'COMP 562'- ' Introduction to Machine Learning'
[('COMP 562', ' Introduction to Machine Learning'), ('GEOG 406', ' Atmospheric Processes II'), ('STOR 836', ' Stochastic Analysis'), ('STOR 734', ' Stochastic Processes'), ('COMP 767', ' Geometric and Solid Modeling'), ('INLS 530', ' Young Adult Literature and Related Materials'), ('COMP 734', ' Distributed Systems'), ('INLS 718', ' User Interface Design'), ('COMP 535', ' Introduction to Computer Security'), ('INLS 161', ' Tools for Information Literacy'), ('INLS 725', ' Electronic Health Records'), ('COMP 824', ' Functional Programming')]

QUERY COURSE 4 'INLS 541'- Information Visualization
[('INLS 541', ' Information Visualization'), ('INLS 501', ' Information Resources and Services'), ('INLS 582', ' Systems Analysis'), ('COMP 631', ' Computer Networks'), ('COMP 410', ' Data Structures'), ('INLS 576', ' Distributed Systems and Administration'), ('STOR 472', ' Short Term Actuarial Models'), ('STOR 734', ' Stochastic Processes'), ('INLS 556', ' Introduction to Archives and Records Management'), ('GEOG 491', ' Introduction to GIS'), ('GEOG 790', ' Spatial Analysis and Computer Modeling'), ('STOR 63', ' FYS: Statistics, Biostatistics, and Bioinformatics: An Introduction to the Ongoing Evolution')]

QUERY COURSE 5 COMP 475 2D Computer Graphics
[('COMP 475', ' 2D Computer Graphics'), ('STOR 852', ' Nonparametric Inference: Rank-Based Methods'), ('GEOG 391', ' Quantitative Methods in Geography'), ('COMP 116', ' Introduction to Scientific Programming'), ('COMP 844', ' Advanced Design of VLSI Systems'), ('STOR 654', ' Statistical Theory I'), ('STOR 854', ' Statistical Large Sample Theory'), ('STOR 112', ' Decision Models for Business'), ('HIST 834', ' The United States in the Middle Period, 1815-1860'), ('INLS 781', '

Proposal Development'), ('COMP 775', ' Image Processing and Analysis'), ('STOR 824', ' Computational Methods in Mathematical Programming')]

QUERY COURSE 6 STOR 445 Stochastic Modeling
[('STOR 445', ' Stochastic Modeling'), ('COMP 740', ' Computer Architecture and Implementation'), ('INLS 740', ' Digital Libraries: Principles and Applications'), ('STOR 835', ' Point Processes'), ('INLS 672', ' Web Development II'), ('COMP 841', ' Advanced Computer Architecture'), ('STOR 842', ' Control of Stochastic Systems in Operations Research'), ('INLS 520', ' Organization of Information'), ('COMP 730', ' Operating Systems'), ('COMP 788', ' Expert Systems'), ('INLS 513', ' Resource Selection and Evaluation'), ('COMP 487', ' Information Retrieval')]

QUERY COURSE 7 INLS 500 Human Information Interactions
[('INLS 500', ' Human Information Interactions'), ('INLS 202', ' Retrieval and Organizing Systems'), ('INLS 151', ' Retrieving and Analyzing Information'), ('GEOG 123', ' Cultural Geography'), ('GEOG 440', ' Earth Surface Processes'), ('STOR 642', ' Stochastic Models in Operations Research II'), ('INLS 754', ' Access'), ('INLS 704', ' Humanities Information'), ('INLS 748', ' Health Sciences Environment'), ('INLS 702', ' Social Science Information'), ('GEOG 442', ' River Processes'), ('COMP 455', ' Models of Languages and Computation')]

QUERY COURSE 8 COMP 844 Advanced Design of VLSI Systems
[('COMP 844', ' Advanced Design of VLSI Systems'), ('COMP 116', ' Introduction to Scientific Programming'), ('STOR 654', ' Statistical Theory I'), ('INLS 781', ' Proposal Development'), ('HIST 834', ' The United States in the Middle Period, 1815-1860'), ('COMP 475', ' 2D Computer Graphics'), ('GEOG 577', ' Advanced Remote Sensing'), ('STOR 851', ' Sequential Analysis'), ('STOR 855', ' Subsampling Techniques'), ('GEOG 391', ' Quantitative Methods in Geography'), ('COMP 580', ' Enabling Technologies'), ('STOR 635', ' Probability')]

