**Career Village Recommender Engine**

**CKME136 – Capstone Project**

**Literature Review**

By Mona David

**Introduction**

These days young people have many decisions to make, many more students stay in education beyond the end of compulsory schooling and they are making these choices in the conditions of a rapidly changing labor market. Extensive engagement from people in work (employers, employees and professional bodies) ensures effective career guidance. Such exposure to the world of work enriches career guidance by presenting young people with authentic and trusted insights into jobs and how it relates to decisions about education and training.

There has always been a link between education and income. Salaries increase with each level of education and unemployment decreases with each level of education. Research has shown that providing career counseling interventions specifically for low-income students provides the support these students need to be successful in reaching their goals.

CareerVillage.org is remodeling how young people prepare for the world of work. With their crowdsourcing model, CareerVillage makes it possible for underprivileged youth to get the answer to any question about any career. Answers to students' questions are provided by an online body of volunteers from various walks of life. The content this community generates is indexed and searchable, making it possible for anyone to review the advice later on. The platform uses a Q&A style similar to StackOverflow or Quora to provide students with answers to any question about any career. Already over 500,000 people have used the career advice on CareerVillage to begin preparing for their futures.

Our Objective for the project is to develop a Recommender system to recommend suitable questions to the professionals who are most likely to answer them.

**Literature Review**

To understand the influence of e-mentoring on underprivileged youth I reviewed this study on Protégé Career Aspirations [3]. The study is the data from a 9 month e-mentoring program conducted by icouldbe.org(ICB) which is not for profit organization that creates and manages on-line adult-youth mentoring programs. ICB matches adult mentors with 10,000 students from schools internationally. This study indicated that career aspirations increased as a result of the mentoring process and specifically as a result of the increases in both general and career-based self-efficacy.

To better tackle our objective for the project I researched on how to process the natural language to interpret the questions and answers authored by the students and professionals. This paper by Elizabeth D. Liddy [6] lays the foundation for better understanding what Natural Language Processing (NLP) is and how its main purpose is Natural Language Understanding (NLU). This paper begins with defining NLP and briefs it’s history. The levels of NLP and different approaches to NLP are elucidated. Information Retrieval and Information Extraction are mentioned as the most frequent applications that use NLP.

The first step in Information Retrieval to isolate the “English Stop Words” i.e. the most frequently occurring words in English. I came across this paper by Christopher Fox from AT&T Bell Laboratories [4], where an assignment was conducted to generate a list of “Stop Words” from the Brown corpus. The Brown corpus was compiled in 1960 by Henry Kucera and W Nelson Francis at the Brown University, it consists of a million and fourteen words. As a result, we have 421 stop words which are highly effective in filtering the most frequently occurring and semantically neutral words in general English literature. The words have been attached in the appendix of the paper and raked for their frequency of occurrence.

The next step in the language was to lemmatize the words ,i.e. to return the word to its base form or dictionary form .I found this article by Vimala Balakrishnan, and Ethel Lloyd-Yemoh[1] on lemmatization which throws light on the importance of language modeling techniques and also compares the performance of Stemming and Lemmatization. The main questions of research in the paper were i) To compare the document retrievals using stemming and lemmatization and ii) compare the stemming and lemmatization techniques against a baseline ranking algorithm (i.e. no language processing). In this experiment the Porter’s stemming algorithm for stemming, LemmaGen for lemmatization and tf-idf was used for the baseline ranking algorithm. The retrieval performances were tested and comparisons were made using the Communications of the Association for Computing Machinery (CACM) collection. The MAP (Mean Average Precision) was calculated by dividing the average precisions with the number of queries. As a result, the study found that the language processing techniques improved the document retrieval and Lemmatization performed better compared to Stemming.

