EECS6414: Data Analytics & Visualization

Sample Past Project Ideas
Group Pattern Discovery of Pedestrian Trajectories

Project 1
Pedestrian trajectories
Network-based Analysis of Public Transportation Systems

Project 2
Transit Networks as Graphs

- Graph model result
- Example:
  - San Francisco
Analysis of the DBLP Collaboration Network

Project 3
Core Behavior Analysis of NYC Taxi Trips

Project 4
Motivation

• Taxi trip data analysis
  – Study individual taxi
  – Find local patterns
  – Find global patterns
Visualization trials and Empirical Observations
Scene Classification

Project 6
Motivation

Parking lot
Highway
Beach
Garage
Street
Itinerary Planning Using Genetic Algorithm

Project 7
Itinerary Planning Problem

Sights
- Central Park
- Statue of Liberty National Monument
- The High Line
- Empire State Building
- Time Square
- Brooklyn Bridge
- Rockefeller Center
- The Metropolitan Museum of Art
- One World Trade Center
- Grand Central Terminal
User Profiling on News Media

Project 7
Motivation
Interactive Analysis and Visualization of Large-scale Transit Data

Project 8
Motivation
Modeling and Forecasting Cryptocurrency

Project 9
Bitcoin & Other Cryptocurrencies
CDiscount Image Classification Challenge

Project 10
Motivation

1\textsuperscript{st} - $20,000
2\textsuperscript{nd} - $10,000
3\textsuperscript{rd} - $5,000

OR

Automatically classifying incoming products based on visual information
Winter 2019
Kaggle Kernel Mining

Project 1
GitHub vs. Kaggle

GitHub

• For software developers in general
• Create, fork, pull ... on repositories
• Get involved in large OSS projects

Kaggle

• Mostly for data scientists
• Create, fork ... on kernels
• Take part in competitions
Generating Tactile Graphs in MsExcel

Project 2
Methods

A macro for Microsoft Excel to automatically export data to an OpenSCAD program, which would generate and print a tactile visualization of the data.

A code that converts data into a 3D printable file

Picture of the pie chart, ready for 3D-printing.
Classification of Brain-wave Data for a Virtual Reality Application

Project 3
Methods

**PROBLEM DEFINITION**

Validating the precision rates of the meditation levels achieved from MUSE by comparing them with our deterministic model.

This is a classification problem!!
UK Housing Prices Analysis and Prediction

Project 4
## Results

### Forecast for top 10 major cities from 2019 to 2020

<table>
<thead>
<tr>
<th>City</th>
<th>December 2020</th>
<th>Forecast (2019-2020)</th>
</tr>
</thead>
<tbody>
<tr>
<td>London</td>
<td>£882732</td>
<td>12.98%</td>
</tr>
<tr>
<td>Bristol</td>
<td>£306110</td>
<td>9.10%</td>
</tr>
<tr>
<td>Manchester</td>
<td>£194135</td>
<td>7.61%</td>
</tr>
<tr>
<td>Southampton</td>
<td>£228141</td>
<td>6.67%</td>
</tr>
<tr>
<td>Birmingham</td>
<td>£200505</td>
<td>6.26%</td>
</tr>
<tr>
<td>Leeds</td>
<td>£195442</td>
<td>6.13%</td>
</tr>
<tr>
<td>Sheffield</td>
<td>£173098</td>
<td>4.26%</td>
</tr>
<tr>
<td>Nottingham</td>
<td>£150079</td>
<td>3.91%</td>
</tr>
<tr>
<td>Liverpool</td>
<td>£137795</td>
<td>1.91%</td>
</tr>
<tr>
<td>Sunderland</td>
<td>£116784</td>
<td>1.83%</td>
</tr>
</tbody>
</table>
Visualizing Deep Neural Networks Decision for Medical Imaging

Project 5
Results – Prediction Difference Analysis

• In this method, we occlude different parts of image and then track the changes in the output probability of the network

• By doing so, we can generate heatmaps of the parts of the image which are contributing for or against a class
Canadian Real Estate Rental market - Artificial Hype or Reality? Airbnb vs Traditional Rental Market

Project 6
Airbnb Demand Analysis (2015 to 2019)

D3 radial column graph. It shows the percentage of listings that become unavailable in Toronto per month, through 2015 to 2019 with spikes from May to September at Summer times.
Go-transit Data Analysis and Visualization

Project 7
Motivation

GO Transit serving Greater Golden Horseshoe carrying 70 million passengers over the year

Why GO Transit Data analysis?

