Syracuse University, School of Information Studies

M.S in Applied Data Science

Portfolio Milestone

Yibo Feng

SUID: 57533926

<https://github.com/yfeng0308/MSADS_Portfolio>

**1. Introduction**

The Applied Data Science program at Syracuse University’s School of Information Studies provides students the opportunity to collect, manage, analyze, and develop insights using data from a multitude of domains using various tools and techniques. Courses like Database Administration (IST659), Natural Language Processing (IST664), Data Analytics (IST707) and Big Data Analytics (IST718). Project reports and presentations deliver insight using tools such as SQL Server Management Studio, R Studio, Spyder. During the time of program, these courses also give me opportunities to be familiar with popular programming language such as Python and data analyzing language like R. The Applied Data Science program has seven learning objectives, and these will be exemplified by the project review in this portfolio:

* Describe a broad overview of the major practice areas in data science.
* Collect and organize data.
* Identify patterns in data via visualization, statistical analysis and data mining.
* Develop alternative strategies based on the data.
* Develop a plan of action to implement the business decisions derived from the analyzes.
* Demonstrate communication skills regarding data and its analysis for relevant professionals in their organization.
* Synthesize the ethical dimensions of data science practice.

**2. IST 659: Database Administration**

**Project Description**

This project focuses on designing a database for Syracuse University iSchool course catalog and course registration system. The iSchool is offering a course catalog which entails available classes of different levels pertinent to various information areas. Currently, the course catalog can be accessed through 2 separate routes which simply display a list of available courses and their basic information (course code, instructor, overview, and pre-requisite course). Apart from the course catalog, the course registration system is integrated with the class searching system called ‘Myslice’ where users can check their current credit records and view available courses. In conclusion, the current course catalog and course registration system are separate, and they are not capable of providing one single platform for comprehensive data sources, resulting in user inconveniences. Therefore,the purpose of this system development is to establish an integrated iSchool course catalog database and link it to the course registration system for enhancing user conveniences.For instance, students are expected to shorten their time for browsingand selecting courses they are interested in or required to take and make better course registration quality-wise. Also, administrators can analyze such data more conveniently and make use of the analysis for improving iSchool academic course design.

Logical model and ER diagram were developed to organize the relationships between Job, JobOffered, Company, ClassRoom, Instructor, Class, TechTool, Course, Student, Registration. Tables were created in SQL Server Management Studio and data population was accomplished by Microsoft Access. As Figure 1 shows that, this is the visualization network of ten tables. Building several forms to display multiple information. One is displaying details of classes for a specific course. One is displaying available job information for students.

Chart, box and whisker chart

Description automatically generated

**Figure 1:** Logical Model/ER Diagram

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated

**Figure 2:** Form Examples

**Reflection & Learning Goals**

This course provides opportunity to develop a data management solution which reveals the importance of data is stored and accessed. This project contributed to the ability to deliver actionable insight to the field of program management. This database help students to view their profile as well as credit status and help instructors notify course updates and companies could use this system to find potential employees by querying it. This is also significant for data analysts and data scientists. Overall, this project contributed successful application of learning goals such as collect and organize the data and identify the pattern via visualization. Thinking of three perspective point (Student, Instructor and Company) reveals there is a plan of action to implement business decisions.

**3. IST 664: Natural Language Processing**

**Project Description**

With more and more text data generated by the Internet, the problem of text information overload is becoming more and more serious. Text summaries are designed to transform text or text collections into short summaries containing key information. According to the output type, it can be divided into extractive abstract and generative abstract. Extractive abstract extracts keywords from the source document to form an abstract. Generative abstracts generate new words and phrases based on the original text to form an abstract. Broadly speaking, there are two ways to summarize the text: Extraction-based summarization and Abstraction-based summarization. This project is to use extraction-based summarization technique based on two algorithms: TF-IDF and TextRank. The goal is to figure out which algorithm is suitable to summarization of News.

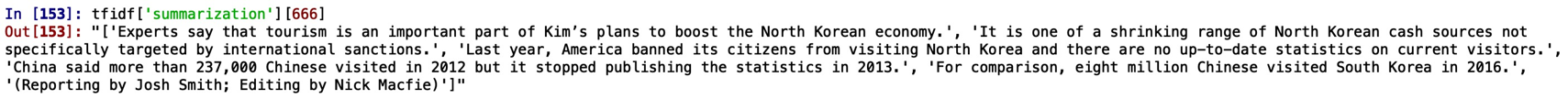
Based on article “Text Summarization Techniques: A Brief Survey”, text summarization got attraction in the 1950s. There was a significant research which explained a method to extract sentences from text or documents using features like word or phrase frequency. TFIDF is such an algorithm mainly using word frequency to measure sentences’ weights and extract sentences. The reason we choose TextRank because the core of it is to build similar matrix. Similarity of two sentences in similar matrix is calculated by computing cosine similarity with TFIDF weights for words. It could conclude that two algorithms have connections but run with different tracks. Therefore, our goal using them is to see how different result they could bring. Figure 3 is specific process for TF-IDF and Figure 4 is for TextRank.



