Paciolan Project Proposal

Project Title: Paciolan Custom Fields
Student Names: Yihan Wang, Sirui Hu, Yadi Yang

1. Project Summary

Paciolan is a company aiming to serve university sport/performing art organizations (**hereinafter called "clients"**), providing the clients' fans and patrons a better place to enjoy the ticketing service. The main goal of this project is to derive business insights from the custom fields/keywords/tags assigned by clients to patrons and donors. One question we can address is "How do clients group or name the custom fields?" To answer this question, our primary approach is to find out the similarities within the users assigned under some common custom fields. By processing and training the given results in a machine learning pipeline, we will be able to gain some useful insights around. We are expecting to find out the naming convention of the custom fields, and possibly unify the custom fields for better classifications. Our ultimate goal is to automate the classification/tagging process and optimize the user experience by recommending tagging suggestions based on the subjects to be grouped.

2. Proposed Technical Approach

The data we will be using in this project will be provided by Paciolan's internal database. The main components of the data are divided into three main parts: 1) five user's custom fields, 2) clients, patrons, and donors' demographics, 3) ticket purchasing history, and 4) ticket information. With these given datasets, next step we are going to run a data analysis pipeline, applying machine learning methodologies to explore potential solutions of user tagging optimization (Fig I.).

Our data processing pipeline will include three major parts: 1) data cleansing, 2) data modeling and predicting, and 3) data visualization and evaluation. In the first step, we plan to generally clean the raw data and handle some missing info issues to make sure the data is organized and ready to go into the modeling process. Next step, we plan to proceed with both supervised and unsupervised learning analysis, fitting our data into different models such as Logistic Regression (supervised), Random Forest (supervised), and KMeans clustering (unsupervised). Specifically, in supervised learning, we will try to predict whether a patron will be given a certain tag with the patron's demographic and transaction data. Also, we would like to do a PCA analysis beforehand, since feature reduction could be a good approach to remove some biased variables that may affect our later predictions. In the final step, we will be generating visualizations such as cluster plots with classified common tags/user groups and the AUC-ROC curve for data evaluation, etc. Lastly, we will be utilizing these visual outputs to make some decisions about the tagging recommendations.

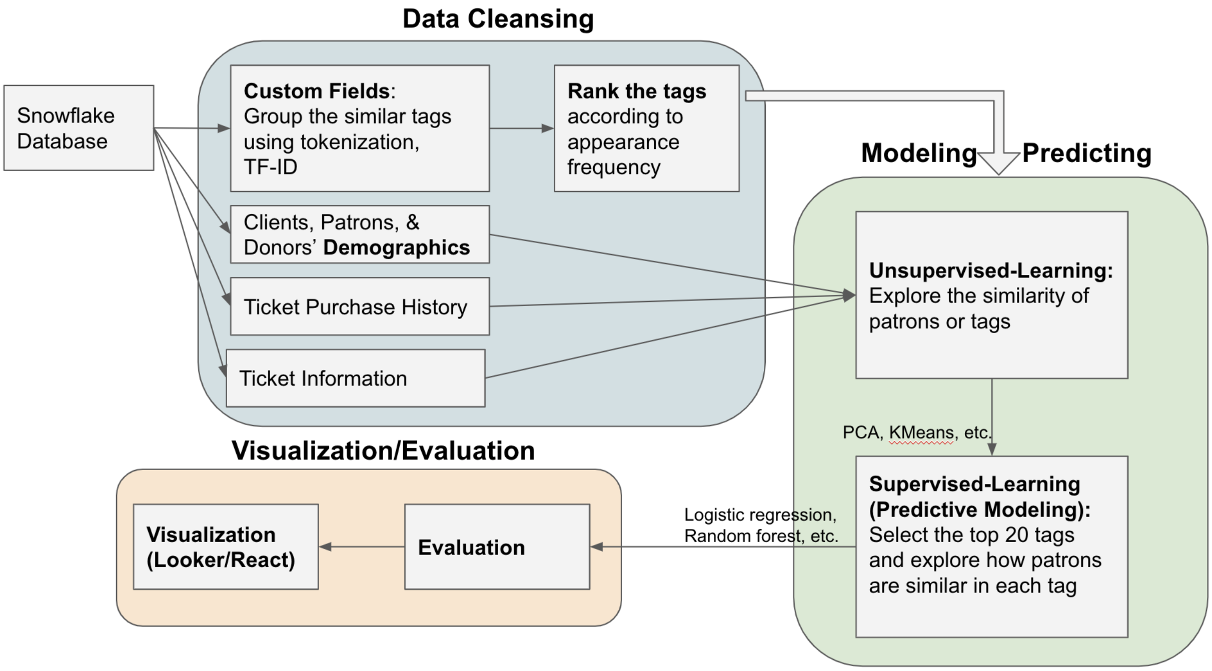
Data Cleansing
Snowflake Database
Custom Fields: Group the similar tags using tokenization, TF-ID
Rank the tags according to appearance frequency
Clients, Patrons, & Donors' Demographics
Ticket Purchase History
Ticket Information
Modeling
Predicting
Unsupervised-Learning: Explore the similarity of patrons or tags
PCA, KMeans, etc.
Supervised-Learning (Predictive Modeling): Select the top 20 tags and explore how patrons are similar in each tag
Logistic regression, Random forest, etc.
Visualization/Evaluation
Visualization (Looker/React)
Evaluation

