NINKASI: Beer Recommender

Alex Li, Andrew Wu,
Nelson Chen, Luke Chu

Dec 18, 2016

Outline

- Introduction
- Data scraping
- Exploratory data analysis
- Recommender system
  - Content-based method
  - Collaborative filtering method
- Python Flask App
- Future improvements

Introduction

NK

Introduction

- Build a recommender system for beer lovers!
- Recommender can find similar beers based on reviews
- Recommender can also give suggestions based on other user explicit ratings
- Models tied together in a Flask App

Project Workflow

Data Scraping

Data Preprocessing & EDA

- NLP Processing
- EDA

Recommender System Algorithm

- Content-Based
- Collaborative Filter

Flask App

Web Scraping & Data Cleaning

NK

Data Scraping

- Website: www.ratebeer.com
- About 280,000 reviews
- Limited data scope to top 20-25 beers per state

Scrapy Fields

Beer Information	Beer Review Information
Beer Name	User Name
Brewer Name	User Location
Weighted Average	Time of Review
Beer Image	User Rating
State Beer Produced	Aroma
Overall Score	Appearance
Beer Style	Taste
Alcohol by Volume	Palate
Estimated Calorie	Overall
IBU (Int'l Bitter Unit)	Review
Beer Description	

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

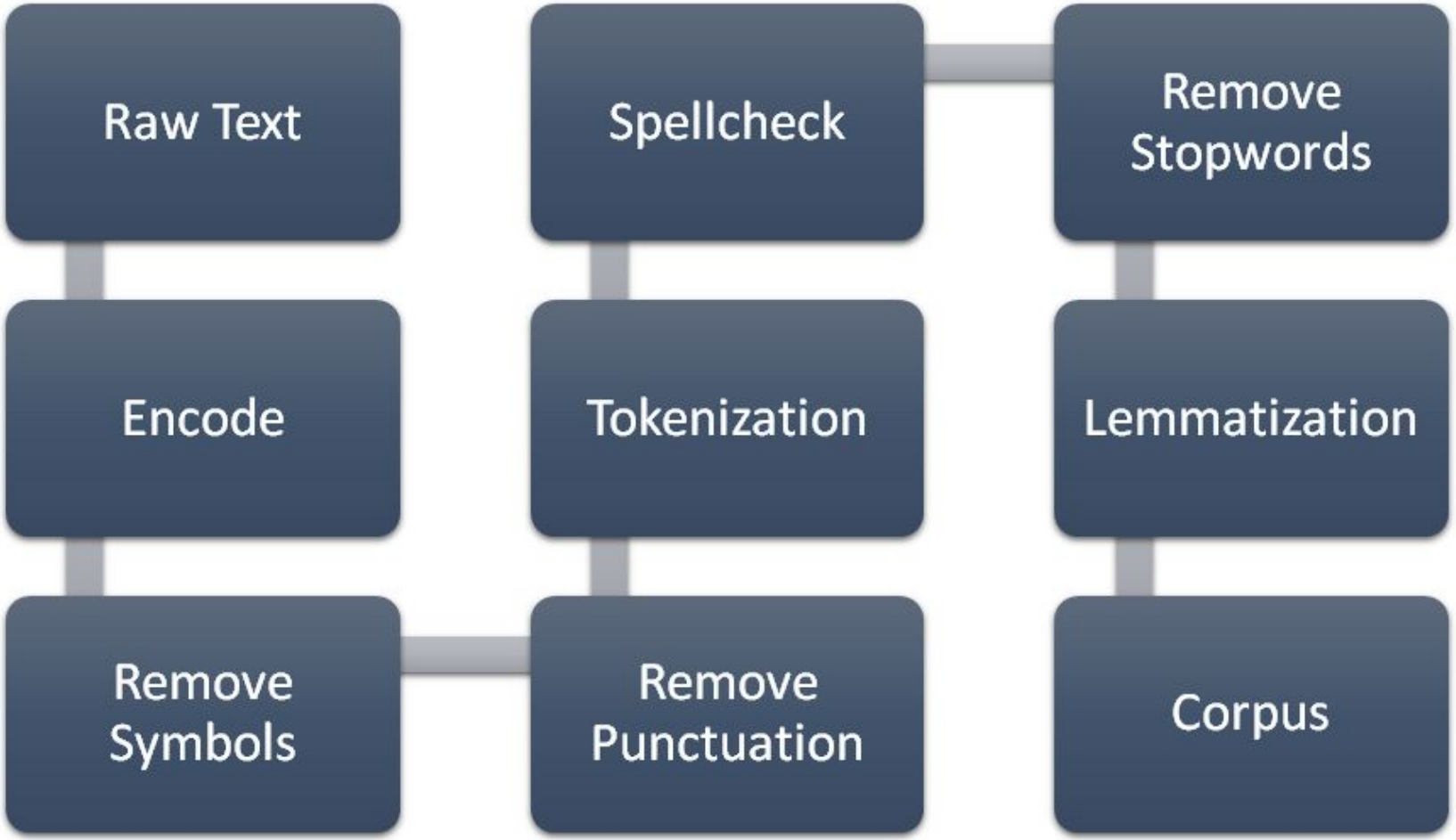
Backlog, ocena przepisana z untappd, w ramach uzupełniania profilu na ratebeer. Genialne. Perfekcyjnie gładkie, oleiste, kremowe, eleganckie. Piękna wanilia, szlachetny alkohol, potężne nuty Bourbonu, odbeczkowy kokos i genialna jak na tę moc pijalność. Wybitne piwo, jeden z moich ulubionych stoutów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