Scenario 1: Jam Packed Buses

Scenario 2: Empty Buses

- Facilitate strategic decision making
- To understand weaknesses and strengths of organization
- Provide better and logical insights of operations
- To optimize daily operations
- To understand traffic patterns and direct influencers
- Patterns of spatial and temporal movement of passengers
York Community: Demographics, Analysis and Housing Location Recommendation for New Students

Project 9
Results

Distribution of YorkU Students

Commuting time needed for each Region
Modeling and Visualization of User Interest in Reddit

Project 10
Overall distribution of geographical location in all the submissions
Mining Opportunities in an Online Shopping Store

Project 11
Mining Online Shopping Store

01 Introduction & Research Questions
02 Data Sources and Cleansing
03 Speed and Security
04 Recommendation System
05 Next Steps
Drone Trajectory Analysis and Visualization

Sensors Raw Data

- UWB Receiver (40 Hz)
- IMU (100 Hz)

Ranges

- Pre-Process Filtering
- Multilateration Algorithm
- Levenberg–Marquardt Algorithm

Clean Ranges

- Height

Optimized XYZ

Post-Process Filtering

XYZ

Data Acquisition System for UWB-based UAV Positioning

- UWB Tag
- IMU

Data
- UWB ranges
- IMU data
- Ground-truth UAV position

Open Field

Indoor

Inside Tunnel

Under Bridge

Near Building

UWB anchor

Set 1

Set 2

Set 3

Set 4

Set 5
Yelp Dataset Analysis for Restaurant Improvement Recommendations

Project 2
Restaurant Insights

How is my business doing?

Total Restaurants (#): 17,941
Reviewed by (# users): 622,280
Total Reviews (#): 1,580,000

User Sentiments:
- Positive (1,020,000, 64.96%)
- Negative (0.550,000, 35.04%)

Find a Restaurant:
- #1 Sushi
- .99 Cent Pizza Place
- @ Dawn
- 1 Plus 2 Pizza & Wings
- 1 Pot
- 100% Natural Mexican Grill
- 1000 Grammes
- 100s...

Sentiment | Year | Month | Day | Review Text
--- | --- | --- | --- | ---
Positive | 2004 | December | 19 | Frequently busy due to their great food, but the wait moves quickly. If you go at a peak time, know that you aren't going to have time to sit around and be coddled by the wait staff as quickly as possible, because there are likely other people who have just come in. Even so, the staff is always courteous and helpful, and they are always happy to help you order something spicy, make sure that you ask for extra sauce, and not overboard on flavor. There's plenty of it in the mix, it's perfect. The large amount of spice in their dish. Think more along the lines of Chinese food. It's good, but perhaps not really spicy.
Positive | 2004 | December | 19 | Not the best part of town. Not particularly good or particularly bad. But the variety is awesome, and unlike many places...

Business Health (based on User's Opinion):
- Average of Star Ratings by Users

YEAR

Average Ratings: 1.2, 1.8, 2.4, 3.0, 3.6, 4.0, 3.8, 3.6
Enhanced Fraud Detection using IEEE CIS dataset

Project 3
Enhanced Fraud Detection

Classifier Selection

- Logistic Regression
- KNN
- SVM
- Random Forest
- XGBoost
- Neural Networks
Toronto BikeShare Data Analysis

Project 4
Bike Sharing Analytics

• Who is using Bike Share? Are there distinct differences between members vs. casual users behaviours?
• When are people using Bike Share? How does usage change across the year, the week, and the day?
• How is bike share being used? Commute to work or to explore the city? How does the weather change the way people
Do Torontonians need the Proposed Pickering Airport

Project 5
Structural Analysis of the Github Network

Project 6
Network Analytics

- Can we detect important developers and structure of the communities in github?
- Can we recommend possible projects that a user would be interested in watching?
- What are the underlying insights in the Github data?

