LDA Model for Analyzing Course Descriptions

Master's Project Paper

Saipavan Tadikonda

Department of Computer Science
University of North Carolina at Greensboro
Greensboro, USA
email : s_tadikonda@uncg.edu

Dr.Lixin Fu

Department of Computer Science
University of North Carolina at Greensboro
Greensboro, USA
email: l_fu@uncg.edu

***Abstract*—The purpose of this paper is to demonstrate a way to generate a topic model using the LDA topic modeling on the course descriptions of the selected universities in the United States of America. Considered universities for this paper are University of North Carolina at Chapel Hill (UNC), North Carolina State University (NCSU), University of North Carolina at Charlotte (UNCC) and Georgia State University (GSU). This paper starts by showing how to web scrap the required course details from each considered university followed by building the LDA Topic model on the scrapped data which generates the topics where each topic is associated with topic keywords. The generated topic model on the above considered four universities will help to the find how each course is related i.e., (with what weight) to the obtained topics of the model. Further this paper shows how we can group the obtained LDA Topics into different communities with help of the modularity class concept in the visualization tool Gephi. The obtained communities helps to find different courses from all the considered universities that fall into those communities. It also helps to find the departments which are related to each other in those communities.**

I. Introduction

In the recent years with the increasing amount of unstructured data it is difficult to obtain the required information out of unstructured data but it is not impossible to obtain. With the help of technology, we can use some powerful methods to mine the data and fetch the required information. One such method in this aspect is Topic Modeling. Topic Modeling is a process to identify the topics present in the text objects and to derive the hidden patterns exhibited by a text corpus. It is an unsupervised approach which is used for finding and observing the bunch of words called topics in large clusters of texts. This approach is very useful for the purpose for document clustering, organizing large blocks of textual data, information retrieval from the unstructured text and feature selection.

Each university website contains the details of their courses offerings. Topic Modeling on these course offerings helps to analyze the course offerings and gives the topics from those offerings. The obtained topics will help to compare the courses across the universities. So this would act like a beneficial tool for students who would like to compare the courses which they are gonna take with other courses in other considered universities. One of the popular unsupervised algorithm used for topic modeling is Latent Dirichlet Allocation (LDA). This paper shows the implementation of this LDA algorithm on the considered universities. The best way of implementing this LDA model is by using the python's Gensim library. Also the other python libraries which are used in this process are Beautiful Soup, Natural Language Processing Tool Kit (NLTK) and Networkx. For visualizing the courses and their relationship with the topics obtained the visualization tool Gephi is used.

The paper is organized as follows: Section 2 contains the methodology which was followed. Section 3 contains the Implementation details. Section 4 contains the results which are obtained and Section 5 contains the conclusion of the paper.

II. Methodology

Below is the figure which contains the workflow followed to generate the LDA Model on the considered universities:

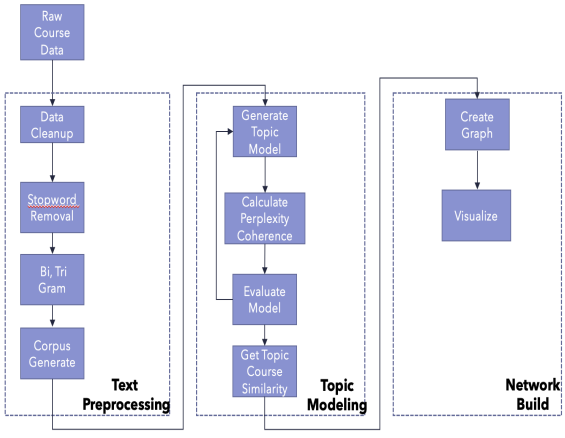

Fig. 1. Workflow

A. Raw Courses Data

In order to generate the LDA model first we need to obtain all the course details from the considered universities. Each university has its own website, and, in those websites, we can observe what all courses that university is offering for both