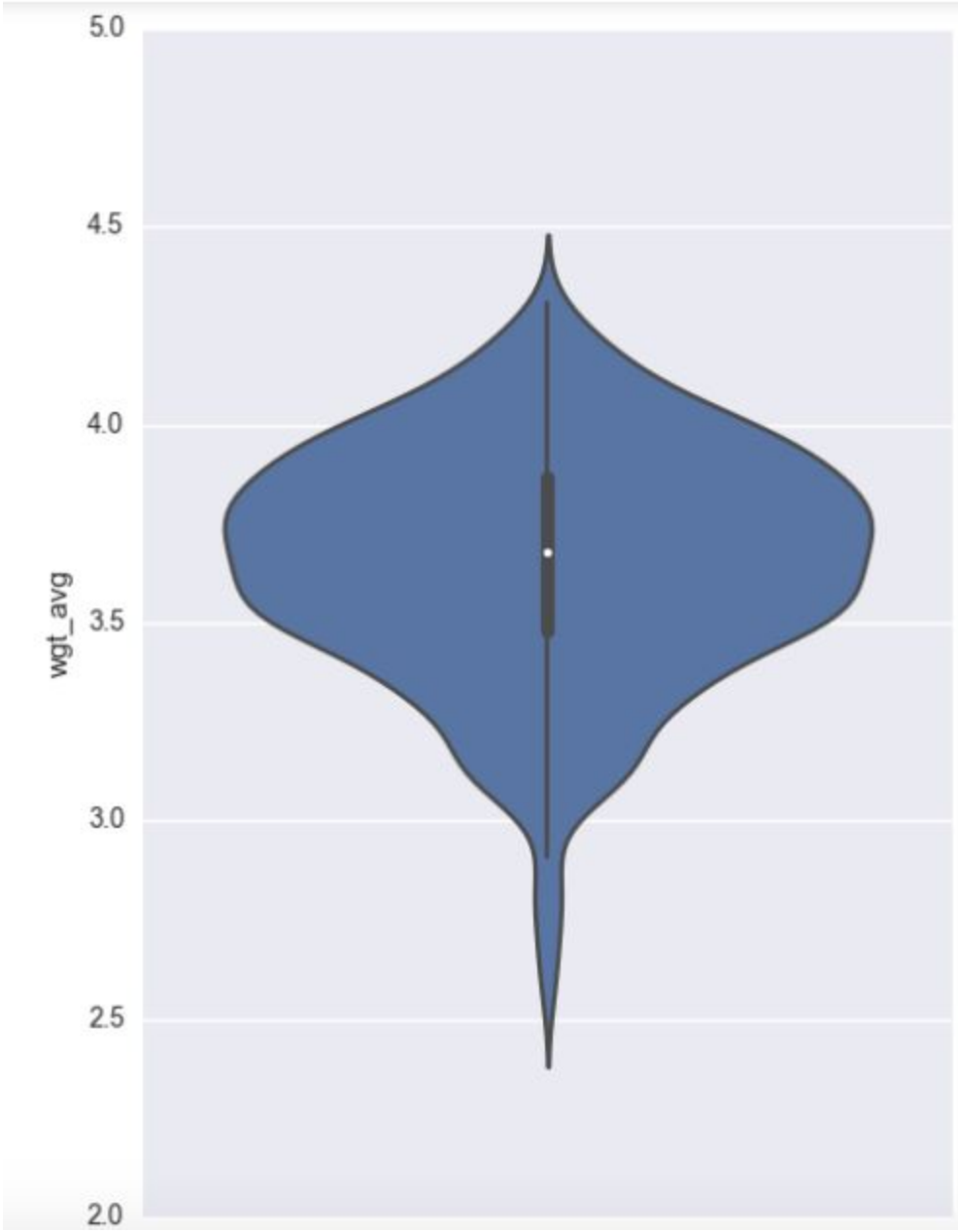
Exploratory Data Analysis

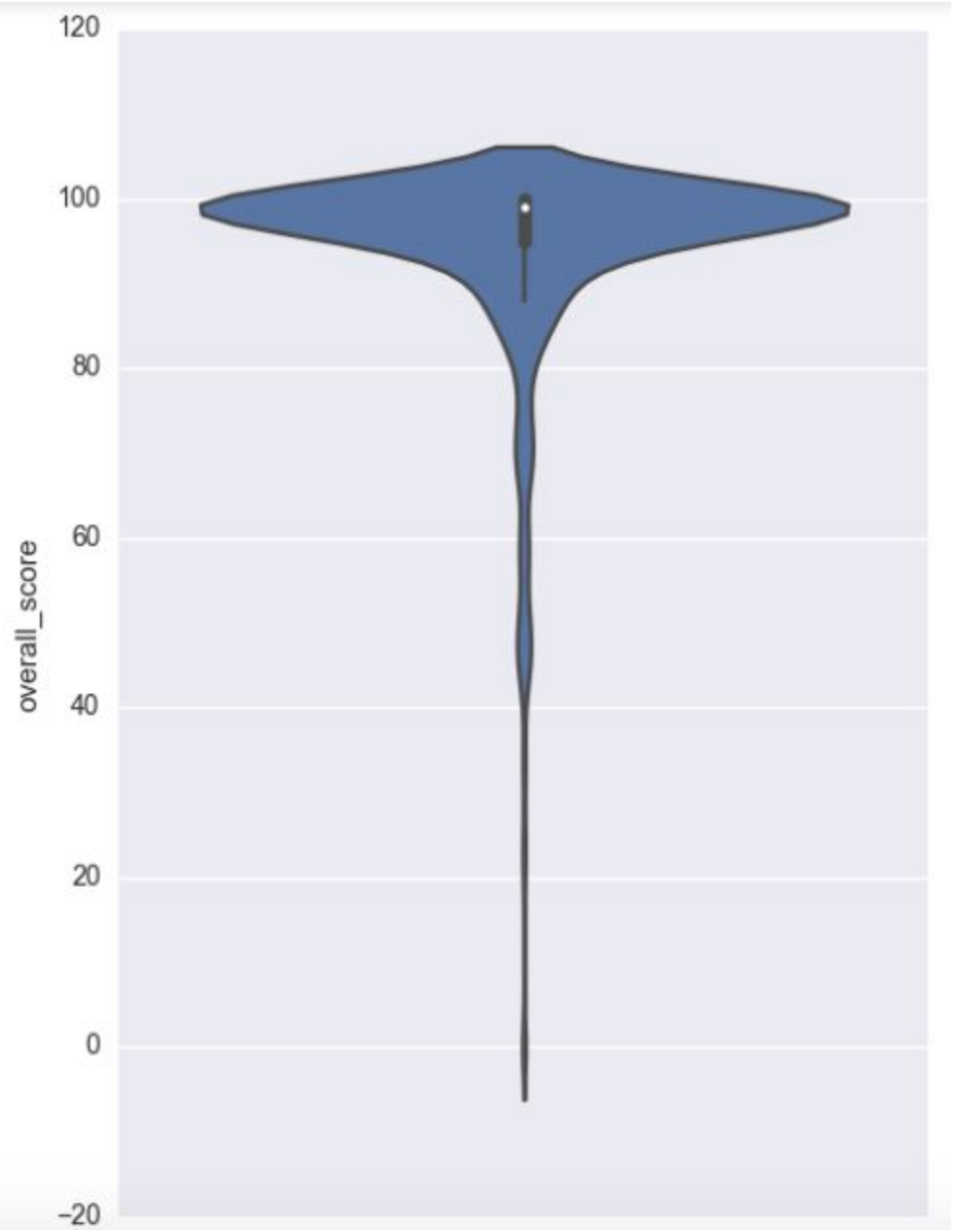
NK

EDA: Beer Items

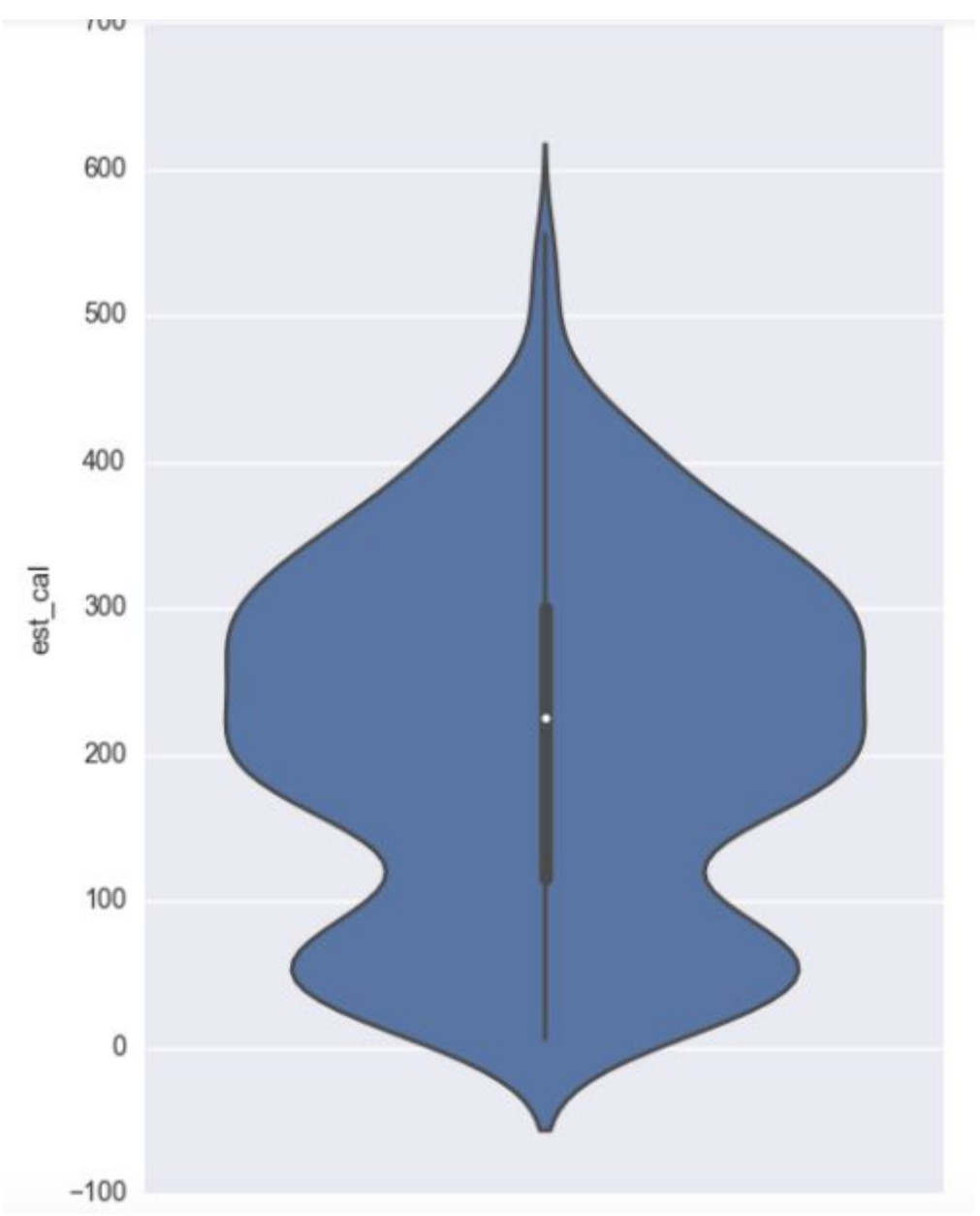
- ~1200 beers
- 58 styles, 335 unique brewers
- ~270,000 reviews