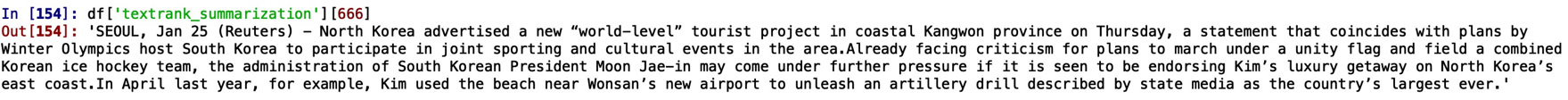
**Figure 3:** Text Summarization Process in TFIDF



**Figure 4:** Text Summarization Process for TextRank



**Figure 5:** Example Summarization for TFIDF



**Figure 6:** Example Summarization for TextRank

**Reflection & Learning Goals**

Based on Figure 5 and Figure 6, It is clear that summarization for TextRank extract summarization is more accurate than it for TFIDF. The reason is that the first sentence of summarization for TextRank shows key word such as North Korea, World-Level and tourist project. However, it is hard to extract these words if we see the first several sentences from summarization for TFIDF. Also, I have to say there exist limitation for these two algorithms and dataset. It is because it is hard to use existing columns from dataset to evaluate the accuracy of all summarization based on extraction summarization.

Text summarization is a common problem in natural language processing. In this project, we preprocess dataset by sentence tokenization, removing punctuation as well as number and special characters, using all lowercase and removing stopwords. and use ROUGH to evaluate two models’ performance. Compared the results of two model, the TextRank has a better performance than TF-IDF.

Overall, this project contributed successful application of learning goals such as collect and organize the data. Although all collected data is public record, consideration must be made to ensure that only the relevant information is requested to both balance request limitations and user privacy.

**4. IST 707: Data Analytics**

**Project Description**

Through studying Data Analytics under the direction of Prof. Jesse Cases, various data mining techniques were introduced which perform with varying precision and efficiency for applications in regression, classification, and clustering. The project dataset I chose collected top products information including ratings and sale performance from E-commerce platform “Wish”. At the beginning of data mining, I started to work on product segmentation. The goal is to help the platform improve their customer base, work on target areas, and segment customers based on purchase history and interests. Besides, product segmentation will help the platform find out the common points and different types of product, then they will be able to set target sale policy. I was also interested in what factors contribute to high rates of rating with five count which means the percentage of rating 5 in total rating count is higher than 50%. This will contribute to improving customer preference. I also want to analyze potential factors that help increase the sale of products. For example, it is interesting whether the merchant rating will influence the sale of products. By analyzing these factors, the platform will be able to predict sales status of each product.

This data required the cleaning and preprocessing of dataset which included removing irrelevant columns and reformatting category feature to numeric ones. This project included analysis from The project included analysis from three directions. The first one was product segmentation. The data mining task is cluster analysis by using K-Means. From the left side of Figure 7, I chose 4 as elbow point and put it into K-Means model. The right side of Figure 7 was clearly showing that the clusters were not clear and there were a lot of outliers. Then, we chose another model hierarchical cluster and used Euclidean distance to build the distance matrix. The final accuracy for this model is 0.998.

Chart, line chart

Description automatically generatedChart

Description automatically generated with medium confidence

**Figure 7:** Cluster Model using K-Means

The second direction was to find factors for High Rating Rates. The specific preprocessing was to reformat category variable as factors and convert numeric variables into category based on their 1st and 3rd quartiles. The data mining task used in this part is Association Rule. Based on following rule from Figure 8, it could conclude that products with badges of product quality would contribute to gaining high rate of 5 ratings. Badges of fast shipping and local product wouldn’t influence the high percentages of the rating 5. Origin country of China would contribute to high level of rating.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| lhs | rhs | support | confidence | lift |
| {badge\_product\_quality=1, badge\_fast\_shipping=0, shipping\_is\_express=0} | {rating\_level=high} | 0.07138092 | 1 | 2.653097 |
| {badge\_local\_product=0, badge\_product\_quality=1, badge\_fast\_shipping=0, shipping\_is\_express=0} | {rating\_level=high} | 0.06871247 | 1 | 2.653097 |
| {badge\_local\_product=0, badge\_product\_quality=1, badge\_fast\_shipping=0, shipping\_is\_express=0, origin\_country=CN} | {rating\_level=high} | 0.06604403 | 1 | 2.653097 |
| {badge\_product\_quality=1, badge\_fast\_shipping=0, shipping\_option\_price=<3} | {rating\_level=high} | 0.06204136 | 1 | 2.653097 |

**Figure 8:** Association Rule result

The final direction was to make sale status prediction. During preprocessing, I selected variables that might contribute to classifying high sales and set high ‘units\_sold\_level’ as 1 while others as 0. The main data mining task was to implement classification algorithm such as SVM and ANN. In this part, four model including KNN, Random Forest, SVM and ANN would be built to compare result with each other. Figure 9 provided result from each model. Since the training accuracy and prediction accuracy of Random Forest differed a lot, there may exist overfitting in the model. Besides, based on dataset description I found the positive value is 0, and what met the goal was to classify unit\_sold\_level=1. Therefore, recall and F-measure could be better to evaluate the models. According to this, SVM will be a better model for this dataset.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Training Accuracy | Prediction Accuracy | Precision | Recall | F-measure |
| KNN | 0.6974428 | 0.6756 | 0.6973 | 0.9100 | 0.78957631 |
| Random Forest | 0.7958761 | 0.7023 | 0.9744 | 0.5700 | 0.71925408 |
| SVM | 0.6691738 | 0.6689 | 0.6689 | 1 | 0.80160585 |
| ANN | 0.6791207 | 0.6622 | 0.7071 | 0.8450 | 0.76992397 |

**Figure 9:** Classification result

**Reflection & Learning Goals**

From the first direction, I found the unexpected result compared with class example using cluster model. It illustrated it is important to test different data mining techniques to find out the simplest, most accurate prediction models. Using alternative strategies and weighting the benefit of each technique with combination of characteristics of dataset could make the whole process more effective and convincible and provide higher precision in data mining tasks. This project contributed to successful application of the learning goal by using alternative strategies based on data and demonstrate communication skills regarding data and its analysis. Combination of data mining and visualization to identify patterns in the data were used in classification tasks.