undergraduates and graduate students in course description pages. Our aim is to scrap the required details i.e., Course Number, Course Title, Course Description and Course Department from each course description page. We can achieve this by using the python library Beautiful Soup. First, we need to fetch the web page of course description using the requests library and we need to apply the html parser of the beautiful soup library on the fetched web page as it parses the html code from the text. From the obtained html code, we need to separate and store the required details which helps in navigating through all pages of the course descriptions. Then by inspecting each page of the course description we need to identify the html tags under which our course number, course title, course description and course department information is placed and by using those html tags we need to scrap our required details. After obtaining those details add the respective University Name column and store them into a csv file.

B. Text Preprocessing

1) Data Cleanup: Read the stored csv file using the python pandas library and remove the unwanted spaces in the beginning and ending of each data in each column. Check if there are any null values present in the obtained data. Each university have some common courses like Directed Study, Independent Study, Thesis, Project, etc. which doesn't contribute much to build the LDA model. So, remove the repeated courses from the data frame so that the final LDA model will be efficient. The data under course description column will be in sentence format. To build an LDA model on the course description column we need to convert each course description into words as the model can only understand and get useful information from the words. Use the simple_preprocess () function which is provided by the gensim library and convert each course description sentence into words.

2) Stopwords Removal: From the obtained words of each course description, remove the words which don't add much semantic value. So, by using the stopwords from the NLTK library remove all the words which don't add much meaning.

3) Creating Bi and Trigrams: After removing the stopwords from the words of course description find if there are any bigrams i.e., any two words occurring together frequently and trigrams i.e., any three words occurring together frequently. This can be done by using the Phrases () function provided by the gensim library. So here if there are any bigrams then those words will be joined by '_' and if there are any trigrams even those three words are joined by '_'. After finding bi and trigrams, lemmatize the words of each course description as it would give words for each course description which are nouns, adjectives, verbs, and adverbs. It helps the topic model to categorize things more easily. Lemmatization is nothing but converting the word into its root word.

4) Generating Corpus: Dictionary (Id2word) and corpus are the two essential inputs for the LDA model. Id2word identifies the keywords of the words that we obtained after lemmatization and gives ids to each word. This can be done by using the Dictionary () function from the gensim.corpora library. Coming to the corpus it contains the id and frequency of that word in the form of tuple for each course description. It can be achieved by using the doc2bow () function which is used to convert the collection of words into a bag of word representation.

C. Topic Modeling

1) Generate LDA Model: Along with Dictionary and corpus some other inputs such as number of topics, Chunksize, Update_every and passes are required. Chunksize is nothing but the number of documents to be used in each training chunk. Update_every is nothing but how often the model parameters should be updated. Passes are nothing but the total number of the training passes. Basically, the LDA model considers each document as a collection of topics in a certain proportion and each topic as a collection of keywords in a certain proportion. If provided the LDA model with a number of topics, it rearranges the topics distribution within the documents. But the efficiency of the LDA model depends on the number of topics given. Optimal number of topics for the LDA model can be obtained by plotting perplexity vs coherence plot.

2) Calculate Perplexity and Coherence: Perplexity and Coherence score is used as a convenient measure to judge how good the trained model is performing. Perplexity tries to measure how the trained model is surprised when it is given with the new data. Generally, the lower the perplexity the better is the model. Coherence score is used to distinguish between the good and the bad topics. Generally, a score more than 0.4 is considered good.

3) Optimal Topics of LDA Model: To identify the optimal number of topics for the final LDA model plot the scatter plot for both perplexity and coherence values for different values of number of topics. From the scatter plot obtained see at what point the both perplexity and coherence intersect, that number will be the input for the optimal number of topics in the final LDA Model. The generated LDA model will be effective. With the help of the Gensim Library, infer the topics and their keywords of the generated LDA model.

4) Topic to Courses relationship: After inferring the topics from the LDA model created, see how each course is related to a topic and its contribution in a significant way. Each course can have a relation with sort of multiple topics with a contribution.