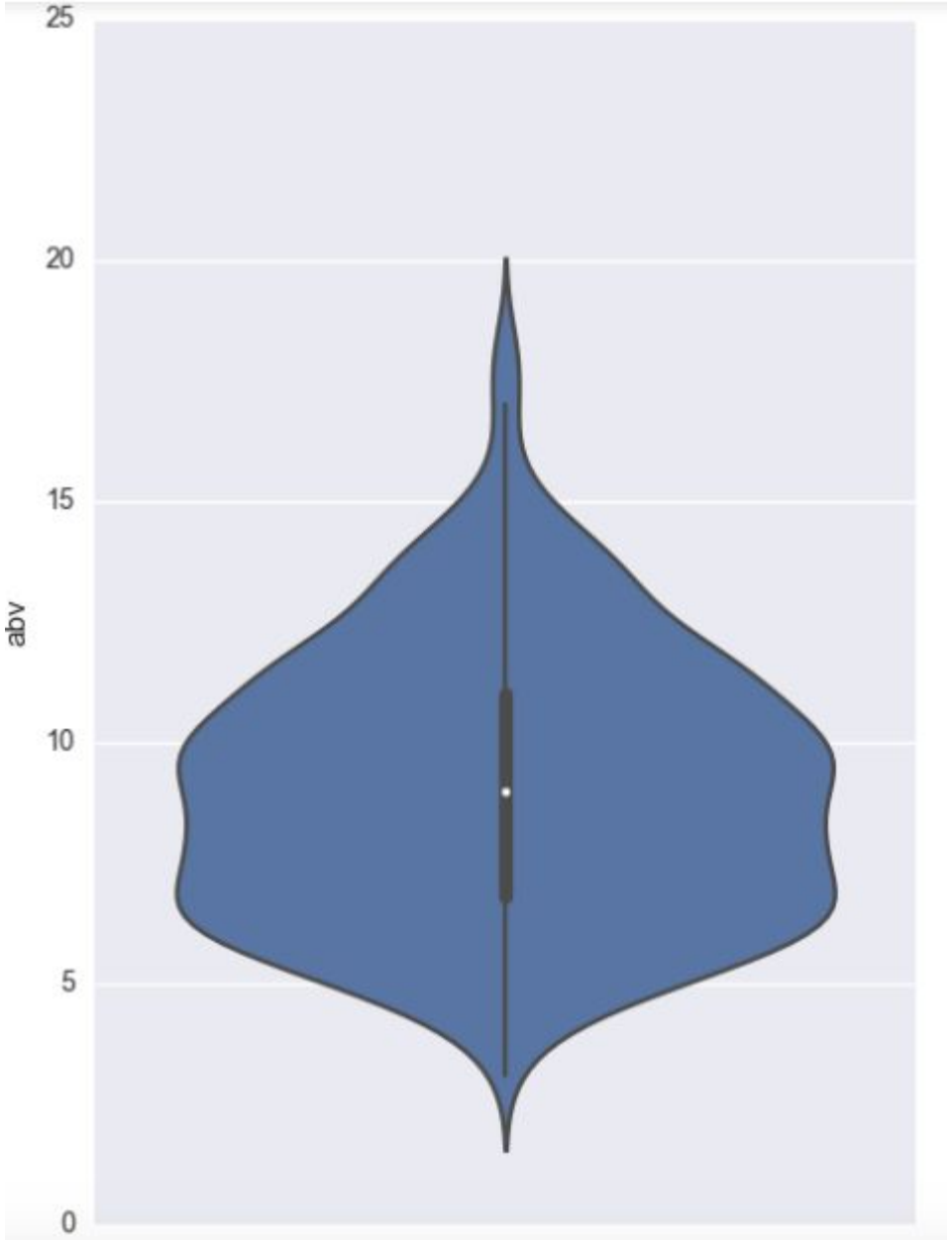

NK


wgt_avg
5.0
4.5
4.0
3.5
3.0
2.5
2.0

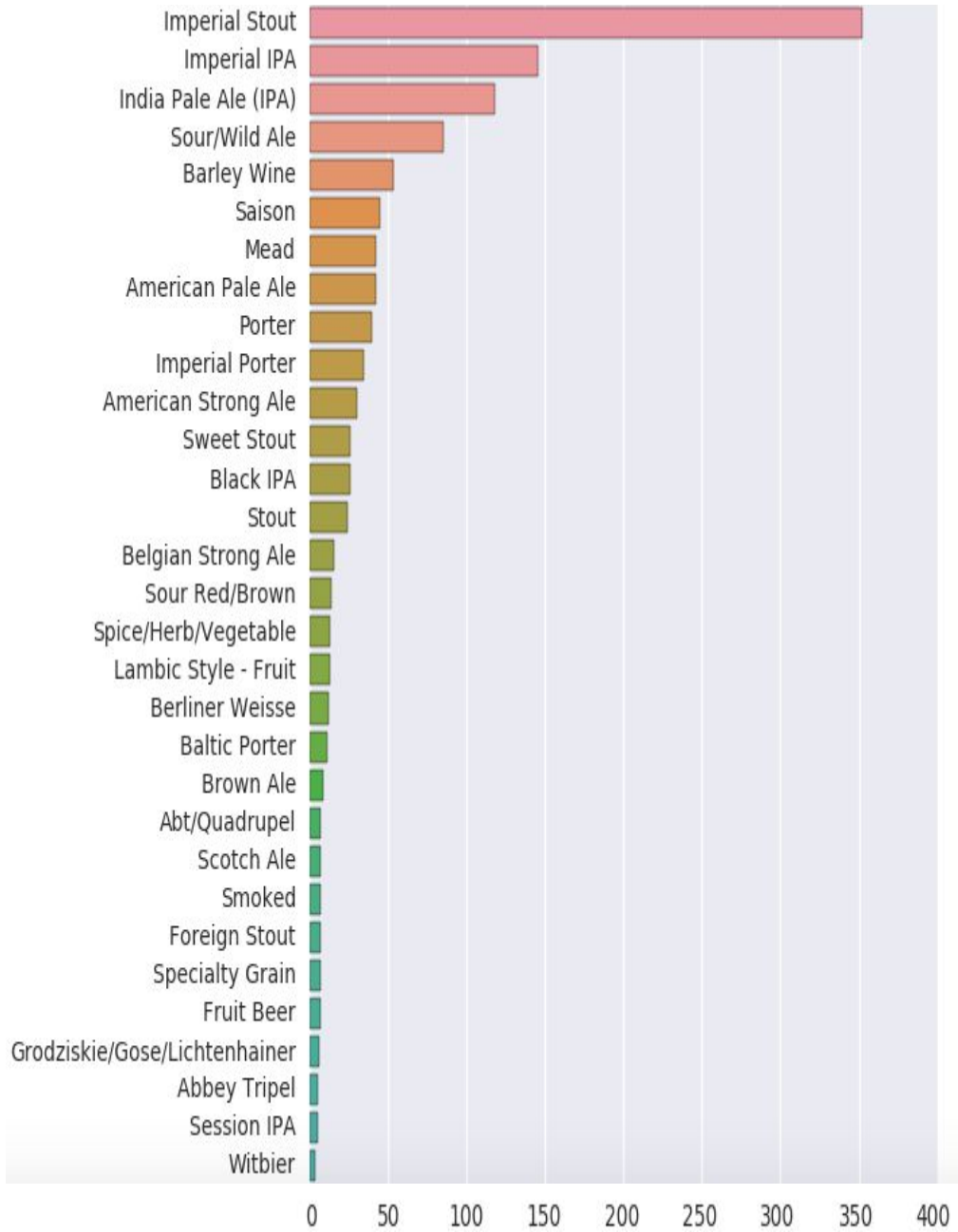

overall_score
120
100
80
60
40
20
0
−20

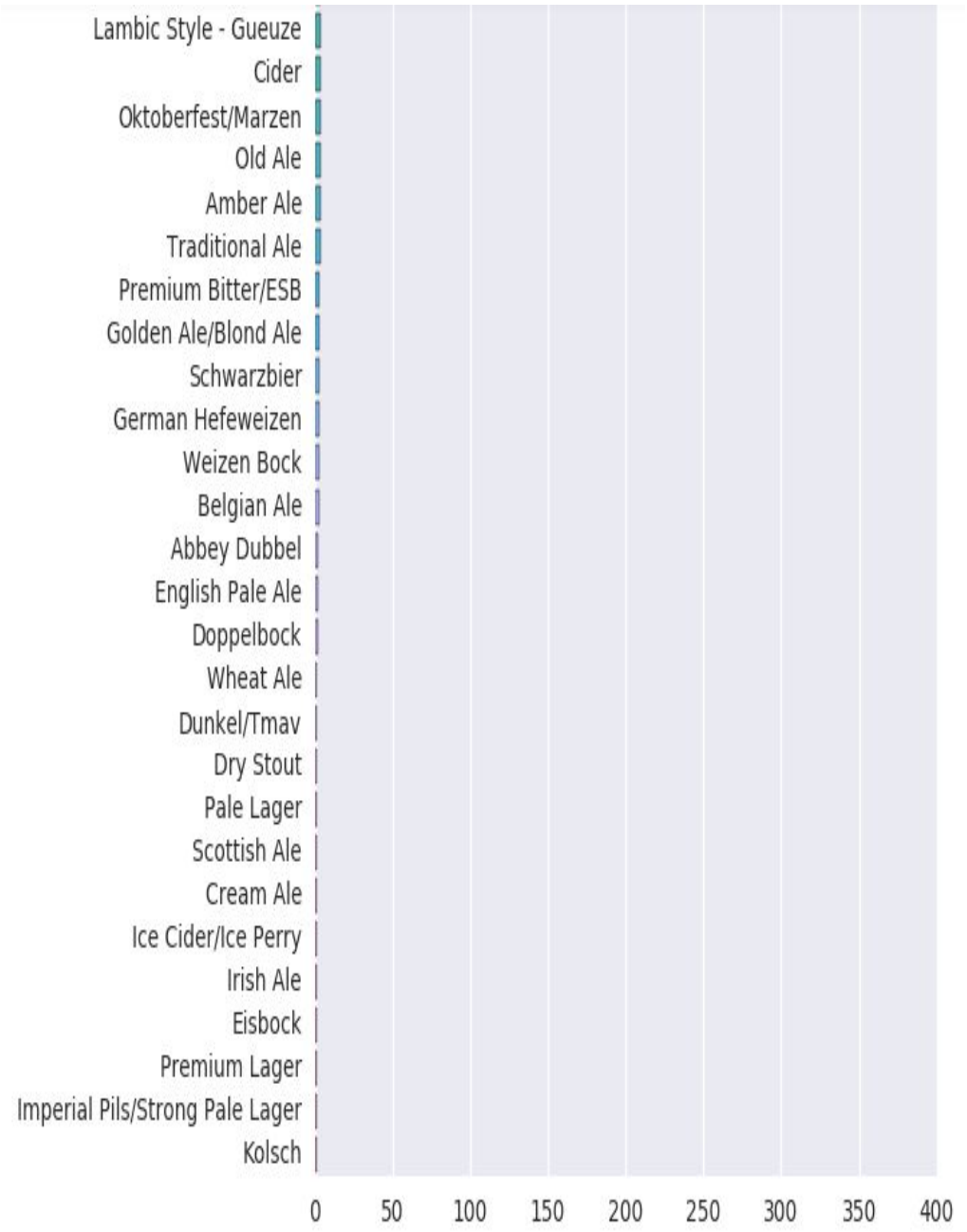

NK

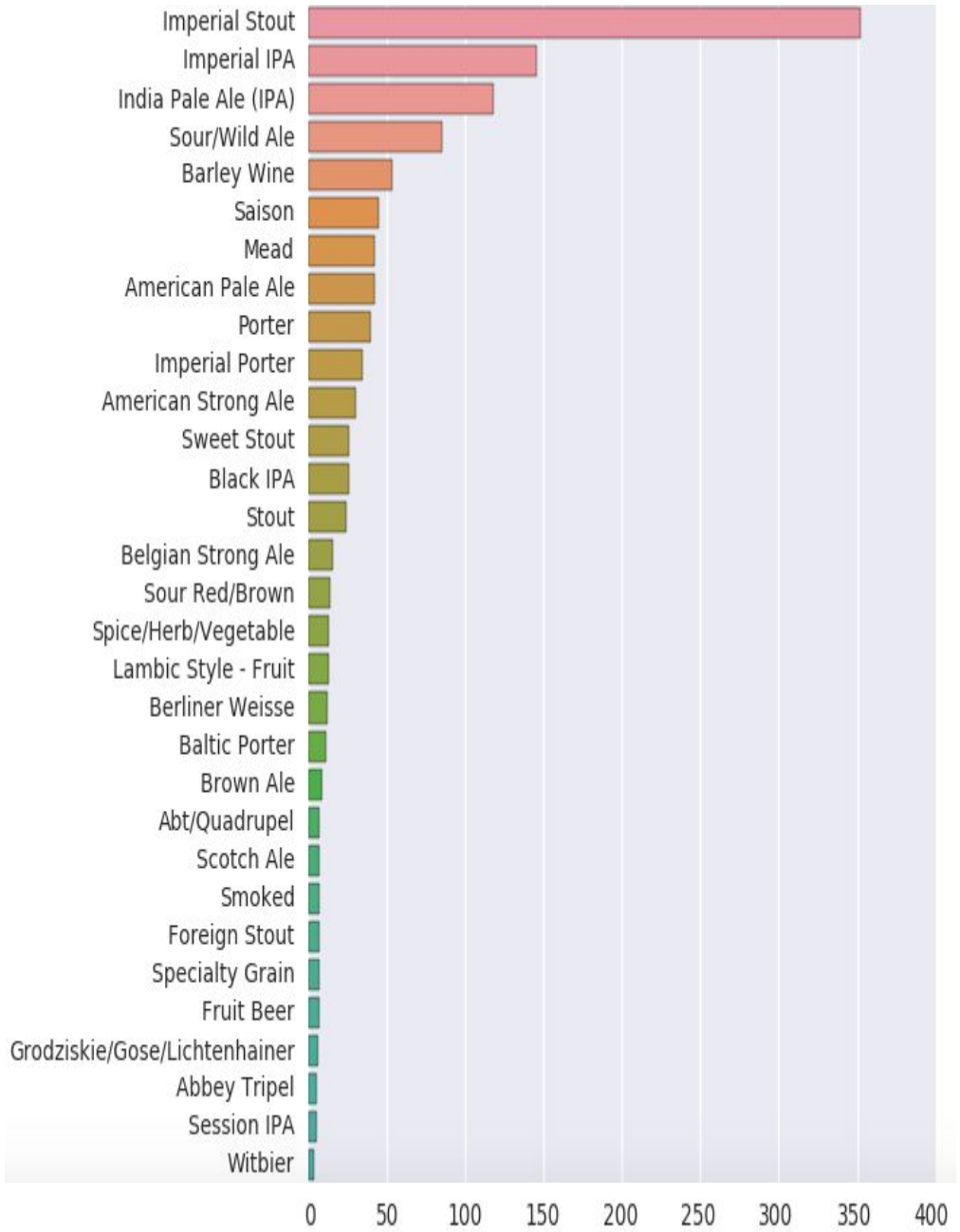

est_cal
600
500
400
300
200
100
0
−100


abv
25
20
15
10
5
0


NK


Imperial Stout
Imperial IPA
India Pale Ale (IPA)
Sour/Wild Ale
Barley Wine
Saison
Mead
American Pale Ale
Porter
Imperial Porter
American Strong Ale
Sweet Stout
Black IPA
Stout
Belgian Strong Ale
Sour Red/Brown
Spice/Herb/Vegetable
Lambic Style - Fruit
Berliner Weisse
Baltic Porter
Brown Ale
Abt/Quadrupel
Scotch Ale
Smoked
Foreign Stout
Specialty Grain
Fruit Beer
Grodziskie/Gose/Lichtenhainer
Abbey Tripel
Session IPA
Witbier
0
50
100
150
200
250
300
350
400


Lambic Style - Gueuze
Cider
Oktoberfest/Marzen
Old Ale
Amber Ale
Traditional Ale
Premium Bitter/ESB
Golden Ale/Blond Ale
Schwarzbier
German Hefeweizen
Weizen Bock
Belgian Ale
Abbey Dubbel
English Pale Ale
Doppelbock
Wheat Ale
Dunkel/Tmav
