NINKASI: Beer Recommender

Alex Li, Andrew Wu,
Nelson Chen, Luke Chu

Jan 17, 2016

Roadmap

- Introduction
- Data Scraping
- EDA
- Recommender System
- Python Flask App
- Future Improvements

Introduction

NK

Motivation

- Ever get tired of drinking common beers like Budweiser or Corona? Move into craft beers?
- Want to find a similar beer?
- Want to find a different beer without taste testing too many before finding a good one?
- Let us recommend you some!

Competitor Research

- Only found systems to find similar beers
- Most sophisticated system, *Next Glass*, uses mass spectrometry to classify beers
- No recommenders that finds similar users

Our Goal:

Build hybrid beer recommendation system that finds both similar ***items*** and ***users***

Project Workflow

Web Scraping

Data Cleaning

Content-Based Recommender

Collaborative-Filtering Recommender

Flask App

NINKASI BEER APP

Web Scraping

NK

Data Scraping

Home > Breweries > United States: Illinois > Goose Island Beer Company (AB-InBev)

Goose Island Bourbon County Stout

* picture credits
copyright may apply

overall **100**

100 style

Brewed by Goose Island Beer Company (AB-InBev)
Style: Imperial Stout - Top 50
Chicago, Illinois USA
Serve in Snifter

bottled common | on tap common | Broad Distribution

send corrections | shelftag | edit barcodes | update pic

RATINGS: 2809 WEIGHTED AVG: 4.26/5 IBU: 60 EST. CALORIES: 426 ABV: 14.2%

COMMERCIAL DESCRIPTION

"I really wanted to do something special for our 1000th batch at the original brewpub. Goose Island could have thrown a party. But we did something better. We brewed a beer. A really big batch of stout-so big the malt was coming out of the top of the mash tun. After fermentation we brought in some bourbon barrels to age the stout. One hundred and fifty days later, Bourbon County Stout was born-A liquid as dark and dense as a black hole with a thick foam the color of bourbon barrels. The nose is a mix of charred oak, vanilla,carmel and smoke. One sip has more flavor than your average case of beer. It overpowers anything in the room. People have even said that it's a great cigar beer, but I haven't yet tried a cigar that would stand up to it." Brewmaster Greg Hall;
IBU's 60-High Color - Midnight
Was 11% abv,
2007 and 2008 - 13% abv
2011 - 14.5% abv
2012 - 15% abv
2013 - 14.9% abv
2014 - 14.4% abv
2015 - 14.2% abv

Editor's Note: Baudoinia Fulton & Wood Series offering (And the Low Storage entry from FoBAB 2012) is simply Bourbon County Stout aged in the same barrels. It has been alted as it offers no different recipe or barrel type simply the presence of a distillers fungus on the barrels that adds no distinct characteristic to the beer from the fungus itself except for possible oxygen exposure changes. While Brewers Intent indicates that they consider it a new beer, no true distinction aside from this oxidation amount and possible aging time differences separates the beers. It's essentially a single barrel Bourbon County version something we have always treated as regular Bourbon County Stout in the past.

Most Recent | Top Raters | Highest Score | Rated By | Ticked By

4 AROMA 9/10 APPEARANCE 3/5 TASTE 8/10 PALATE 4/5 OVERALL 16/20
Sudz4Dayz (70) - Montreal, Quebec, CANADA - DEC 16, 2016

2016 bottle purchased at Fort Point Market in South Boston, MA. Pours jet black, smooth, cappuccino head with small frothy bubbles. Dissipates very quickly. Smells like burnt wood, vanilla, roastiness with bourbon sweetness and booze as well. Love it. Nice roasty flavor, smooth vanilla. Those come first and are quickly replaced by ample, everlasting dark chocolate flavor. Sweet bourbon barrel throughout and slight charred wood. Not the thickest mouthfeel but I wouldn't say light either. Medium. Overall not all too complex, balance is ok. The flavor is great and it definitely saves this one from being mediocre. In terms of overall quality it's definitely missing balance, a little complexity and a little on the mouthfeel. Delicious nonetheless. I enjoyed this one.

4.1 AROMA 8/10 APPEARANCE 3/5 TASTE 9/10 PALATE 4/5 OVERALL 17/20
chinchill (4448) - South Carolina, USA - DEC 15, 2016

12 oz bottle dated 29 August 2014 served in a Belgian snifter. Removing the cap indicated a secure seal, but this has very little carbonation. This is not ideal for the appearance, but the low carbonation seems about right for the smooth, full, high ABV body. Boozy and woody aroma with dark roasted grains and molasses. Sweet bourbon dominates the rich flavor. Moderately woody with faint vanilla and coffee in the finish. Mild bourbon barrels and dark bread in the aftertaste.