D. Network Build

1) Create Network Graph: Network graph can be created with the help of the python's networkx library. The graph created using networkx library contains nodes and edges. In this case nodes are nothing but course numbers in each university and the LDA topics which are obtained. For each node set the node attributes as document text, course title, course description, course department and university. Edges of the created network graph are nothing but the course number, and the topic to which it is related and with what percentage

contribution i.e., the weight between them. Finally store the created network graph in graphml format.

2) Visualize the network graph using Gephi: Gephi is an open-source software for visualizing and analyzing large network graphs. It is used to explore, analyze, spatialize, filter, manipulate and export all types of the graphs. Open the stored graphml file in this application. While opening the graphml file it shows the import report that contains the information regarding the number of nodes, edges in that graph. By default, the nodes in the graphs will be very close. Some operations should be performed to see the understandable graph. One such operation is adjusting the layout. Layout algorithms set the graph shape. It is the most essential step. In the layout module choose Force Atlas in the dropdown and in the layout, properties set the repulsion strength at 10000 to expand the graph. Run this layout algorithm until the graph expands after the expansion stop the layout algorithm. Here all the nodes move far away from other nodes. Next locate the ranking module which will be in the top left. This ranking module will help in configuring the node's color and size. For that to happen choose the rank parameter as degree and click on apply. The graph color will be changed based on the degree.

Next locate the Statistics module on the right panel. Under this module click the run button beside average path length. It computes the path length for all possible pairs of nodes and gives information about how nodes are close to each other. On running the average path length three new values will be obtained. Those are betweenness centrality, closeness centrality and eccentricity. After obtaining these three values go back to the ranking module and select betweenness centrality in the list as this metric indicates the influential nodes for the highest value. Click on the diamond icon in the toolbar for size and set the minimum size and maximum size of the nodes. On clicking on apply will result in network graph with colored nodes by the degree and size of nodes by betweenness centrality metric. Now go back to the layout module and check the adjust by sizes box in the layout properties and run the layout algorithm for a bit. It results in network graph without any overlapping nodes.

Now find the 'T' button in the bottom of the Gephi interface and click on it as it displays the node labels and set the label size proportional to the node size so that the label of the smaller node appears small, and the label of the larger node appears big. Also set the label size with the scale slider which is available next to the label size. Now the graph obtained contains all the nodes with their respective sizes and edges between the nodes.

3) Community Detection: One of the goal of this paper is to show how to group the topics into communities. Find the modularity in the statistics panel of the Gephi interface and click the run button next to it. In the popup check on the randomize and click ok to launch the detection. In the background Gephi will implement the Louvain method to detect the communities in the graph. Now locate the Partition module which will be left to the Ranking module. In the dropdown select the Modularity class which is obtained by running the Modularity from the statistics panel before. A random color would be set for each community identifier and on clicking on apply the graph will be updated with the respective colors based on the community they belong to.

III. Implementation

A. Courses Data

In order to generate a LDA Model we need to first have all the course details from the each consisdered university i.e., University of North Carolina at Chapel Hill (UNC), North Carolina State University (NCSU), University of North Carolina at Charlotte (UNCC) and Georgia State University (GSU). In order to obtain the course details, web scrapping the required course number, course title, course description, course department and university name from the respective html tags of each university website using the python Beautiful Soup library. After obtaining for each university combining all the course details of each university into one dataframe. The final dataframe contains all the course details from the considered universities and it looks as follows:

	CourseNumber	CourseTitle	CourseDescription	CourseDepartment	University
0	ACCT 5220	Income Tax	An introduction to the Federal income tax syst...	Accounting	UNCC
1	ACCT 5311	Intermediate Financial Reporting I	Analysis of the financial reporting requiremen...	Accounting	UNCC
2	ACCT 5312	Intermediate Financial Reporting II	A continuation of ACCT 5311 with emphasis on ...	Accounting	UNCC
3	ACCT 6110	Tax Research and Planning	Tax research techniques applicable to federal ...	Accounting	UNCC
4	ACCT 6120	Taxation of Corporations and Shareholders	Examines the federal and state tax law applica...	Accounting	UNCC
...	...	...	...	...	...
27149	WLC 8320	Cultural and Literary Analysis and Production	Cultural and Literary Analysis and Production3...	World Languages and Cultures	GSU
27150	WLC 8330	The Media, the Arts, and Popular Culture	The Media, the Arts, and Popular Culture3 Cred...	World Languages and Cultures	GSU
27151	WLC 8340	Learning a World Language	Learning a World Language3 Credit Hours Descr...	World Languages and Cultures	GSU
27152	WLC 8350	Identities, Borders, and Social Justice in Lan...	Identities, Borders, and Social Justice in Lan...	World Languages and Cultures	GSU
27153	WLC 8360	The Profession, Career and Intercultural Exper...	The Profession, Career and Intercultural Exper...	World Languages and Cultures	GSU

27154 rows × 5 columns

Fig. 2. Raw Courses Data of considered universities

Obtained total of 27154 courses data from the considered universities.

B. Data cleanup

From the obtained raw courses data removing the extra spaces if there exists any by applying the strip() function to each column in the data frame. Also removing if there any nulls present in the obtained raw data. Removing the repeated courses from the original raw courses data. The final courses data frame obtained after cleaning the raw courses data is as follows:

	CourseNumber	CourseTitle	CourseDescription	CourseDepartment	University	collated_cno	deptcode	cno	Course Level
0	ACCT 5220	Income Tax	An introduction to the Federal income tax syst...	Accounting	UNCC	0	ACCT	5220	Graduate
1	ACCT 5311	Intermediate Financial Reporting I	Analysis of the financial reporting requiremen...	Accounting	UNCC	0	ACCT	5311	Graduate
2	ACCT 5312	Intermediate Financial Reporting II	A continuation of ACCT 5311 with emphasis on ...	Accounting	UNCC	0	ACCT	5312	Graduate
3	ACCT 6110	Tax Research and Planning	Tax research techniques applicable to federal ...	Accounting	UNCC	0	ACCT	6110	Graduate
4	ACCT 6120	Taxation of Corporations and Shareholders	Examines the federal and state tax law applica...	Accounting	UNCC	0	ACCT	6120	Graduate
...	...	...	...	...	...	...	...	...	...
27149	WLC 8320	Cultural and Literary Analysis and Production	Cultural and Literary Analysis and Production3...	World Languages and Cultures	GSU	0	WLC	8320	Graduate
27150	WLC 8330	The Media, the Arts, and Popular Culture	The Media, the Arts, and Popular Culture3 Cred...	World Languages and Cultures	GSU	0	WLC	8330	Graduate
27151	WLC 8340	Learning a World Language	Learning a World Language3 Credit Hours Descr...	World Languages and Cultures	GSU	0	WLC	8340	Graduate
27152	WLC 8350	Identities, Borders, and Social Justice in Lan...	Identities, Borders, and Social Justice in Lan...	World Languages and Cultures	GSU	0	WLC	8350	Graduate
27153	WLC 8360	The Profession, Career and Intercultural Exper...	The Profession, Career and Intercultural Exper...	World Languages and Cultures	GSU	0	WLC	8360	Graduate

26857 rows × 9 columns

Fig. 3. Courses Data after data cleaning

After data cleanup and removing the repeated courses the total courses from all the considered universities are 26857.

C. Tokenization using Gensim

Now converting the course description of each course into words i.e., nothing but the tokenization. This is obtained by using the simple_preprocess() function of the gensim package. The next step is forming the bigrams and trigrams using the Phrases() function of the gensim package. Now removing the stopwords from the tokens of the each description using the NLTK library. After removing the stop words finding the lemmatized words of the original words in the course description followed by generating dictionary and corpus which are the two essential inputs to generate the LDA Model.