Dry Stout
Pale Lager
Scottish Ale
Cream Ale
Ice Cider/Ice Perry
Irish Ale
Eisbock
Premium Lager
Imperial Pils/Strong Pale Lager
Kolsch
0
50
100
150
200
250
300
350
400



	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

EDA: Reviews

After Cleaning:

- ~11,370,000 words
  - approximately the number of words in 150 350-page books
- ~278,000 reviews
  - ~average of 40 words per review
- ~127,600 unique words, only 1.1% of number of words.

Top 100 Most Frequent Words

	word	count		word	count		word	count		word	count		word	count
0	head	202336	20	bitter	70922	40	thick	44537	60	sweetness	32477	80	malty	24569
1	aroma	181938	21	fruit	69217	41	tan	43302	61	strong	32197	81	fruity	24222
2	chocolate	157036	22	note	68554	42	full	42951	62	palate	31967	82	hazy	24017
3	flavor	151063	23	roasted	67037	43	orange	42928	63	small	31822	83	38	23809
4	dark	140109	24	vanilla	63879	44	little	41814	64	thanks	31686	84	sour	23459
5	beer	137862	25	body	63641	45	bitterness	41209	65	much	30509	85	glass	23335
6	malt	133818	26	medium	63570	46	really	40971	66	grapefruit	29271	86	poured	22830
7	hop	120424	27	caramel	61161	47	big	40794	67	41	29075	87	cherry	22524
8	taste	118579	28	like	59020	48	4	39795	68	smell	27644	88	ipa	22334
9	sweet	117786	29	well	54050	49	lot	39757	69	hoppy	27111	89	almost	22011
10	coffee	113370	30	color	53194	50	creamy	38704	70	42	27095	90	pretty	21547
11	bottle	112449	31	citrus	53127	51	dry	36666	71	39	27028	91	slight	21434
12	finish	100783	32	one	53013	52	stout	35800	72	thin	26924	92	heavy	21229
13	nice	99787	33	bit	52008	53	oak	35549	73	bodied	26089	93	43	21203
14	black	96429	34	bourbon	51638	54	lacing	35180	74	quite	26087	94	brew	21047
15	pours	94479	35	white	50769	55	hint	34834	75	balanced	25593	95	overall	20993
16	light	84667	36	nose	50315	56	mouthfeel	34820	76	amber	25443	96	wood	20696
17	brown	84258	37	carbonation	49041	57	rich	34449	77	roast	25372	97	golden	20072
18	good	77294	38	great	46202	58	pine	34282	78	deep	25190	98	molasses	19750
19	alcohol	71694	39	smooth	44767	59	pour	32983	79	slightly	24788	99	complex	19715