4.7 AROMA 9/10 APPEARANCE 5/5 TASTE 9/10 PALATE 5/5 OVERALL 19/20
Korcz (1142) - Warsaw, POLAND - DEC 11, 2016

Backlog, ocena przepisana z untappd, w ramach uzupełniania profilu na ratebeer. Genialne. Perfekcyjnie gładkie, oleiste, kremowe, eleganckie. Piękna wanilia, szlachetny alkohol, potężne nuty Bourbonu, odbeczkowy kokos i genialna jak na tę moc pijalność. Wybitne piwo, jeden z moich ulubionych stoutów ever. Jeśli nie ulubiony :)

4.2 AROMA 9/10 APPEARANCE 5/5 TASTE 8/10 PALATE 5/5 OVERALL 15/20
Listigovers (967) - Toronto, Ontario, CANADA - DEC 10, 2016

Pitch black, very strong whisky scent. Taste is very sweet and there is also heat from the 14%, tar like texture, very roasted and bourbony flavour. A true sipper, incredibly heavy and warming beer.

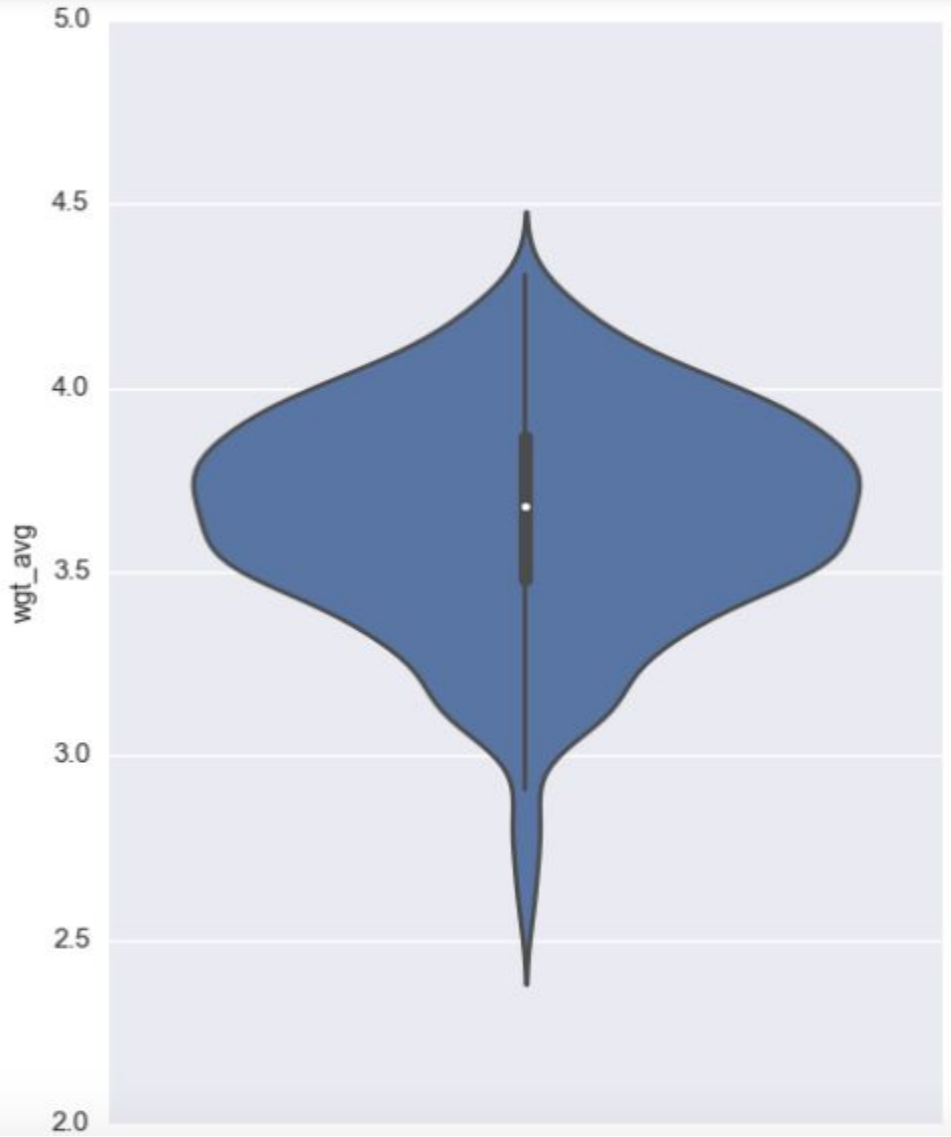
Data Set

- 1275 unique beers
- Top 25 beers per state
- 16,000+ users
- 280,000+ reviews/ratings