D. LDA Model

Choosing the number of topics to be inferred is the first step before building an LDA model. One cannot randomly choose the number of topics and create a model, resulting in repeated topics or very few topics than needed. The solution to this was to calculate the Topic coherence and Perplexity of a test model with a random number of topics and see where these two values intersect. So the following process was to calculate a test model's perplexity and topic coherence by varying the number of topics from the range of 5 to 50 in the step of 5. Now see at what number of topics the coherence and perplexity intersect. The plot is as follows:

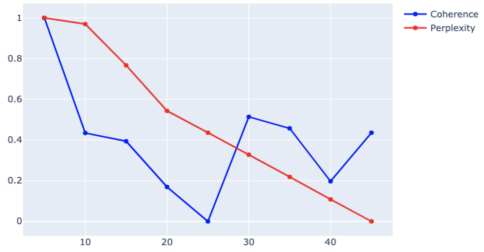

Fig. 4. Optimal number of topics

The coherence and perplexity plot intersects once at 5 and again at 29. Now the LDA model is generated by giving number of topics as 29 so the topics that are inferred from the LDA model will not be repetitive. The LDA topics and the topic keywords obtained are as follows:

	Topic	Topic_Keywords
0	Topic 0	health, course, care, provide, various, address, issue, focus, specific, overview
1	Topic 1	management, system, organization, process, information, plan, decision, planning, operation, control
2	Topic 10	analysis, datum, science, method, technique, approach, use, include, question, basic
3	Topic 11	relate, gender, historical, conduct, environmental, thesis, race, goal, core, comprehensive
4	Topic 12	problem, communication, analytic, value, basis, implementation, hand, programming, visual, image
5	Topic 13	major, theoretical, contemporary, perspective, write, present, discussion, case, framework, influence
6	Topic 14	community, also, different, way, component, think, complex, multiple, software, course
7	Topic 15	description, instructor, design, project, skill, develop, student, variety, take, appropriate
8	Topic 16	research, graduate, student, opportunity, activity, year, pursue, final, prerequisite, public
9	Topic 17	emphasis, review, examine, service, impact, ethic, place, state, integrate, issue
10	Topic 18	policy, social, language, economic, cultural, political, history, change, issue, foundation
11	Topic 19	hour, credit, course, student, include, education, clinical, require, school, psychology
12	Topic 2	introduction, role, procedure, form, emerge, alternative, objective, potential, pertain, prerequisite
13	Topic 20	social, global, relationship, medium, emphasize, survey, resource, identity, advance, network
14	Topic 21	student, learn, program, seminar, knowledge, critical, read, effective, scientific, presentation
15	Topic 22	course, behavior, apply, reading, develop, skill, explore, make, text, utilize
16	Topic 23	physical, structure, evaluate, cover, include, age, time, system, computer, digital
17	Topic 24	development, field, business, technology, cover, course, create, manage, product, examine
18	Topic 25	study, requirement, special, area, individual, direct, depth, interest, concentration, enable
19	Topic 26	strategy, identify, limited, creation, drive, rule, phase, limitation, compliance, handle
20	Topic 27	topic, current, include, class, issue, investigation, subject, representation, view, elective
21	Topic 28	assessment, performance, financial, market, interpretation, assess, offer, security, quality, capstone
22	Topic 3	advanced, high, grade, level, literature, philosophy, art, discipline, course, consequence
23	Topic 4	human, select, woman, modern, behavioral, number, evolution, genetic, significant, theorie
24	Topic 5	student, course, experience, work, provide, practice, understand, teach, professional, culture
25	Topic 6	international, examination, standard, reporting, prerequisite, politic, region, entry, patient, moral
26	Topic 7	model, base, week, tool, modeling, outcome, integration, use, expect, point
27	Topic 8	theory, application, introduce, practice, concept, principle, basic, method, practical, evidence
28	Topic 9	need, treatment, biological, intensive, depend, prerequisite, math, patient, dynamic, consent