Top 100 TF-IDF words (distinct, no particular order)

	word	word	word	word	word	word	word	word	word	word
0	bluejacket	terminal	hemp	hibiscus	raspberry	abita	aquavit	fermentoren	neapolitan	cranberry
1	founder	cognac	boysenberry	jalapeno	maltcaramel	himmeriget	blueberry	zakoon	reno	brett
2	tequila	tg	coriander	strawberry	dubbel	firehouse	belgian	chardonnay	triple	mead
3	queen	bourbon	aluminum	tiramisu	oktoberfest	apple	chocolate	apricot	spruce	3708
4	sage	rice	betty	esb	plum	gose	pecan	whiskey	mint	montana
5	dorothy	meridian	kauai	skyview	raisin	maple	ginger	sour	peach	oyster
6	doughnut	pumpkin	bozeman	porter	brandy	blackberry	honolulu	cab	rye	potato
7	nectarine	coffee	crawler	whaleman	cinnamon	envie	hazelnut	cbc	chicory	nelson
8	sockeye	cedar	currant	rum	cherry	mosaic	quad	coconut	abraxas	smoked
9	peanut	knot	port	corn	gingerbread	gin	chili	gesho	papaya	saison

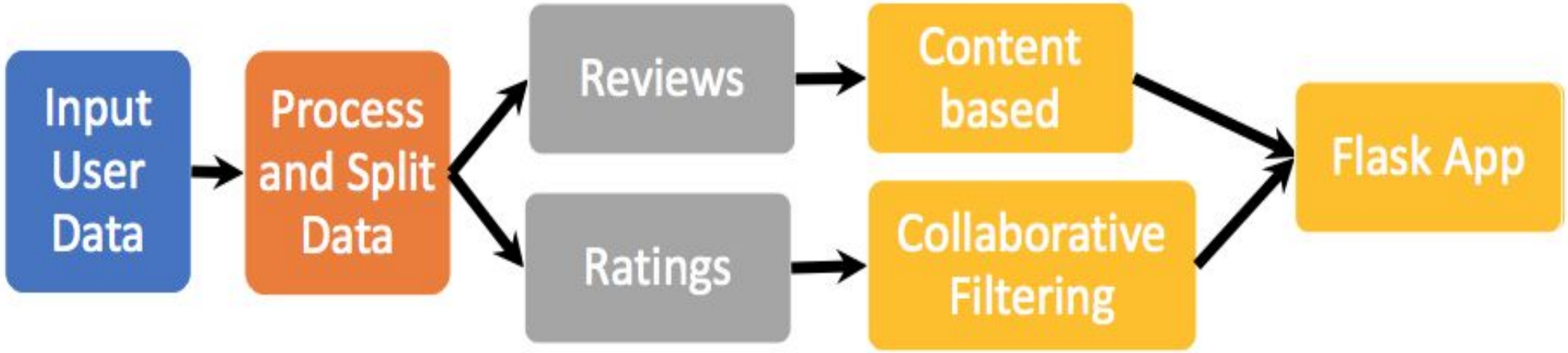
Latent Dirichlet Allocation

1. 0.012*"chocolate" + 0.010*"coffee" + 0.009*"bourbon" + 0.006*"black" + 0.006*"roasted" + 0.004*"roast" + 0.004*"brown" + 0.004*"vanilla" + 0.003*"dark" + 0.003*"stout"
2. 0.006*"pine" + 0.006*"citrus" + 0.006*"grapefruit" + 0.005*"orange" + 0.004*"ipa" + 0.004*"tropical" + 0.004*"hop" + 0.004*"golden" + 0.003*"funk" + 0.003*"mango"

Recommender System

NK

Recommender System



Content-Based Algorithm

- Recommendation based on user review.
- Two algorithms were implemented:
  - Term Frequency-Inverse Document Frequency (TF-IDF) to produce the document-term matrix
  - Latent Semantic Analysis (LSA) does dimension reduction on the document-term matrix

Term Frequency-Inverse Document Frequency (TF-IDF)

- Calculate the “importance” of every word to a review in a corpus.
- Produce a document-term matrix
- 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 much information each word provides, or 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

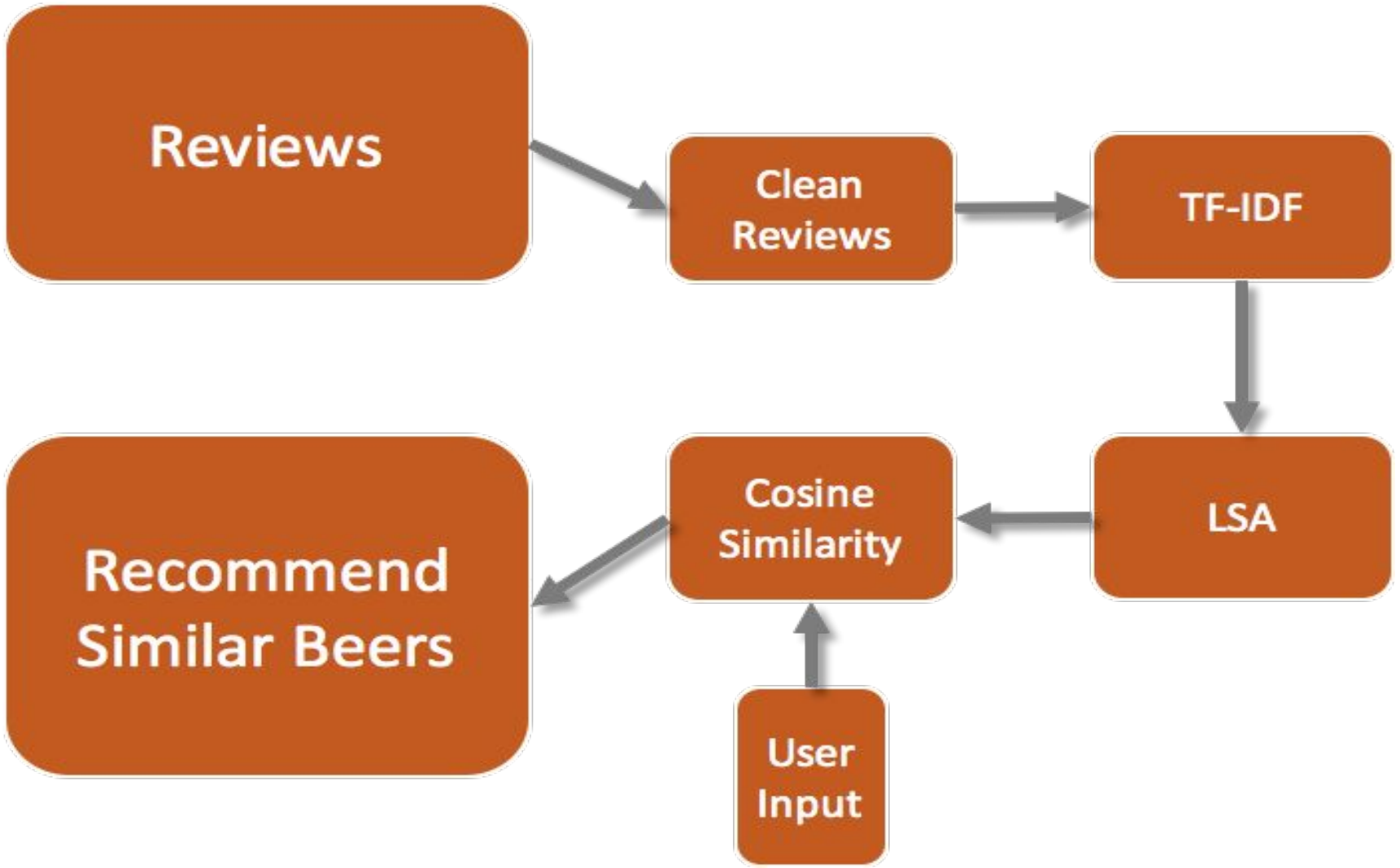
$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D)$$