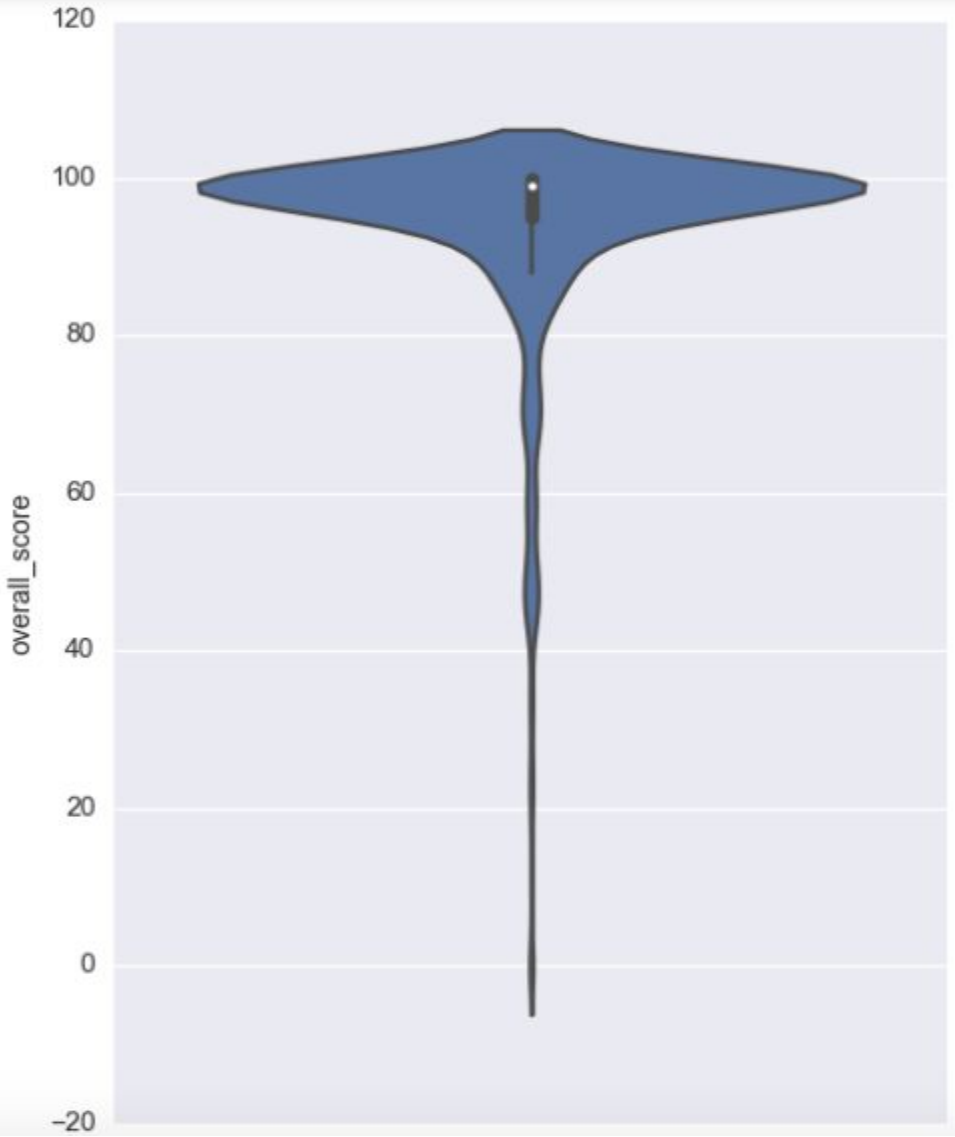
Exploratory Data Analysis

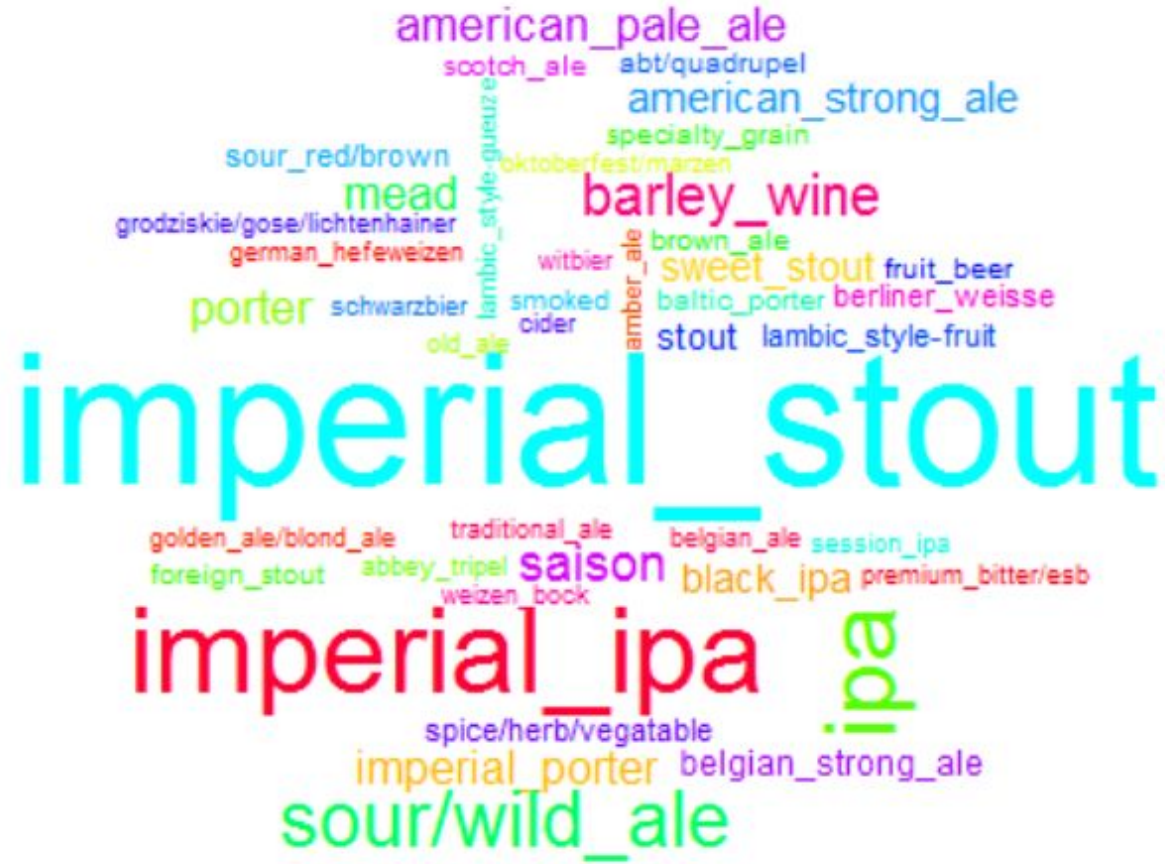
NK

Weighted Avg



Overall Score





	count
Imperial Stout	352
Imperial IPA	146
India Pale Ale (IPA)	118
Sour/Wild Ale	85
Barley Wine	54
Saison	45
Mead	42
American Pale Ale	42
Porter	40
Imperial Porter	35
American Strong Ale	30
Sweet Stout	26
Black IPA	26
Stout	24
Belgian Strong Ale	16
Sour Red/Brown	14
Spice/Herb/Vegetable	13
Lambic Style - Fruit	13
Berliner Weisse	12
Baltic Porter	11

Latent Dirichlet Allocation

Top Two Topics:

Stout

Word	Proportion
chocolate	0.012
coffee	0.010
bourbon	0.009
black	0.006
roasted	0.006
roast	0.004
brown	0.004
vanilla	0.004
dark	0.003
stout	0.003

Fruity IPA

Word	Proportion
pine	.006
citrus	.006
grapefruit	.006
orange	.005
ipa	.004
tropical	.004
hop	.004
golden	.004
funk	.003
mango	.003