Fig. 5. LDA topics and keywords

IV. Results and Discussions

A. Topic to Courses Relationship

Now for each course finding with what LDA topics out of 29 LDA topics it is related and with what probability. The final data frame will contain the document number, document text, the topic it is related to, percentage contribution and topic keywords. It is as follows:

	Document_no	Document_Text	Topic	Perc_Contribution	Topic_Keywords
0	0	[introduction, federal, income_tax, system, em...	17	0.3035	emphasis, review, examine, service, impact, et...
1	0	[introduction, federal, income_tax, system, em...	2	0.2035	introduction, role, procedure, form, emerge, a...
2	0	[introduction, federal, income_tax, system, em...	1	0.1035	management, system, organization, process, inf...
3	0	[introduction, federal, income_tax, system, em...	8	0.1034	theory, application, introduce, practice, conc...
4	0	[introduction, federal, income_tax, system, em...	11	0.1034	relate, gender, historical, conduct, environme...
...	...	...	...	...	...
239784	26856	[profession, career, intercultural, experience...	24	0.1058	development, field, business, technology, cove...
239785	26856	[profession, career, intercultural, experience...	18	0.0493	policy, social, language, economic, cultural, ...
239786	26856	[profession, career, intercultural, experience...	12	0.0493	problem, communication, analytic, value, basis...
239787	26856	[profession, career, intercultural, experience...	11	0.0493	relate, gender, historical, conduct, environme...
239788	26856	[profession, career, intercultural, experience...	23	0.0492	physical, structure, evaluate, cover, include,...

Fig. 6. Topic to Courses Relationship

B. Network graph

As the number of courses across the considered universities are huge, it is difficult to analyze how the courses are related to the LDA topics. It would be efficient and easy if we create a network graph. Creating a network graph where the nodes of the graph are nothing but the course numbers and the LDA topics and edges of the graph are nothing but the degree range with which each course is connected to the topic. Below is the network graph obtained after many manipulations using the Gephi visualization tool.

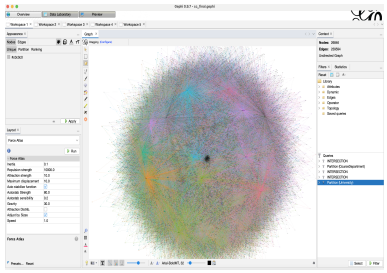

Fig. 7. Network graph after manipulations

C. Finding Data Science Cluster

By running the Modularity class in the statistics panel of the Gephi interface in the background the gephi implements the Louvian method to detect the communities in the graph. It sets each unique color to each community. Each community consists of group of topics. From the obtained communities the Modularity class 10 grouped the topic 10, topic 12, topic 23 and topic 7. On observing the topic keywords of above mentioned topics theoretically one can come to a conclusion that these topics are related to the data science field. So this modularity class 10 cluster can be called as data science cluster. Separting the modularity class 10 cluster from the network graph. It is as follows.