<table>
<thead>
<tr>
<th>Rank</th>
<th>Username</th>
<th>PageRank $\times 10^4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>torvalds</td>
<td>8.82</td>
</tr>
<tr>
<td>2</td>
<td>JakeWharton</td>
<td>6.93</td>
</tr>
<tr>
<td>3</td>
<td>michaelliao</td>
<td>5.25</td>
</tr>
<tr>
<td>4</td>
<td>githubpy</td>
<td>4.44</td>
</tr>
<tr>
<td>5</td>
<td>Tj</td>
<td>3.84</td>
</tr>
<tr>
<td>6</td>
<td>mojombo</td>
<td>3.30</td>
</tr>
<tr>
<td>7</td>
<td>paulirish</td>
<td>3.24</td>
</tr>
<tr>
<td>8</td>
<td>defunkt</td>
<td>2.93</td>
</tr>
<tr>
<td>9</td>
<td>ruanyf</td>
<td>2.90</td>
</tr>
<tr>
<td>10</td>
<td>gaearon</td>
<td>2.76</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Username</th>
<th>Followers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>torvalds</td>
<td>80184</td>
</tr>
<tr>
<td>2</td>
<td>JakeWharton</td>
<td>48120</td>
</tr>
<tr>
<td>3</td>
<td>ruanyf</td>
<td>39102</td>
</tr>
<tr>
<td>4</td>
<td>Tj</td>
<td>37402</td>
</tr>
<tr>
<td>5</td>
<td>addyosmani</td>
<td>32666</td>
</tr>
<tr>
<td>6</td>
<td>paulirish</td>
<td>29690</td>
</tr>
<tr>
<td>7</td>
<td>yyx990803</td>
<td>29200</td>
</tr>
<tr>
<td>8</td>
<td>gaearon</td>
<td>27415</td>
</tr>
<tr>
<td>9</td>
<td>sindresorhus</td>
<td>25701</td>
</tr>
<tr>
<td>10</td>
<td>mojombo</td>
<td>25112</td>
</tr>
</tbody>
</table>
Earthquake Analysis

Project 7
Earthquake Analysis

Analyzing the behaviors and patterns of earthquakes can help us in better preparation and crisis management.
Question Answering using Deep Learning

Project 8
Question Answering using Deep Learning

- Multi-hop QA task: for every question, the model needs at least two sentences (supporting facts) in two different documents to answer
- What do we need to do?
  - Find gold documents in the given 10 documents.
  - Find all sentences that help to answer the questions in the gold docs.
  - Reason over these supporting facts to predict the answer.

**Paragraph A, Return to Olympus:**
[1] Return to Olympus is the only album by the alternative rock band Malfunkshun. [2] It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990. [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

**Paragraph B, Mother Love Bone:**

Q: What was the former band of the member of Mother Love Bone who died just before the release of “Apple”?
A: Malfunkshun

**Supporting facts:** 1, 2, 4, 6, 7
Exploring Stack Overflow: Answer Time Prediction and User Profile Analysis

Project 9
Problems

Unanswered Questions
- Answer time
- Prediction [1]

Tags Accuracy
- Tags
- Prediction [2]

Answer Quality
- User Profile
- Analysis [3]


Graph-based Identification of Attacks in Anomalous Traffic

Project 10
**Motivation**

DDoS attacks have increased in sophistication and intensity, threatening the availability of enterprise services, applications, websites, and networks.

Most survey respondents (86%) report experiencing at least one DDoS attack in the past 12 months.
Data Analytics and Route Navigation in Toronto Based on Accidents Data

Project 11
Navigation that Avoids High-risk Areas

Accident areas after clustering

Suggested Navigation Route
Impact of Industry on Computer Vision Research

Project 12
Results

Affiliation network: 2010
Bitcoin Network Analytics and Machine Learning

Project 13
Motivation

Bitcoin (₿)

• Decentralized digital currency without a central bank that can be sent from user to user on the peer-to-peer bitcoin network without the need for intermediaries
• Exponential rise in use and interest since 2010
• High volatility in stock prices
Adaptive Traffic Signal Control

Project 14
Motivation

Two traffic movements (dashed arrows) are allowed in phase #2

\[ p_2 = \{ (l_W, m_N), (l_E, m_S) \} \]
Prediction of Hit Songs

Project 15
Motivation

Now

So far, we have a predictive model with **good performance** in **classifying** hit/non-hit songs, based on **high level** and **low level** features.