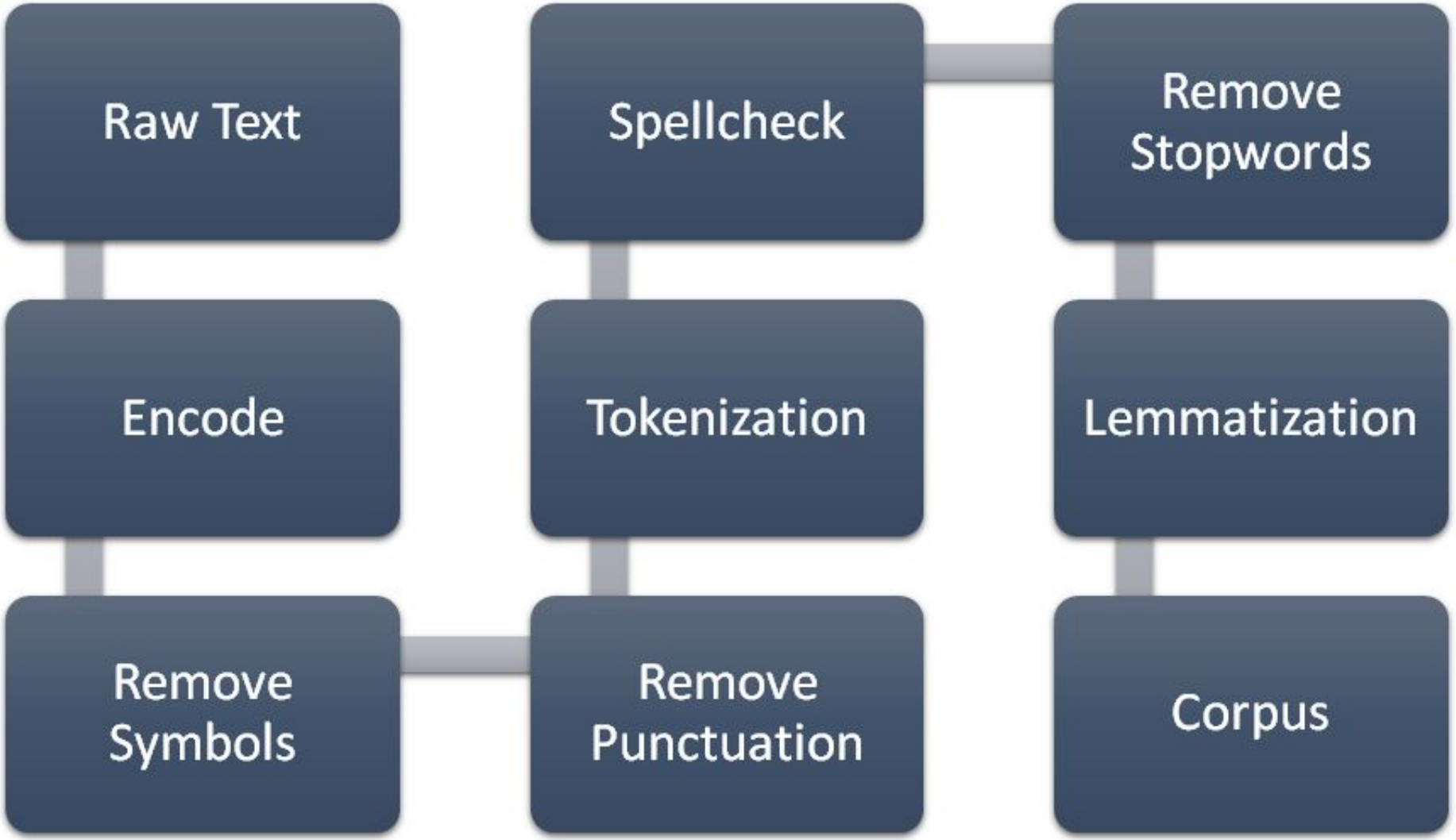
Recommender System

NK

Recommender System

Reviews	Ratings
Text Processing	SVD++, RBM
TF-IDF, LSI	Ensemble

Content-Based: NLP



Content-Based: Algorithms

- Term Frequency-Inverse Document Frequency (TF-IDF): calculates importance of terms in each document
- Latent Semantic Indexing (LSI): does dimension reduction on the document-term matrix

Collaborative Filtering Algorithms

- Recommendation based on beer rating (explicit information)
- Two separate models were used
  - Singular Value Decomposition++ (SVD++)
  - Restricted Boltzmann Machine (RBM)
- Ensemble models together (Average)

Collaborative Filtering Results

- SVD++ : 91.4% Accuracy
- RBM : 90.7% Accuracy
- Ensemble: 92.1% Accuracy

Prediction Method for New Users

- Neighborhood approach using cosine similarity
- Impute missing ratings with weighted average
- Rank and output top 5 as recommendations

Flask App Demonstration

NK

ninkasibeer.herokuapp.com

Future Steps

NK

Future Steps

- Expand beer database
- Add in more features (i.e style, palate, appearance, aroma, taste, and time)
- Build better single models (hyperparameter tuning, try new algorithms)
- Ensemble smarter (use minimizer or stacking)
- Develop App functionalities further

Thank you, Cheers!

