Team Name: Team 4

Team Members:

- Jae Woong Lee
- Ward Huang
- Somya Menghani

Purpose of project / Abstract

In modern days location-based recommender systems are able to fit users' interests more accurately the old "trial-and-error" approach. This is possible through large-scale data mining and machine learning algorithms. In this project, we will focus on the recommendation algorithms to implement such a system.

The ultimate goal of this project is to recommend appropriate places to a user using the Google local reviews dataset. The recommendations to a user will be based on their current location and/or place of residence. To give better personalized recommendation to each user, we will consider a combination of algorithms like Matrix factorization and content based filtering.

We will start with an EDA phase to give us insights on how to proceed with data preprocessing. After data preprocessing, clustering and recommendation algorithms will be employed, and visualizations can be utilized to efficiently deliver results.

Currently the places dataset includes various places. For example, hospitals, parks, garage etc. For proper recommendation, we will use content-based similarities to cluster places based on their "genre" so that algorithms can predict and recommend without being "out-of-topic".

Dataset overview

	Places	**Reviews**	**Users**
Entries	3,114,353 (1.4 gb)	11,453,845 (276 mb)	3,747,937 (178 mb)
Features	● GPS Coords ● Address ● Hours	● Review ● Ratings ● Timestamp ● Categories	● Places Lived ● Occupation ● Current Place ● Education
Overview	Information about redundant location, such as address or area code may be removed.	Per-user ratings will be normalized. Review timestamps will be discretized to an hourly format.	EDA will need to be performed on this dataset to better understand each feature.

Exploratory Data Analysis

- Visualization
    - Plotting review’s place location + (price, rating or category)
    - Check distributions of each feature (need discretization for numerical features)

Data Preprocessing

- Places Dataset
    - Drop “Address” Field
    - Drop “closed” restaurants
    - GPS: Normalize and/or Discretize (for using them as features or clustering)
    - Merge / Drop / Discretize Phone
- Reviews Dataset
    - Review Time: Normalize and/or Discretize

Clustering (by GPS coordinates)

- Algorithms
    - K-mean
    - DB-scan
    - Mean-Shift

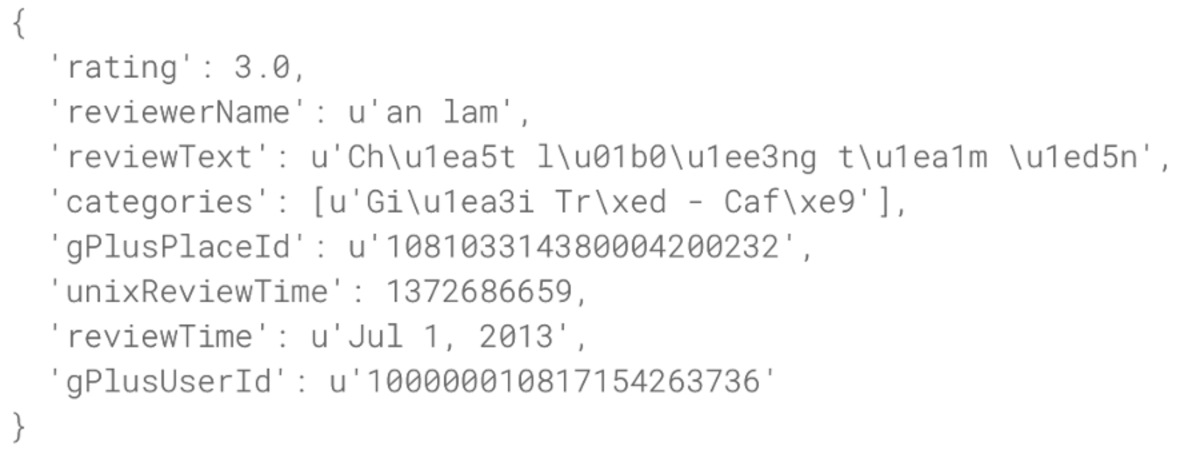
Recommender System

- Algorithms
    - Neighborhood Collaborative Filtering
        - User-Based
        - Item-Based
    - Content-Based Filtering
    - Matrix Factorization
- Possible input parameters
    - GPS
    - Reviews
- Possible output parameters
    - PlaceID(s)
    - Rating for PlaceID(s)