After the basic data processing techniques, now we get into the real deal of using similarity measure for Information Retrieval .This paper by 4 college students from Bharati Vidyapeeth’s college of engineering published in International Journal of Computer Applications[5] lays foundational background for better understanding of the Vector Space Model along with the TF/IDF weighting scheme, also the Cosine and Jaccard similarity measures with TF/IDF scheme are elucidated .It also introduces us to the another improved term weighting scheme TF-ATO, Term Frequency with Average Term Occurrences. The paper states the advantages of Vector Space Model over the Boolean Model and also describes the limitations of Vector Space Model.

Statistical Language models such as N-gram model has been dominating the language modeling field for almost three decades due to their simplicity and low computational complexity. The paper by William B. Cavnar and John M. Trenkle explains N-Gram Based Text Categorization [2]. It starts with an introduction to N-gram based similarity measures and takes use through the N-gram frequency statistics. They also test the N-gram based text categorization on a language classification task and discuss the advantages of N-gram categorization over other possible approaches. The key benefit of N-gram is, since every string is decomposed into small parts, any errors that are present tend to affect only a limited number of those parts, leaving the remainder intact. If we count N-grams that are common to two strings, we get a measure of their similarity that is resistant to a wide variety of textual errors. The authors collected 3713 language samples from the soc.culture newsgroup hierarchy of the Usenet to test the N-gram based text categorization. The results were the system yielded its best performance at a profile length of 400 N-grams. The system misclassified only 7 articles out of 3478, yielding an overall classification rate of 99.8%.

In recent times Neural network-based language modeling is gaining popularity. The authors Tomas Mikolov and colleagues describe how to effectively train neural network-based language models on large datasets [7]. For their study they use the Recurrent Neural Network language model (RNN LM). They performed recognition on the English Broadcast News (BN) NIST RT04 task provided to them by IBM containing 430 hours of audio which was roughly 400M tokens. Using the concept of ‘incremental learning’ that is learning simple patterns in the data before complex patterns can be learnt, the approach they took was to begin the training with out-of -domain data and end with the most important in-domain data. The study shows that training RNN model with direct connections can lead to good performance both on perplexity and word error rate, even if very small hidden layers are used.

I also wanted to look into Collaborative Filtering for Recommendation systems and that’s when I stumbled upon this survey by Xiaoyuan Su and Taghi M. Khoshgoftaar [8].In this paper they list out the main challenges of collaborative filtering such as data sparsity, scalability, synonymy, gray sheep, shilling attacks, privacy protection, etc., and their possible solutions. They present the 3 main categories of CF techniques which are memory-based, model based, and hybrid CF algorithms. To evaluate the prediction performance of CF, MAE (Mean Absolute Error) is used. Precision and recall, ROC sensitivity is used as a decision support accuracy metric.

This [9] paper of improved K-NN for Text classification based on clustering interested me due to the fact that it had enhanced results than the traditional KNN text classification algorithms. The researchers used an algorithm that doesn’t use all the training samples and got rid of the uneven distribution that causes multipeak effect. Afterwards it trained all the categories by K-mean clustering to get the cluster centers and they are used as new training samples and new weights were introduced to each cluster center and this was used in the algorithm. The training corpus had 20 categories with 9804 documents and the test set had 9833 documents. The experiments confirmed the effectiveness of this algorithm

In addition to the above articles I also found courses in E-learning platform Udemy to be quite helpful. The courses are mentioned below.

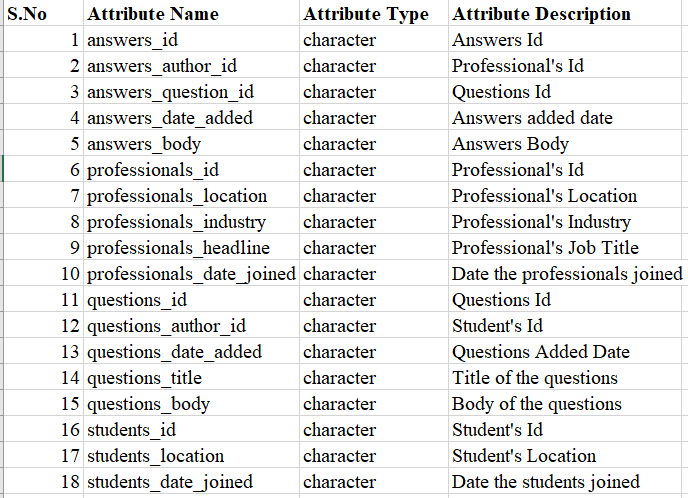
1.Building Recommender Systems with Machine Learning and AI