Scrapy

python™
Natural Language Analyses
with NLTK

gensim
topic modelling for humans

TensorFlow

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

matplotlib

APACHE Spark™

NumPy

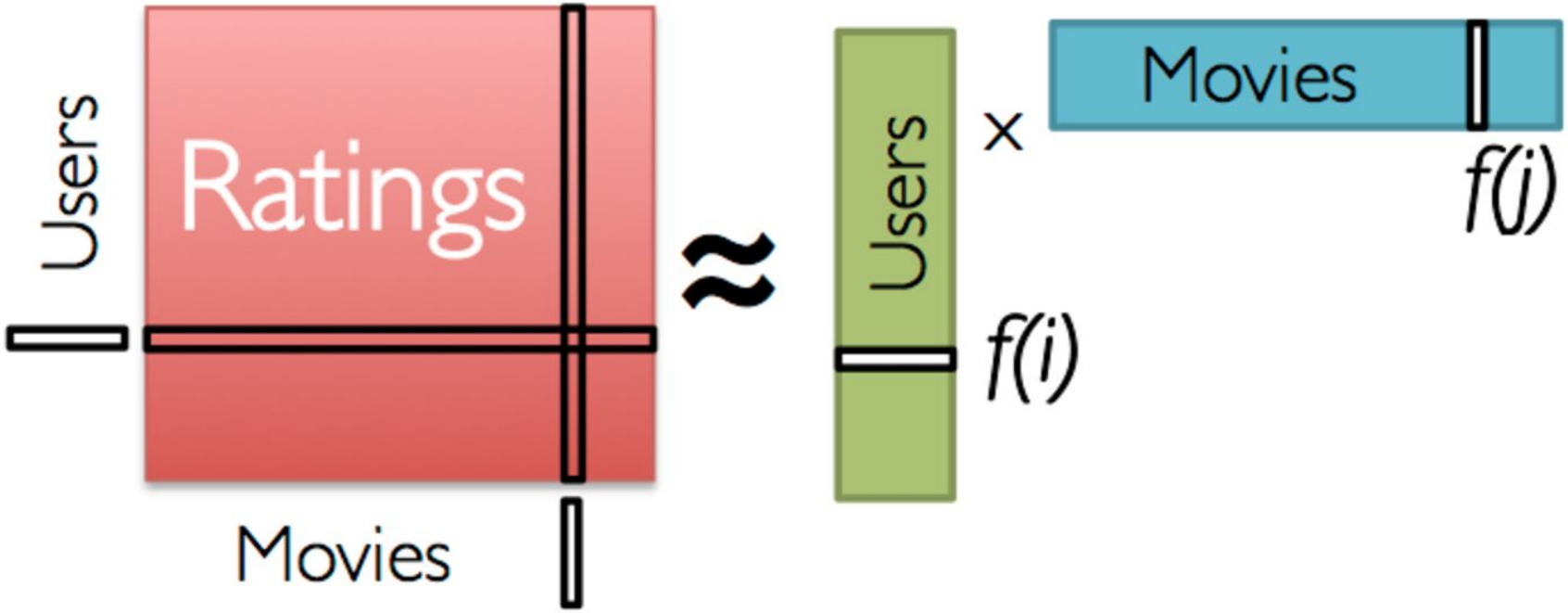
Flask

Appendix

NK

SVD For Collaborative Filtering

- Latent Matrix Factorization of user-item matrix to find *latent* features
- Used Spark to implement as baseline: RMSE = 2.15



https://databricks-training.s3.amazonaws.com/movie-recommendation-with-mllib.html

SVD++ (Implicit Feedback Version)

- Developed by Netflix Challenge winners
- Decomposes rating equation further
- Users that have less ratings are penalized more

$$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T \left(p_u + \frac{1}{\sqrt{|N(u)|}} \sum_{j \in N(u)} y_j\right)$$