Latent Semantic Analysis (LSA)

- Similar to PCA, LSA does dimension reduction by performing SVD on the document-term matrix

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

Content-Based Recommender System

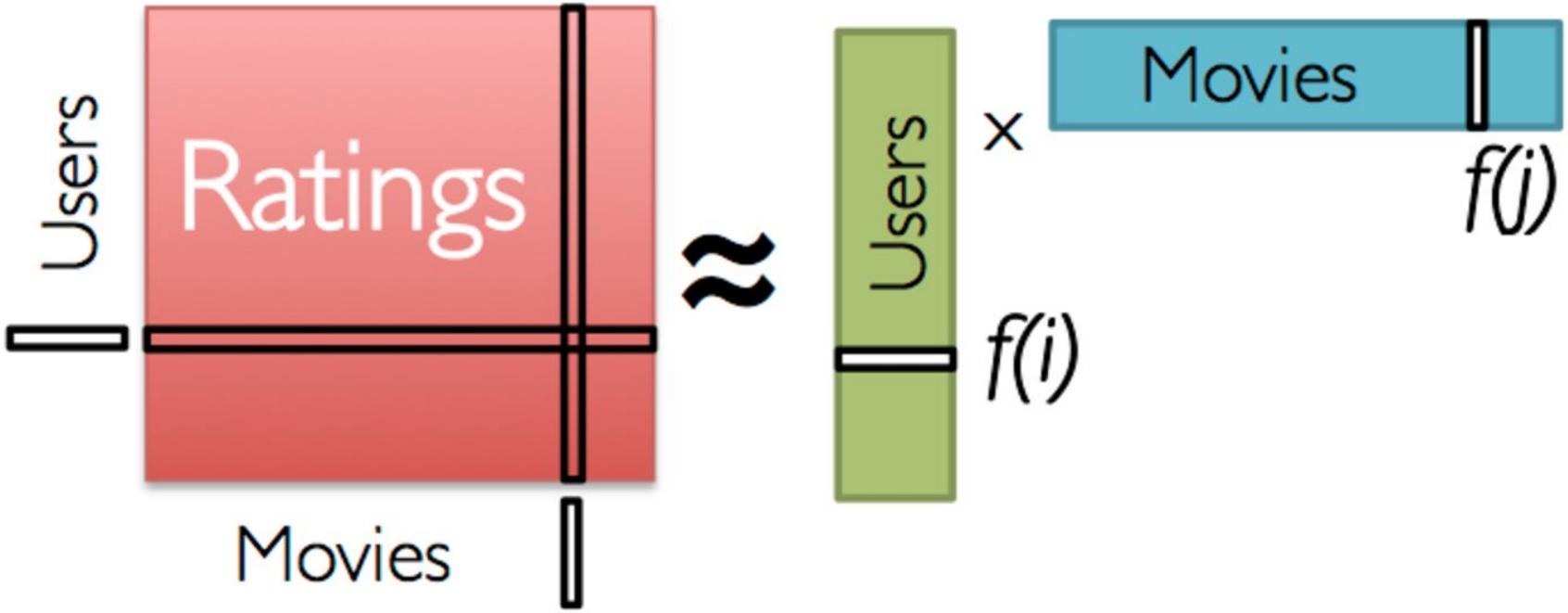


Collaborative Filtering

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

SVD For Collaborative Filtering

- Latent Matrix Factorization of user-item matrix to 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
- Further decompose ratings to global average, user/item biases, implicit feedback, and latent features
- Remove bias of each user and item to center data
- Users that have less ratings are penalized more (given rating closer to average)