	Topic	Topic_Keywords
0	Topic 0	health, course, care, provide, various, address, issue, focus, specific, overview
1	Topic 1	management, system, organization, process, information, plan, decision, planning, operation, control
2	Topic 10	analysis, datum, science, method, technique, approach, use, include, question, basic
3	Topic 11	relate, gender, historical, conduct, environmental, thesis, race, goal, core, comprehensive
4	Topic 12	problem, communication, analytic, value, basis, implementation, hand, programming, visual, image
5	Topic 13	major, theoretical, contemporary, perspective, write, present, discussion, case, framework, influence
6	Topic 14	community, also, different, way, component, think, complex, multiple, software, course
7	Topic 15	description, instructor, design, project, skill, develop, student, variety, take, appropriate
8	Topic 16	research, graduate, student, opportunity, activity, year, pursue, final, prerequisite, public
9	Topic 17	emphasis, review, examine, service, impact, ethic, place, state, integrate, issue
10	Topic 18	policy, social, language, economic, cultural, political, history, change, issue, foundation
11	Topic 19	hour, credit, course, student, include, education, clinical, require, school, psychology
12	Topic 2	introduction, role, procedure, form, emerge, alternative, objective, potential, pertain, prerequisite
13	Topic 20	social, global, relationship, medium, emphasize, survey, resource, identity, advance, network
14	Topic 21	student, learn, program, seminar, knowledge, critical, read, effective, scientific, presentation
15	Topic 22	course, behavior, apply, reading, develop, skill, explore, make, text, utilize
16	Topic 23	physical, structure, evaluate, cover, include, age, time, system, computer, digital
17	Topic 24	development, field, business, technology, cover, course, create, manage, product, examine
18	Topic 25	study, requirement, special, area, individual, direct, depth, interest, concentration, enable
19	Topic 26	strategy, identify, limited, creation, drive, rule, phase, limitation, compliance, handle
20	Topic 27	topic, current, include, class, issue, investigation, subject, representation, view, elective
21	Topic 28	assessment, performance, financial, market, interpretation, assess, offer, security, quality, capstone
22	Topic 3	advanced, high, grade, level, literature, philosophy, art, discipline, course, consequence
23	Topic 4	human, select, woman, modern, behavioral, number, evolution, genetic, significant, theorie
24	Topic 5	student, course, experience, work, provide, practice, understand, teach, professional, culture
25	Topic 6	international, examination, standard, reporting, prerequisite, politic, region, entry, patient, moral
26	Topic 7	model, base, week, tool, modeling, outcome, integration, use, expect, point
27	Topic 8	theory, application, introduce, practice, concept, principle, basic, method, practical, evidence
28	Topic 9	need, treatment, biological, intensive, depend, prerequisite, math, patient, dynamic, consent

Fig. 8. Topic keywords for Data Science cluster

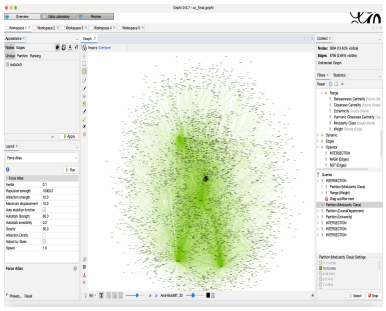

Fig. 9. Modularity class 10 of the network graph

D. Graph Representations

In the Modularity class 10 showing what courses from the University of Chapel Hill (UNC) are falling in this cluster. It is as follows:

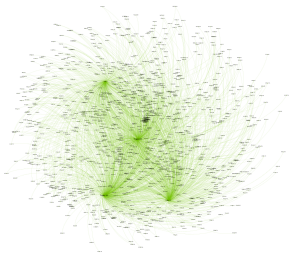

Fig. 10. UNC courses falling in Modularity class 10

In the Modularity class 10 showing what courses from the North Carolina State University (NCSU) are falling in this cluster. It is as follows:

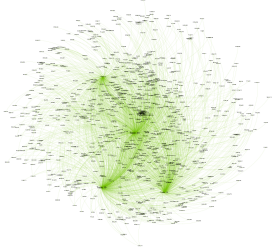

Fig. 11. NCSU courses falling in Modularity class 10

In the Modularity class 10 showing what courses from the Georgia State University (GSU) are falling in this cluster. It is as follows:

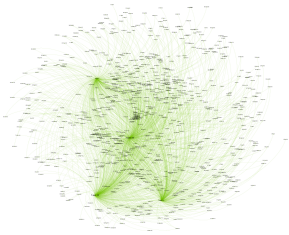

Fig. 12. GSU courses falling in Modularity class 10

In the Modularity class 10 showing what courses from the University of North Carolina at Charlotte (UNCC) are falling in this cluster. It is as follows:

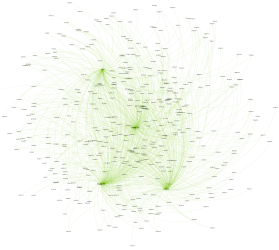

Fig. 13. UNCC courses falling in Modularity class 10

In the Modularity class 10 showing the Computer Science department courses from the University of North Carolina at Chapel Hill (UNC) that are falling in this cluster. It is as follows:

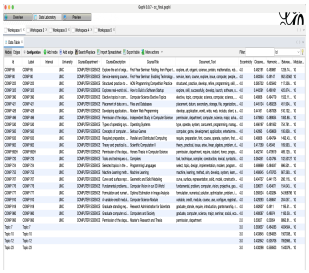

Fig. 14. UNC Computer Science courses falling in Modularity class 10

In the Modularity class 10 showing the Computer Science department courses from the North Carolina State University (NCSU) that are falling in this cluster. It is as follows:

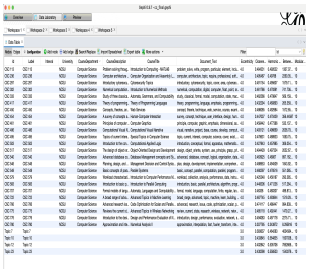

Fig. 15. NCSU Computer Science courses falling in Modularity class 10

In the Modularity class 10 showing the Computer Science department courses from the Georgia State University (GSU) that are falling in this cluster. It is as follows:

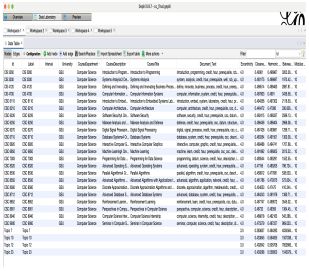

Fig. 16. GSU Computer Science courses falling in Modularity class 10

In the Modularity class 10 showing the Computer Science department courses from the University of North Carolina at Charlotte (UNCC) that are falling in this cluster. It is as follows:

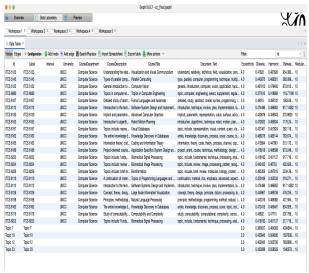

Fig. 17. UNCC Computer Science courses falling in Modularity class 10

V. Conclusion

This paper demonstrated a way how to web scrape the required details from the websites of the considered universities using the Beautiful Soup python library and how to use Topic Modeling on the scrapped course details and how courses from different universities relates to the topics that are inferred from the generated LDA model. In the later stage the paper also explained how to visualize the created network graph using the visualization tool Gephi and also how to group the obtained LDA topics into the clusters to obtain different communities. At the end it also showed the courses from each university falling under the clusters.

The future work to this paper would be implementing a website on this which helps students to relate courses across the multiple universities. Also it leads a way to explore how the obtained clusters relates to particular field more efficiently.

References

[1] https://libres.uncg.edu/ir/uncg/f/Kallem_uncg_0154M_13722.pdf
[2] https://beautiful-soup-4.readthedocs.io/en/latest/
[3] https://radimrehurek.com/gensim/intro.html
[4] https://www.nltk.org/
[5] https://networkx.org/documentation/networkx-1.10/reference/introduction.html
[6] https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/
[7] https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/9createbigramandtrigrammodels
[8] https://datascienceplus.com/evaluation-of-topic-modeling-topic-coherence/
[9] https://gephi.org/users/quick-start/
[10] https://www.crummy.com/software/BeautifulSoup/bs4/doc/quick-start
[11] https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/
[12] https://catalogs.gsu.edu/
[13] https://catalogs.gsu.edu/content.php?catoid=13navoid=1427
[14] https://catalog.unc.edu/courses/
[15] http://catalog.ncsu.edu/course-descriptions/
[16] https://catalog.uncc.edu/content.php?catoid=19navoid=1167