Predicted rating

Global average

User bias

Item bias

Item Latent feature

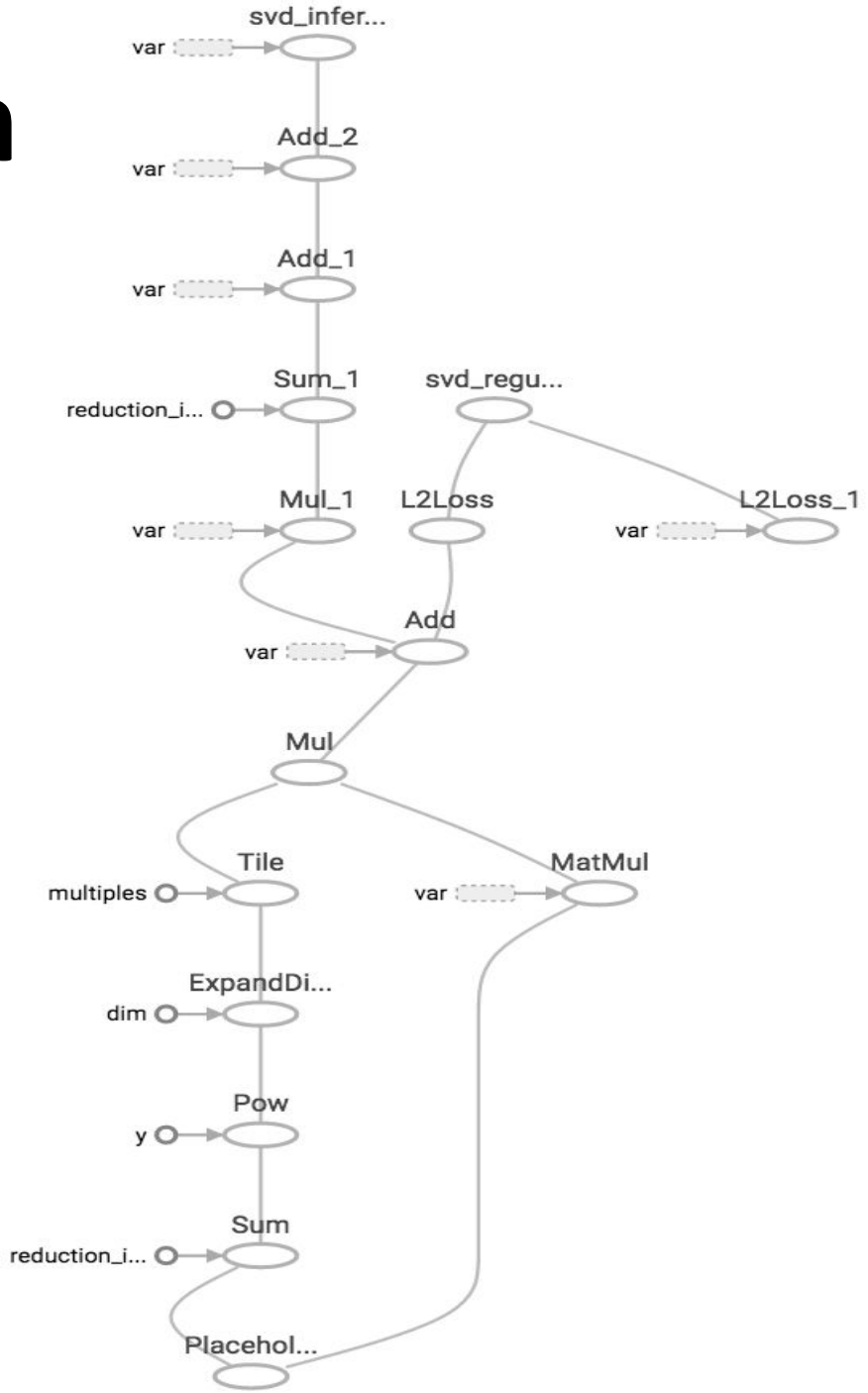
User latent feature

Implicit feedback

Implicit feedback parameters

SVD++ Implementation

- No fast publicly available Python package
- Realized SVD++ can be highly optimized using ***TensorFlow***
- Accelerated by C++ backend, GPU computation
- **Results: RMSE = 1.69**

