Data Set and Features

Data
- Scrape a portion of the user ids (randomly) and their starred dishes from Xia Chu Fang, a community where users can publicly post recipes of dishes, and save dishes they are interested in
- Utilize parallel crawling and proxies to fetch data more efficiently, and ran a spider on Google Cloud
- Scraped ~230,000 valid users and 2.5 million dishes in their starred list (~12,000 unique dish names)

Ratings
- User ratings are necessary to implement methods such as collaborative filtering
- Rating is defined as $R_{ui}^{(0)}$ = count of dish u in user i's starred list
- Ratings from 5 to 10 is kept for simplification

Dish Name Mapping
- Map dishes to an online database of Chinese dish names with English translations as a dictionary

Objective
- Obtain user and dish data from the web, and provide English translation
- Build a recommendation system for dishes through rating predictions
- Examine the prediction performance of our selected algorithms
- Focus on Chinese food for specialization

Methods

Word2Vec - Skip Gram
- Trains a neural network with a single hidden layer to perform, and outputs words most relevant to the input
- Minimizes the loss function $E$ in each training iteration:

$$E = - \log p(w_{u1}, w_{u2}, ..., w_{u|w|})$$

$$= - \log \prod_{w \in W} \frac{\exp(u_{w}^{T} v_{w})}{\sum_{w' \in W} \exp(u_{w'}^{T} v_{w'})}$$

$$= - \sum_{w \in W} u_{w}^{T} v_{w} + C \cdot \log \sum_{w \in W} \exp(u_{w}^{T} v_{w})$$

- Returns a similarity score for each output word

Collaborative Filtering
- Matrix-factorization (MF)-based approaches prove to be highly accurate and scalable in addressing CF problems
- Implements non-negative matrix factorization (NMF) and singular value decomposition (SVD) for comparison
- NMF
  - Utilizes Python library ‘Surprise’
  - Uses regularized stochastic gradient descent update rule
  - Uses $\lambda = 0.06$ and $\lambda = 0.06$

- SVD
  - Minimizes by gradient descent
  - Uses learning rate $\gamma = 0.005$ and regularization factor $\lambda = 0.02$

Results

Skip-Gram
- Creates t-SNE graphs to represent similarity between dish names
- The closer the dishes are, the more similar they are

NMF Predictions

SVD Predictions

Error Analysis
- RMSE
  - $RMSE = \sqrt{\frac{1}{N} \sum_{u \in U} (\hat{R}_{u} - R_{u})^2}$
- Recall
  - $\frac{TP}{TP + FN}$

Conclusions
- Word2Vec directly gives recommendations, but it is hard to conceptualize or quantify errors
- SVD model performs the best for CF as it has the lowest RMSE and highest Recall on dev set, and the test set error is close to the dev set error, which means it does not overfit and is fairly robust

Future Work
- try other recommendation systems (hybrid system, item-based CF, memory based algorithm, ...)
- obtain data from other dish websites to examine stability
- create user interface
- References available upon request