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

Predicted rating

Global average

User bias

Item bias

Item Latent feature

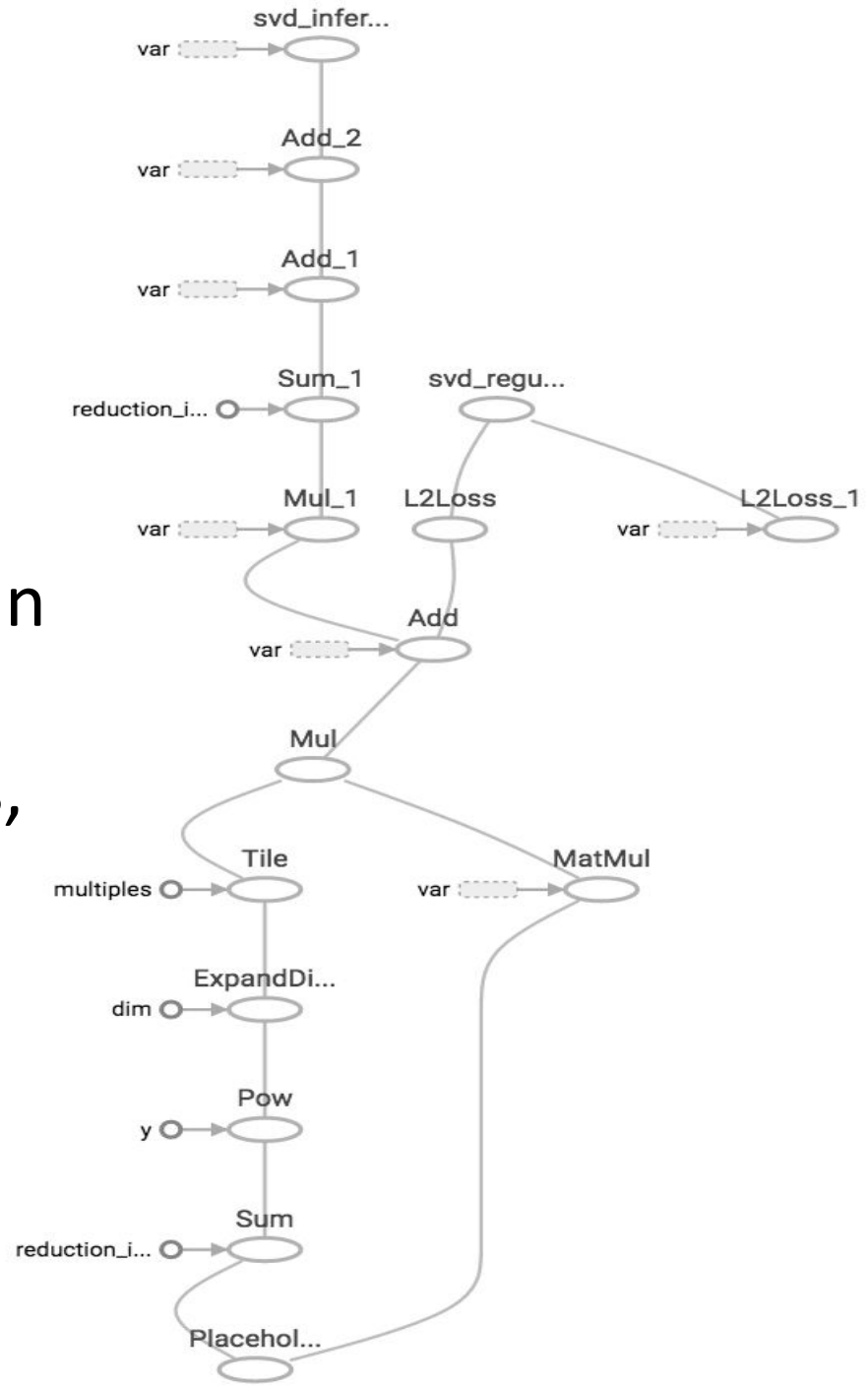
User latent feature

Implicit feedback

Implicit feedback parameters

SVD++ Implementation

- No fast publicly available package
- Realized SVD++ can be highly optimized using ***TensorFlow***
- Found vanilla SVD algorithm built in Tensorflow
- Augmented code to include biases, implicit feedback, k-fold cross validation, and early stopping
- Accelerated by C++ backend, GPU computation
- **Results: RMSE = 1.69**

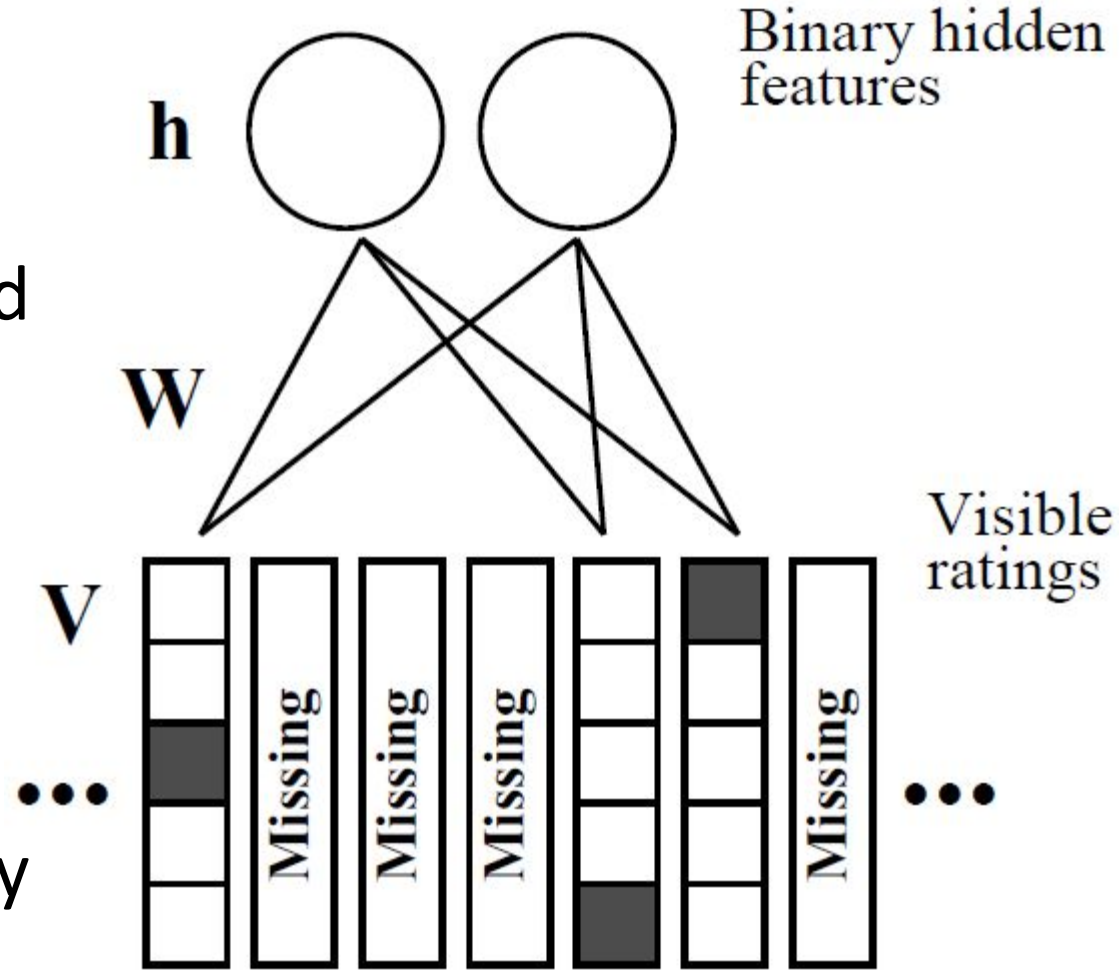


Restricted Boltzmann Machine (RBM)

- RBM is an unsupervised, 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

Prediction Method for New Users

- Neighborhood approach
- Compute cosine similarity between new user and all users
- Impute missing ratings of new user by a weighted average proportional to similarity metric
- Rank recommendations and output top 5

Flask App Demonstration

NK

Lessons Learned & Future Steps

NK

Lessons Learned

- Text is fundamentally messy to work with
- When an algorithm is not available, implement it yourself
- Hacking other people's code can be time consuming (RBM), document your code for readability
- Workflow is important, also good to document workflow
- Don't try to build Flask app from scratch in two days

Future Steps

- Add in more features (i.e style, palate, appearance, aroma, taste, and time)
- Add in more data
- Tune hyperparameters to get optimal single models
- Ensemble smarter (use minimizer or stacking)
- Develop App aesthetics and functionalities further

We hope you enjoyed our presentation, Cheers!

NK