QUERY COURSE 9 STOR 614 Linear Programming
[('STOR 614', ' Linear Programming'), ('INLS 747', ' Special Libraries and Knowledge Management'), ('COMP 651', ' Computational Geometry'), ('INLS 719', ' Usability Testing and Evaluation'), ('STOR 435', ' Introduction to Probability'), ('STOR 854', ' Statistical Large Sample Theory'), ('STOR 60', ' First-Year Seminar: Statistical Decision-Making Concepts'), ('HIST 834', ' The United States in the Middle Period, 1815-1860'), ('INLS 781', ' Proposal Development'), ('STOR 112', ' Decision Models for Business'), ('GEOG 591', ' Applied Issues in Geographic Information Systems'), ('STOR 455', ' Methods of Data Analysis')]

QUERY COURSE 10 INLS 578 Protocols and Network Management
[('INLS 578', ' Protocols and Network Management'), ('INLS 624', ' Policy-Based Data Management'), ('INLS 586', ' Project Management'), ('INLS 752', ' Digital Preservation and Access'), ('ENGL 307', ' Studies in Fiction and Poetry: Stylistics'), ('INLS 765', ' Information Technology Foundations for Managing Digital Collections'), ('HIST 493', ' Internship in History'), ('INLS 465', ' Understanding Information Technology for Managing Digital Collections'), ('GEOG 704', ' Communicating Geography'), ('ENGL 317', ' Networked Composition'), ('ENGL 706', ' Rhetorical Theory and Practice'), ('INLS 750', ' Introduction to Digital Curation')]

QUERY COURSE 11 HIST 110- Introduction to the Cultures and Histories of Native North America
[('HIST 110', ' Introduction to the Cultures and Histories of Native North America'), ('HIST 81', ' First-Year Seminar: Diaries, Memoirs, and Testimonies of the Holocaust'), ('ENGL 359', ' Latina Feminisms'), ('HIST 268', ' War, Revolution, and Culture: Trans-Atlantic Perspectives, 1750-1850'), ('ENGL 860', ' Seminar in 20th-Century Literature, English and American'), ('ENGL 387', ' Canadian

Literature'), ('ENGL 723', ' Later Middle English Literature'), ('ENGL 662', ' History of Literary Criticism'), ('ENGL 835', ' 18th-Century Fiction'), ('HIST 568', ' Women in the South'), ('ENGL 360', ' Contemporary Asian American Literature and Theory'), ('HIST 346', ' Dictators in the 20th Century')]

QUERY COURSE 12 GEOG 265 - Eastern Asia
[('GEOG 265', ' Eastern Asia'), ('GEOG 53', " First-Year Seminar: Battle Park: Carolina's Urban Forest"), ('ENGL 381', ' Literature and Cinema'), ('ENGL 281', ' Literature and Media'), ('GEOG 50', ' First-Year Seminar: Mountain Environments'), ('ENGL 830', ' Studies in Renaissance Literature: Primarily Nondramatic'), ('ENGL 70', ' First-Year Seminar: Courtly Love, Then and Now'), ('GEOG 125', ' Cultural Landscapes'), ('ENGL 685', ' Literature of the Americas'), ('GEOG 263', ' Environmental Field Studies in Siberia')]

QUERY COURSE 13 ENGL 282 - Travel Literature
[('ENGL 282', ' Travel Literature'), ('ENGL 472', ' African American Literature--Contemporary Issues'), ('ENGL 373', ' Southern American Literature'), ('HIST 257', ' Politics, Society, and Culture in Postwar Germany'), ('ENGL 390', ' Studies in Literary Topics'), ('HIST 345', ' Comparative Strategies of Empire'), ('ENGL 868', ' African American and African Diasporan Literature, 1930-1970'), ('ENGL 410', ' Documentary Film'), ('HIST 64', ' First-Year Seminar: Gorbachev: The Collapse of the Soviet Empire and the Rise of the New Russia'), ('HIST 782', ' Readings in Soviet History')]

QUERY COURSE 14 HIST 120 Sport and American History
[('HIST 120', ' Sport and American History'), ('HIST 463', ' Germany since 1918: Politics, Society, and Culture'), ('HIST 203', ' Empires and Cultures in the Modern World'), ('ENGL 448', ' Philosophies of Life from Classical Antiquity to 1800'), ('HIST 757', ' Late Medieval England'), ('HIST 152', ' European History since 1650'), ('HIST 564', ' The American Revolution, 1763-1815'), ('HIST 462', ' Germany, 1806-1918: Politics, Society, and Culture'), ('HIST 243', ' The United States and Africa'), ('ENGL 659', ' War in 20th-Century Literature'), ('GEOG 464', ' Europe Today: Transnationalism, Globalisms, and the Geographies of Pan-Europe'), ('HIST 288', ' Japan in the 20th Century')]