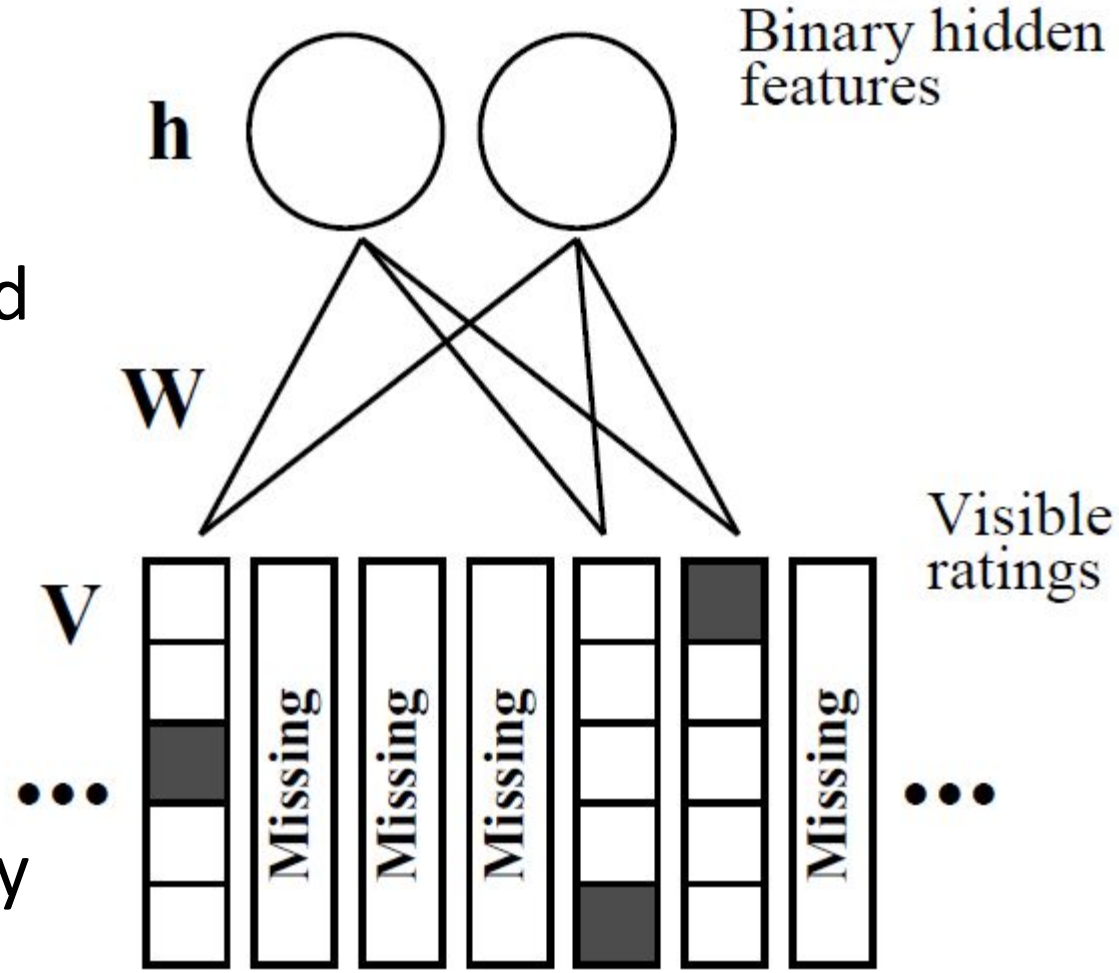


Restricted Boltzmann Machine (RBM)

- RBM is a generative, two-layer neural network
- Inspired by the Boltzmann Distribution from thermodynamics and fluid dynamics to minimize the energy function
- Performs CF by reconstructing the user-item matrix
- Hidden units can be thought of as latent features

Restricted Boltzmann Machine

- To train a RBM:
  - Initialize the visible layer
  - Hidden layer is calculated
  - Missing values are imputed
  - Weights are updated based with gradient descent
- The RBM is trained for every user, but the weights and bias are shared across all users



Ensemble CF models

- Averaged SVD++ and RBM predictions to get final predictions

Term Frequency-Inverse Document Frequency (TF-IDF)

- Calculate the “importance” of every word to a review in a corpus.
- Term Frequency is the number of times a word occurs in a document.

$$\mathrm{tf}(t,d) = 0.5 + 0.5 \cdot \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}}$$

- Inverse Document Frequency measures how rare is the word across all documents.

$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

- The TF-IDF is defined as

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D)$$

Latent Semantic Analysis (LSA)

- LSA performs dimension reduction on the TF-IDF matrix

$$
\begin{array}{cccc}
 & X & U & \Sigma \\
 & (\mathbf{d}_j) & & \\
 & \downarrow & & \\
(\mathbf{t}_i^T) \rightarrow & \begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & x_{m,n} \end{bmatrix} & = & (\hat{\mathbf{t}}_i^T) \rightarrow \begin{bmatrix} \begin{bmatrix} \\ \mathbf{u}_1 \\ \\ \end{bmatrix} & \cdots & \begin{bmatrix} \\ \mathbf{u}_l \\ \\ \end{bmatrix} \end{bmatrix} \cdot \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_l \end{bmatrix}
\end{array}
\cdot \begin{array}{c} V^T \\ (\hat{\mathbf{d}}_j) \\ \downarrow \\ \begin{bmatrix} [& \mathbf{v}_1 &] \\ & \vdots & \\ [& \mathbf{v}_l &] \end{bmatrix} \end{array}
$$