Project Recommendation System

Anonymous ACL submission

1 Problem Statement

Through our academics, we are oftentimes confused about what topics to select for intentions like course project, minor/major project for degree achievement, thesis or research work, etc. Very often we pick up a project from an area where we have a little knowledge and want to just explore more about that area. We generally have knowledge about different domains but cannot find an appropriate topic which makes use of all the domain that we know. In these cases, we might want to refer to previously made projects in that domain. Even if we want to know about the past advancements about that matter such recommendation systems can be quite helpful for an eager learner. If one gets a topic of his/her own interest the person works on that with more passion and determination.

2 Background

Many systems have been made for recommending research papers and research guides etc. But in our project we are trying to recommend projects based on the individuals previous domain knowledge and interests. A project report contains the details of the project all the deep insights used in the implementation as well as the results of the project with all its pros and cons. If some one tries to do a project in that area he/she can get a details of the work already done in that field.

3 Data-set Used

The data set consist of 3266 projects reports from different domains. The input to the code is a CSV file containing Document Id,Project Name, University in which the project was made, Course name for which the project was made and hyper link to the report of the project. We have collected the data-set by scrapping.

Details about data collection: Data set is collected by scrapping the web page of Stanford University.We have scraped the web pages subject wise. We followed the following steps:

a. We have searched for online course pages of different courses of different universities. The course pages contain the details of the projects done by the students during the academic session of current as well as previous years.

b. These projects have their details in the report file. we find the links of the PDF using scraping. The titles of the project and the domain to which the project belongs is scraped from website. These all information is stored in a CSV file along with a document Id. The library ”Beautiful Soup” is used for scraping.

c. The pdf file is accessed using the link in CSV file and then we convert it to .txt files so that we can get all the written text in that file.

d. Then we pre-process each text file to remove stop words, punctuation etc. Then the processed text is stored in Data Frames of Pandas.

4 Proposed Solution Sketch

Database created:.collect data-set from Stanford university.

4.1 preprocessing

1.remove punctuation
2.convert text to lowercase
3.build bigrams and trigrams from text file
4.stop-words removal
5.Lemmatize:Only noun,adjective,verb ,adverb are kept.

4.2 Model

4.2.1 LDA

LDA Topic Model technique is very useful for the purpose for document clustering, organising large blocks of textual data and retrieve information from unstructured text and feature selection. LDA assumes documents are produced from a mixture of topics. Those topics then generate words based on their probability distribution.LDA backtracks and tries to figure out what topics would create those documents in the first place. Once we get the total no of topic then we assign each topic from the top words we get from LDA.

4.2.2 LSI

It is also known as LSA which is latent Semantic Analysis.It will used the term document matrix and perform Singular valued decomposition over that and then take out all the latent topics.LSI is much faster to train then LDA but its disadvantage is having low accuracy.

4.2.3 HDP

It works with mixture of words to assign topics to the document.It is an extension of LDA.In this design the numbers of topics in document modeling is not known in prior.When their is uncertainty in topics count then to capture this we can apply Dirchlet process.In this a common distribution is consider to show the possible number of topics for the corpus,then from this distribution finite distribution of all required topics were get sampled.

Advantage of using HDP is it will learn out maximum number of topics and are not bound instead of specified number of topics in advance.It will not used in the cases where limited number of topics is required.

4.3 Clustering

In clustering there is group formation of similar kind of data in same group on the basis of common property.We use the doc2vec model to create the vector notation of the documents.Then we use the clustering method to cluster the document which are correlated to each other.and then we find the most common words which represent the topic of the clustered document,We can apply two different approach to find the most common words either choose the TF-Idf technique or apply the LDA model on the clusterto find the most important words in the document

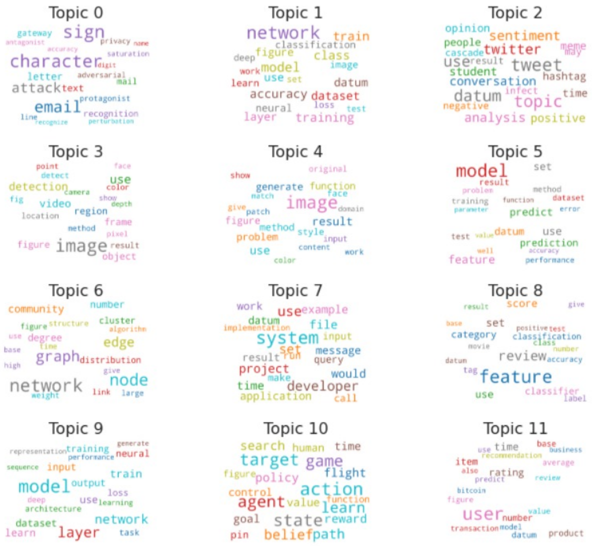


Figure 1: topics taken out by Topic Modeling

5 Literature Review

Topic modeling techniques is one kind of text mining tool that use to extract hidden semantic structure from paragraph or text.Based on LDA topic modeling many research had done.Like to analyse the up gradation or new exploration of different science articles from 1880 to 2000;Dynamic topic models (Blei and Lafferty, 2006),Also many recommendation system like content recommendation in newspapers or articles which was done over New York Times;Collaborative topic models (Wang and Blei, 2011).Popular unsupervised topic models such as Latent Dirichlet Allocation(LDA) and hierarchical models have been successfully applied to various publications such as The American Political Science Review and Science.