QUERY COURSE 15 GEOG 124 Gender and Place: Feminist Geographies
[('GEOG 124', ' Gender and Place: Feminist Geographies'), ('GEOG 452', ' Mobile Geographies: The Political Economy of Migration'), ('HIST 452', ' The Renaissance: Italy, Birthplace of the Renaissance, 1300-1550'), ('ENGL 140', ' Introduction to Gay and Lesbian Culture and Literature'), ('ENGL 80', " First-Year Seminar: The Politics of Persuasion: Southern Women's Rhetoric"), ('ENGL 124', ' Contemporary Literature'), ('HIST 272', ' Contemporary India, Pakistan, and Bangladesh'), ('HIST 276', ' The Modern Middle East'), ('ENGL 875', ' Critical Race Theory-Graduate Seminar'), ('HIST 501', ' The Gender of Welfare: Comparative Perspectives, 19th and 20th Century'), ('ENGL 255', ' Introduction to Media Studies'), ('HIST 581', ' American Constitutional History to 1876')]

QUERY COURSE 16 ENGL 251 Film Performance and Stardom
[('ENGL 251', ' Film Performance and Stardom'), ('GEOG 435', ' Environmental Politics'), ('ENGL 377', ' Introduction to the Celtic Cultures'), ('HIST 722', ' Readings in Contemporary Global History'), ('HIST 924', ' Seminar in Modern European History'), ('ENGL 340', ' Studies in Jane Austen'), ('GEOG 444', ' Landscape Biogeography'), ('ENGL 313', ' Grammar of Current English'), ('INLS 728', ' Seminar in Knowledge Organization'), ('GEOG 212', ' Environmental Conservation and Global Change'), ('INLS 203', ' Human Information Behavior'), ('GEOG 720', ' Cultural and Political Ecology')]

QUERY COURSE 17 HIST 329 An Introduction to the History of Medicine

[('HIST 329', ' An Introduction to the History of Medicine'), ('HIST 469', ' European Social History, 1815-1970'), ('ENGL 724', ' Chaucer'), ('ENGL 850', ' Studies in English and American Poetry of the 20th Century'), ('HIST 106', ' Ancient History'), ('HIST 107', ' Medieval History'), ('ENGL 126', ' Introduction to Drama'), ('GEOG 448', ' Transnational Geographies of Muslim Societies'), ('ENGL 437', ' Chief British Romantic Writers'), ('ENGL 225', ' Shakespeare'), ('ENGL 858', ' Studies in English and American Fiction of the 20th Century'), ('ENGL 333', ' 18th-Century Fiction')]

QUERY COURSE 18 GEOG 341 Hydrology, Ecology, and Sustainability of the Humid Tropics
[('GEOG 341', ' Hydrology, Ecology, and Sustainability of the Humid Tropics'), ('INLS 710', ' Evidence-Based Medicine'), ('INLS 525', ' Electronic Records Management'), ('GEOG 454', ' Historical Geography of the United States'), ('GEOG 230', ' The World at Eight Billion'), ('GEOG 423', ' Social Geography'), ('INLS 758', ' International and Cross-Cultural Perspectives for Information Management'), ('GEOG 720', ' Cultural and Political Ecology'), ('GEOG 111', ' Weather and Climate'), ('GEOG 222', ' Health and Medical Geography'), ('GEOG 715', ' 715 Land Use/Land Cover Dynamics and Human-Environment Interaction'), ('ENGL 79', ' First-Year Seminar: Globalization/Global Asians')]

QUERY COURSE 19 ENGL 720 Old English Poetry
[('ENGL 720', ' Old English Poetry'), ('GEOG 429', ' Urban Political Geography: Durham, NC'), ('ENGL 155', ' The Visual and Graphic Narrative'), ('INLS 613', ' Text Mining'), ('INLS 789', ' Big Data, Algorithms and Society'), ('ENGL 564', ' Interdisciplinary Approaches to Literature'), ('GEOG 419', ' Field Methods in Physical Geography'), ('ENGL 661', ' Introduction to Literary Theory'), ('HIST 50', ' First-Year Seminar: Time and the Medieval Cosmos'), ('STOR 66', ' First-Year Seminar: Visualizing Data'), ('ENGL 147', ' Mystery Fiction'), ('ENGL 852', ' Seminar in Modern Drama')]

QUERY COURSE 20 HIST 880 Readings in the Global History of Capitalism
[('HIST 880', ' Readings in the Global History of Capitalism'), ('HIST 285', ' 20th-Century China'), ('HIST 240', ' Introduction to Mexico: A Nation in Four Revolutions'), ('HIST 261', ' France, 1870-1940'), ('ENGL 877', ' Introduction to Modern Irish II'), ('HIST 151', ' European History to 1650'), ('ENGL 250', ' Faulkner'), ('ENGL 863', ' Seminar in Postcolonial Literature'), ('HIST 565', ' Civil War and Reconstruction, 1848-1900'), ('HIST 335', ' Cracking India: Partition and Its Legacy in South Asia'), ('HIST 770', '' Readings in Modern European Women's and Gender History''), ('HIST 784', ' Readings in East European History')]