Fig I. Data Analysis Pipeline.

Besides the pipeline procedures, we are also considering another feasible approach as our next step (if time allows) that may contribute to the future development of a patron-side recommendation system. This approach would allow the system to suggest relevant products/services based on users' preferences (tags). A desirable recommendation page would be containing popular events (maximum expectation strategy) and events near the users (demographics information) to a new-signup user. Also, it can contain some similar events to those events that users already bought tickets (similar to their preferences) for an old user. For users that already have a lot of information, general steps we can follow:

A collaborative method based on past transactions:

- Use purchasing history to build a user-item interaction matrix
- Calculate the similarities of users by users' distance and correlation
- Provide a recommended content of events based on nearest neighbors, matching each user's preference

This part is our ultimate goal. We are not sure whether we can accomplish that due to the time limits and technical difficulties, but we will try to implement this.

3. Data Sets

The company Paciolan used Snowflake to store their data mainly in JSON format. We will be working with 4 types of data: 1) 5 user custom fields, 2) demographics of clients, patrons, and donors, 3) ticket purchasing history, and 4) ticket information. So far we only obtained samples for 1) and 3), but we are still waiting for samples from Paciolan on 2) and 4). There are in total five custom fields we need to learn the business insights from. Below are the descriptions and sample data for each of them:

1) **Keywords**: Keywords are on the Account records and help Paciolan users' segment accounts into lists for marketing
   Sample:
```
"keywords": [
"MARTIL",
"P9999999999",
"PLMMARTIN4@UH.EDU"
]
```

2) **Tags**: Tags also are on the Account records and help Paciolan's NumPyent accounts into lists for marketing
   Sample:
```
 "tags": [
   "F16REN",
   "F16LINKEDBAL"
 ]
```

3) **User defined fields**: User defined fields allow Paciolan's clients to create new data points on account records to track additional data
   Sample:
```
"userdefinedfields": [
   {
     "value1": "CPP",
     "value2": "0"
   },
   {
     "value1": "APP",
     "value2": "0"
   }
```

```
 ]
```

4) **Donor Categories**: Donor categories allow Paciolan’s clients to categorize their donors with custom labels that introduce the concept of time as well. This means clients can have categories with expiration dates.
Sample:

```
"categories": {
   "FB18": "1",
   "G": "",
   "GP93": "",
   "MAG": "",
   "MC": "",
   "STP17": "640",
   "TKNPS5": "80",
   "TKPS10": "520"
 }
```

5) **Motives**: Motives give Paciolan’s clients the ability to create custom campaign tags and assign a goal to them. The motives can be linked to transaction records to see how they impact the overall goal.
Sample:

<u>Motives Record</u>

```
{
 "dbid": "NCSU",
 "description": {
   "en_US": ""
 },
 "goal": "0",
 "id": "CR",
 "name": {
   "en_US": "Correction"
 },
……
}
```

<u>Motives Tagged on a Transaction</u>

```
{
 "accountdbid": "MSSTATE",
 "accountid": "799439",
 "allocationid": "BDC",
 "allocgroupid": "BDC",
 "batchid": "4",
 "channelid": "MSSTATE",
 "comments": "BDC - GENERAL DONATIONS",
 "creditamount": "0",
 "dbid": "MSSTATE",
 "driveyear": 2003,
 "fundtype": "S",
 "giftid": "55407",
 "id": 1,
 "matchgiftaccountid": "",
 "matchpledgeamount": "0",
 "motiveid": "CR",
 "operation": "I",
 "parenttransactiontype": 0,
 "paymentamount": "25000",
 "paymentapplyamount": "0",
 "pledgeamount": "0",
 "previousdonationamount": "0",
 "previouspledgeamount": "0",
 "printfl": false,
 "receipteddonorid": "799439",
 "receiveddate": "2003-08-07T05:00:00.000Z",
 "renewable": false,
 "rolledover": false,
 "scheduledtoggle": false,
 "sendemailconfirmation": false,
 "seq": 89,
 "sourceid": "",
 "transactionid": "193481",
 "transactionitemtype": "F",
......
}
```

4. Experiments and Evaluation

Our plan made for the unsupervised learning is to cluster the patrons and show the groups of patrons to customers and have qualitative feedback on how much they agree with the clustering. The prediction modeling can be binary, and we plan to evaluate the results by precision and recalls. We will add the regularization to the logistic models to reduce overfitting, and we will apply 5-fold cross-validation to find the best model. This result will also be shown to the company to get feedback from the expertise about the reasonability of the result.

For the recommendation system, we can use metrics such as mean average precision and mean reciprocal rank to evaluate the performance.

5. Software

Public Available Sources:

Database: Snowflake
Compiler/IDE: Jupyter Notebook
Packages:
Data Wrangling: numpy, pandas, json
Data Analysis: scikit-learn, scipy, statsmodel
Data Visualization: matplotlib, seaborn, plotly
Present Final Result: Looker (snowflake) or React.
Repository: GitLab/GitHub
Communication: Jira (Agile project management) + Confluence, Slack, Zoom

Personal Available Sources:

Programming Language: Python, R (Optional)
Compiler/IDE: Visual Studio Code, Jupyter Notebook, R Studio (for statistical analysis only)

6. Milestones

- Winter
  - Weeks 8-10
    - Set up accounts and retrieve the data
    - Explore the raw data
    - Clarify and keep updating milestone goals
    - Research relevant literature and methods to come up with some initial ideas.
    - Draft and prepare for the proposal presentation.
- Spring
  - Weeks 1-2
    - Set up a weekly agenda and recurrent zoom meetings with the Paciolan team.
    - Preprocess the raw data and start cleaning and organizing the data for future use.
    - Brainstorm and experiment on any possible approaches.
    - Evaluate the results of experiments.
    - Discuss with the sponsor and know their preference

- Weeks 3-4
    - Continue brainstorming and experimenting on any possible approaches.
    - Evaluate the results of experiments.
    - Update and deliver any innovative insights with the sponsor.
    - Get suggestions from professors and the Paciolan team
- Weeks 5-6
    - Continue brainstorming and experimenting on any possible approaches.
    - Evaluate the results of experiments.
    - Update and deliver any innovative insights with the sponsor.
- Weeks 7-8
    - Automate predictive modeling processes and gain insights from the results.
    - Evaluate the results of experiments.
    - Use Looker/React to build data visualization for the deliverable results
    - Start writing the final report
- Weeks 9-10
    - Optimize results and prepare for the final presentation
    - Finalize the final report