In Computational Linguistics, the only work of which we are aware is that of Hall et al. 2008 who study the history of ideas using LDA and topic entropy.In our project we are using LDA as a topic modeling technique as this is a process of identifying topics in set of documents.This can be useful for recommending the project by analysing topics which suits to the user requirement.And similar to LDA topic modeling we are using two other modeling name LSI and HDP (hierarchical direchlets principal) with the combination of bag of words and tf-idf as document terms.

6 Baselines Created

We have used implemented Latent Dirichlet Allocation (LDA) from the gensim library. The input

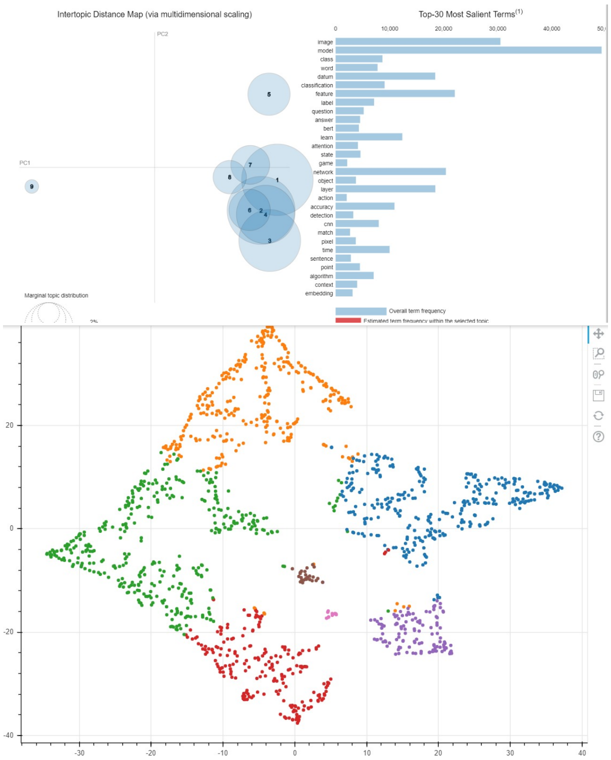


Figure 2: pyLDAvis and t-SNE plot for the document cluster

to this library has the list of word in each document. The list contains frequent bigrams phrases as well. The coherence score is used for evaluating the LDA model. Coherence score measures a single topic by measuring the degree of semantic similarity between topics that are semantically interpret able topics and topic that are artifacts of statistical inference.

7 Result Produced

1.fig3 graph represents the coherence score vs number of topics in LDA using trigram,as here in total 20 topics we get best coherence score hence we use total 20 topics.
2.fig4 graph represents the coherence score vs number of topics in LDA using bigram where 22 topics shows best coherence score 0.39
3.When user write input like subject which he/she require to get all the projets he/she gets all the subject name and links for that corresponding output(fig 6).
4.When we applied LDA afdter clustering then we get coherence score 0.41.