Future

We need to Train and test our model using **more recent songs**, with a revision on the feature set to select out the **most influential features**.

Github Repository: [https://github.com/shahabty/hit_song_prediction](https://github.com/shahabty/hit_song_prediction)
Knowledge Graph Construction from Images

Project 16
Main Approach

Object Detection
Generate Graph from detected Objects
Predicting Magnetic Interactions of Molecules

Project 17
Scalar Couplings
Yelp Network Analytics and Recommendations

Project 18
Main Approach

- Data filtering
  - Limited to Toronto City
  - Focus on “Beauty and Spa” business category
  - Concentrate on user signed up after 01-01-2018
- Goal
  - Focus on the study on a city, a category and limit the number of users

<table>
<thead>
<tr>
<th>Data</th>
<th>Beauty &amp; Spa Business</th>
<th>Review</th>
<th>Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>1618</td>
<td>23620</td>
<td>172</td>
</tr>
</tbody>
</table>
Winter 2021
Analysis of Google Play Store Reviews

Project 1
Our Data Set

App reviews on the Google Play Store

547 apps
8 million reviews

Several languages

Network Topology Based on Comment Sentiment
Human Motion-State Classification from Motion Capture Data

Project 2
Animation & Motion Capture Databases

Current motion capture databases rely on manually annotated tags for querying. This approach has problems:

1) Manual labeling does not scale well
2) Is not always very accurate

Image Source: Gupta et al. (2016) [1]
User Emotional Modeling and Visualization in Social Media Conversations

Project 3
User emotional association to topics

Overall emotion distribution
LLM-based Recommendations on Yelp

Project 4
Yelp Dataset
Deep-learning based Music Recommendations

Project 5
Lyrics embedding

TF-IDF (only considers frequency of words)
\[ \text{dim} = 500 \]

Doc2Vec (considers the order and meaning of the words)
\[ \text{dim} = 100 \]
Knowledge Graph for API Method Recommendation

Project 6
What we propose

**Self-explainable and interpretable** API recommendation system using a **Knowledge Graph**
Movie Metadata Analysis to Guide Recommendations

Project 7
Approach Overview

Data

<table>
<thead>
<tr>
<th>userld</th>
<th>movielid</th>
<th>rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>156</td>
<td>1</td>
<td>5.0</td>
</tr>
<tr>
<td>156</td>
<td>2</td>
<td>5.0</td>
</tr>
<tr>
<td>156</td>
<td>4</td>
<td>3.0</td>
</tr>
</tbody>
</table>

Movie Embedding Matrix (NxK)

User Embedding Matrix (MxK)

Neural Collaborative Filtering

Insights

Visualizing Stories

Project 8
Motivation

Characters and their interactions

“Do authors use similar patterns?”
“Are patterns similar in a genre?”
“What pattern appear in real stories like biography and history books?”

Image is retrieved from Wikipedia.
(https://upload.wikimedia.org/wikipedia/commons/7/73/Paget_holmes.png)

Sherlock Holmes and Dr. John Watson By Sidney Paget

Image is retrieved from Wikipedia.
(https://upload.wikimedia.org/wikipedia/commons/7/73/Paget_holmes.png)
Flight Network During Covid-19

Project 9
Strong Connected Components (SCC)

Travel Restrictions

SCC

SCC1

SCC2
Product-Customer Networks – Final Presentation

Project 10
Motivation & Dataset Information

• General Information:
  – 1067371 Transactions
  – 41 Countries
  – 5943 Customers
  – 5699 Products

• Questions:
  – What are the different communities of products/clients?
  – What new products can be recommended to different clients?
  – What products may be of interest to certain regions?
Movie Casting Problems

Project 10
Graph Autoencoder on Actors Network

- Learning a joint representation of both local graph structure and node features
- MSE on link weight prediction task 0.005406 (for edge weights from 0. to 1)