Possible Tools

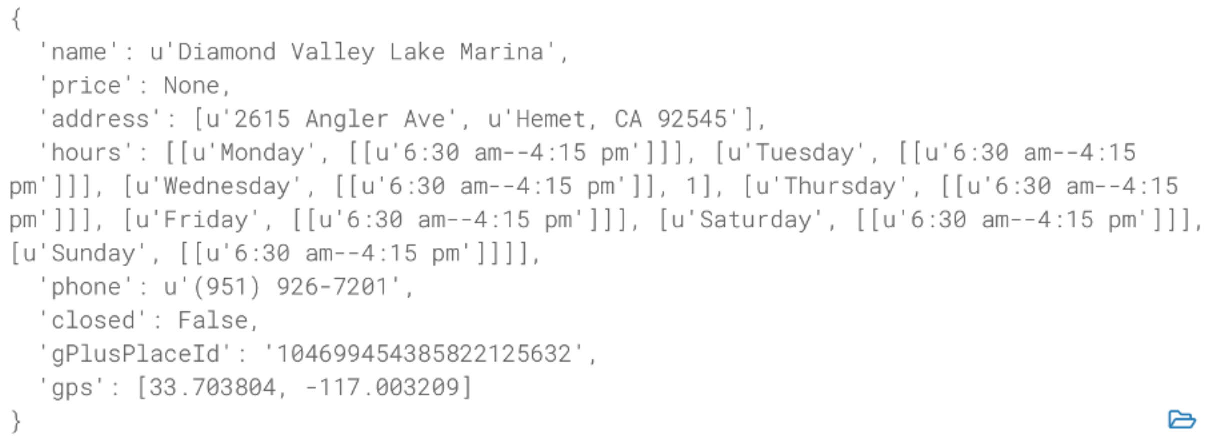
- Scikit-learn, Facets, pyplot, seaborn, ggplot, missingno, Surprise, Universal Recommender, Racoon

Example (review)



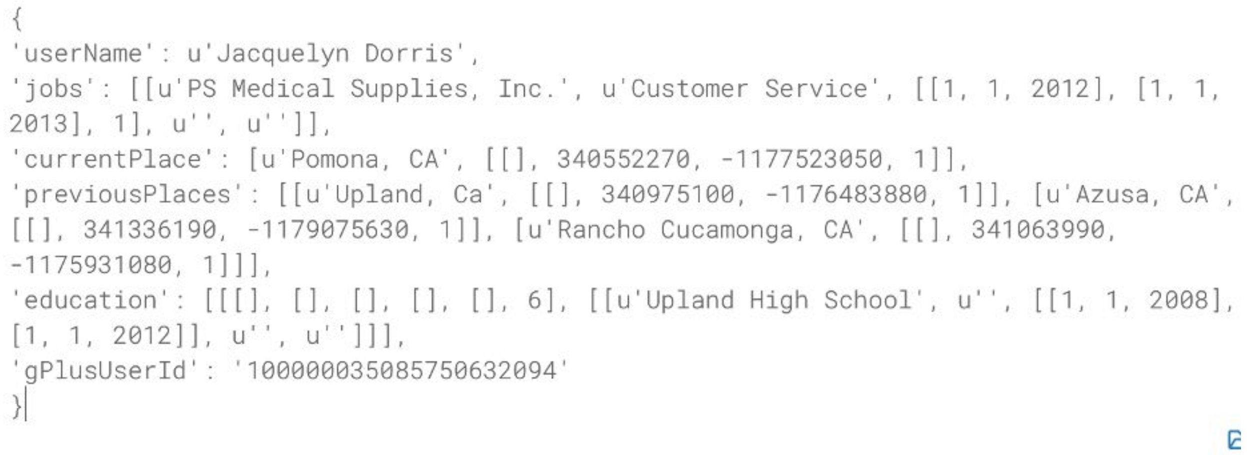
```
{
  'rating': 3.0,
  'reviewerName': u'an lam',
  'reviewText': u'Ch\u1ea5t l\u01b0\u1ee3ng t\u1ea1m \u1ed5n',
  'categories': [u'Gi\u1ea3i Tr\xed - Caf\xe9'],
  'gPlusPlaceId': u'108103314380004200232',
  'unixReviewTime': 1372686659,
  'reviewTime': u'Jul 1, 2013',
  'gPlusUserId': u'100000010817154263736'
}
```

Example (business)



```
{
  'name': u'Diamond Valley Lake Marina',
  'price': None,
  'address': [u'2615 Angler Ave', u'Hemet, CA 92545'],
  'hours': [[u'Monday', [[u'6:30 am--4:15 pm']]], [u'Tuesday', [[u'6:30 am--4:15
pm']]], [u'Wednesday', [[u'6:30 am--4:15 pm']], 1], [u'Thursday', [[u'6:30 am--4:15
pm']]], [u'Friday', [[u'6:30 am--4:15 pm']]], [u'Saturday', [[u'6:30 am--4:15 pm']]],
[u'Sunday', [[u'6:30 am--4:15 pm']]]],
  'phone': u'(951) 926-7201',
  'closed': False,
  'gPlusPlaceId': '104699454385822125632',
  'gps': [33.703804, -117.003209]
}
```

Example (User)



```
{
'userName': u'Jacquelyn Dorris',
'jobs': [[u'PS Medical Supplies, Inc.', u'Customer Service', [[1, 1, 2012], [1, 1,
2013], 1], u'', u'']],
'currentPlace': [u'Pomona, CA', [[], 340552270, -1177523050, 1]],
'previousPlaces': [[u'Upland, Ca', [[], 340975100, -1176483880, 1]], [u'Azusa, CA',
[[], 341336190, -1179075630, 1]], [u'Rancho Cucamonga, CA', [[], 341063990,
-1175931080, 1]]],
'education': [[[], [], [], [], [], 6], [[u'Upland High School', u'', [[1, 1, 2008],
[1, 1, 2012]], u'', u'']]],
'gPlusUserId': '100000035085750632094'
}
```