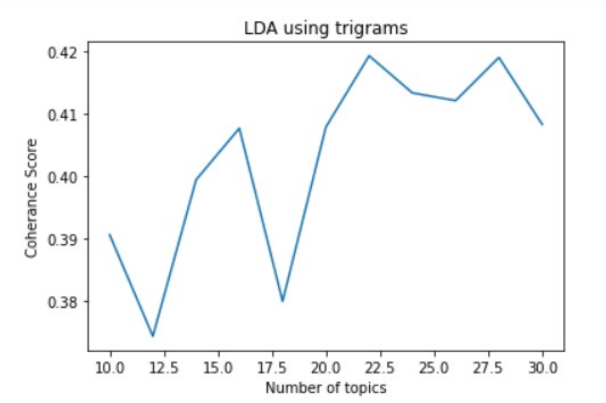


Figure 3: LDA trigram bag of words

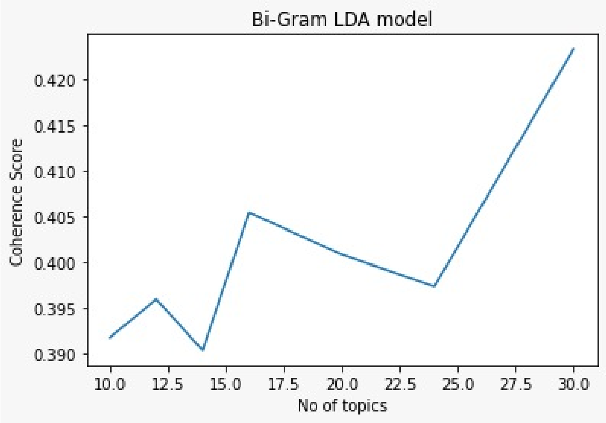


Figure 4: LDA tf-idf of unigrams bag words

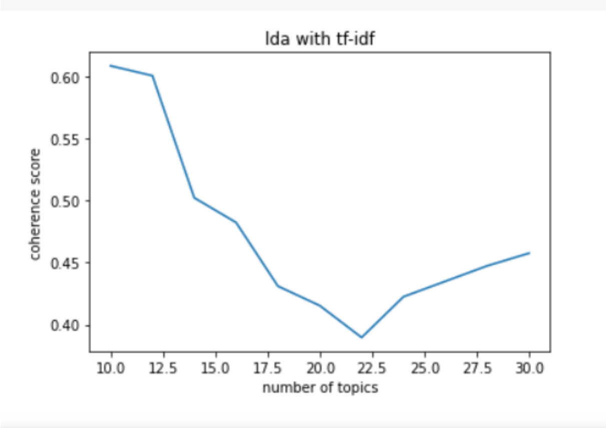
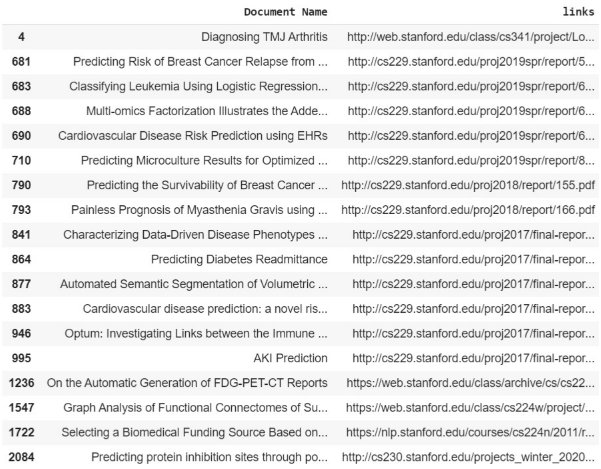


Figure 5: LDA tf-idf of unigrams



	Document Name	links
4	Diagnosing TMJ Arthritis	http://web.stanford.edu/class/cs341/project/Lo...
681	Predicting Risk of Breast Cancer Relapse from ...	http://cs229.stanford.edu/proj2019spr/report/5...
683	Classifying Leukemia Using Logistic Regression...	http://cs229.stanford.edu/proj2019spr/report/6...
688	Multi-omics Factorization Illustrates the Adde...	http://cs229.stanford.edu/proj2019spr/report/6...
690	Cardiovascular Disease Risk Prediction using EHRs	http://cs229.stanford.edu/proj2019spr/report/6...
710	Predicting Microculture Results for Optimized ...	http://cs229.stanford.edu/proj2019spr/report/8...
790	Predicting the Survivability of Breast Cancer ...	http://cs229.stanford.edu/proj2018/report/155.pdf
793	Painless Prognosis of Myasthenia Gravis using ...	http://cs229.stanford.edu/proj2018/report/166.pdf
841	Characterizing Data-Driven Disease Phenotypes ...	http://cs229.stanford.edu/proj2017/final-repor...
864	Predicting Diabetes Readmittance	http://cs229.stanford.edu/proj2017/final-repor...
877	Automated Semantic Segmentation of Volumetric ...	http://cs229.stanford.edu/proj2017/final-repor...
883	Cardiovascular disease prediction: a novel ris...	http://cs229.stanford.edu/proj2017/final-repor...
946	Optum: Investigating Links between the Immune ...	http://cs229.stanford.edu/proj2017/final-repor...
995	AKI Prediction	http://cs229.stanford.edu/proj2017/final-repor...
1236	On the Automatic Generation of FDG-PET-CT Reports	https://web.stanford.edu/class/archive/cs/cs22...
1547	Graph Analysis of Functional Connectomes of Su...	https://web.stanford.edu/class/cs224w/project/...
1722	Selecting a Biomedical Funding Source Based on...	https://nlp.stanford.edu/courses/cs224n/2011/r...
2084	Predicting protein inhibition sites through po...	http://cs230.stanford.edu/projects_winter_2020...

Figure 6: Recommonded result of subject biomedical application

8 References

1.https://github.com/kapadias/mediumposts/blob/master/nlp/publishednotebooks/Evaluate%20Topic%20Models.ipynb

2.https://www.machinelearningplus.com/nlp/topic-modeling-visualization-how-to-present-results-lda-models/ .https://towardsdatascience.com/end-to-end-topic- modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0

4.https://towardsdatascience.com/topic-modeling-and- latent-dirichlet-allocation-in-python-9bf156893c24

5.https://www.aclweb.org/anthology/R09-1061.pdf