2.Word2Vec: Build Semantic Recommendation System with Tensor Flow.

**Data Description:**

The dataset for this project is from Kaggle competition<https://www.kaggle.com/c/data-science-for-good-careervillage/overview>.

The dataset has been provided by CareerVillage.org and contains 15 CSV files but our main area of interest lies in 4 CSV files named Students, Questions, Professionals and Answers. We prepare our dataset using these 4 files and it contains 51123 observations. Below are the attribute name and description.



**Approach:**

**Data Preparation**

**Exploratory Data Analysis**

Feature Selection/Engineering

Recommender Engine

Evaluation & Conclusion

**1.Data Preparation**

In this step the data is loaded and the initial cleaning of the raw data is performed. Any missing values are looked for. Since my data is already clean with a few missing values in professionals\_location and professionals\_industry, the missing values will be taken care of.

**2.Exploratory Data Analysis**

This stage will involve detailed analysis of the attributes.EDA will be performed on the data to answer trivial questions such as

a) Which location are most Number of Professionals from.

b) Which Industry do most professionals belong to.

c)What is the most common Job Title for the professionals.

d)Which location are most of the students from.

e) Create Word clouds to see the most common words in the questions\_body and answers\_body

**3.Feature Selection/Feature Engineering**

This step involves feature selection based on the Exploratory Data Analysis and features that do not provide any useful information such as date will be dropped off.

And then in this stage we remove the stop words, the alphanumeric characters and lemmatize the columns. Then we merge the columns to create the corpus of text.

**4.Build Recommender Engine**

In this step the corpus from the above step will be used to build the below engines.

1. Content based Recommender with TF-IDF.
2. Content Based Recommender with KNN.
3. Word2Vec.

Different models will be built based on the corpus.

**5.Evaluation and Conclusion**

The Recommender systems will be evaluated for accuracy and conclusions will be drawn based on the evaluations.

**Bibliography**

1. Balakrishnan, Vimala, and Ethel Lloyd-Yemoh. "Stemming and lemmatization: a comparison of retrieval performances." (2014): 174-179.
2. Cavnar, William B., and John M. Trenkle. "N-gram-based text categorization." *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*. Vol. 161175. 1994.
3. DiRenzo, Marco S., Christy H. Weer, and Frank Linnehan. "Protégé career aspirations: The influence of formal e-mentor networks and family-based role models." *Journal of Vocational Behavior* 83.1 (2013): 41-50.
4. Fox, Christopher. "A stop list for general text." *Acm sigir forum*. Vol. 24. No. 1-2. New York, NY, USA: ACM, 1989.
5. Jain, Abhishek, et al. "Information retrieval using cosine and jaccard similarity measures in vector space model." *International Journal of Computer Applications* 164.6 (2017): 28-30.
6. Liddy, Elizabeth D. "Natural language processing." (2001).
7. Mikolov, Tomáš, et al. "Strategies for training large scale neural network language models." *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2011.
8. Su, Xiaoyuan, and Taghi M. Khoshgoftaar. "A survey of collaborative filtering techniques." *Advances in artificial intelligence* 2009 (2009).
9. Yong, Zhou, Li Youwen, and Xia Shixiong. "An improved KNN text classification algorithm based on clustering." *Journal of computers* 4.3 (2009): 230-237.