Ok F🍩🥝die!

Recommendation App

Restaurant Decision Tool For Two!

September 2018 ◆ NYC-Data Science Academy

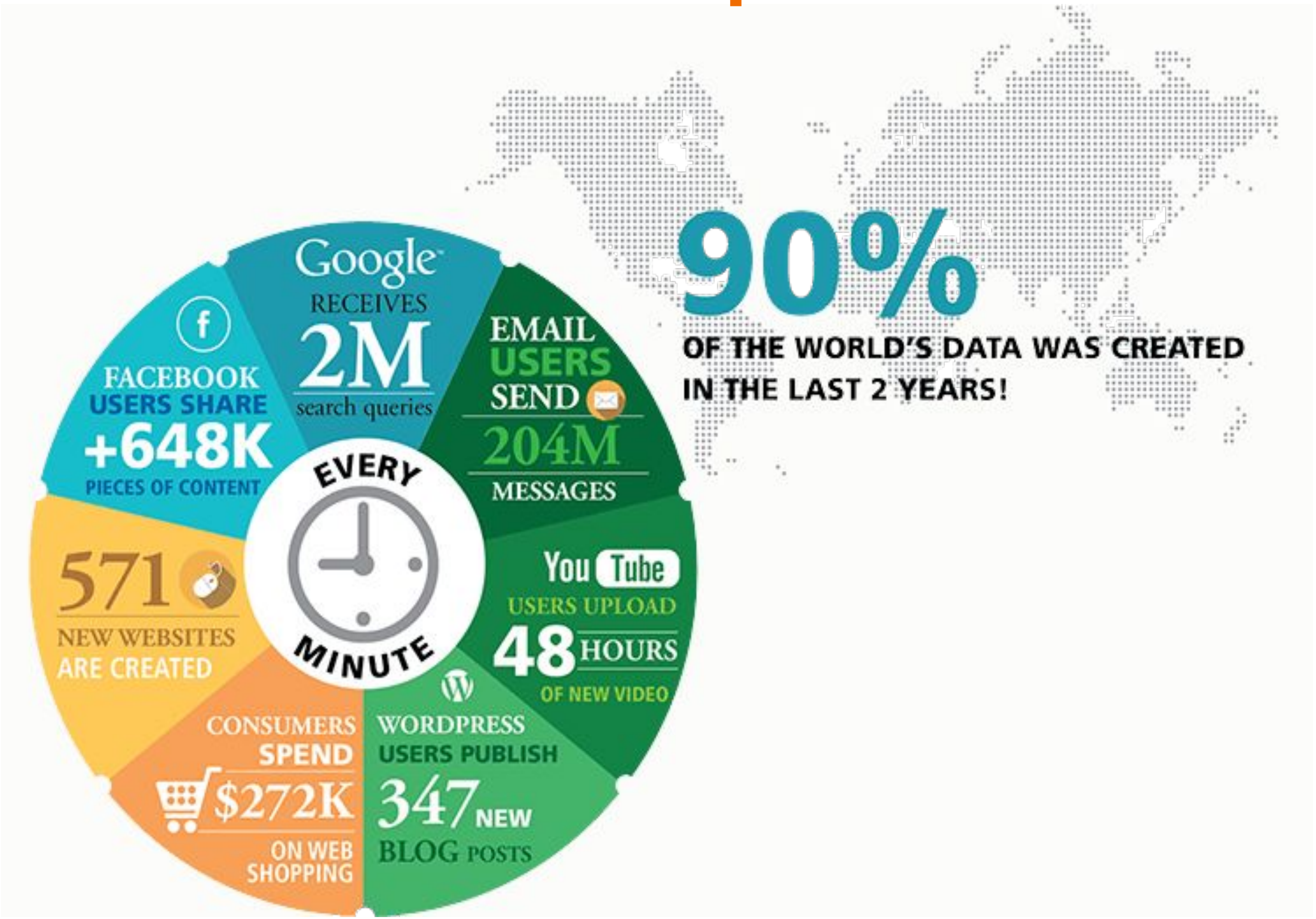
Unsupervised Foodies ◆ Ariani Herrera, Elsa Vera, Eric Moore, Erin Dugan, Murugesan Rangasamy

Outline

- Introduction (Ariani)
- Project Dataset (Muru)
  - • Understanding the Features
  - • EDA
- Restaurant Recommendation Algorithms
  - TF-IDF/NLTK (Elsa)
  - Doc2Vec Methods (Erin)
  - Collaborative Filtering/Hybrid (Ariani)
- Results and Comparison of Methods (Eric)
- Ok Foodie! App Demonstration (Muru)
- Conclusions & Future Directions

Ever had this conversation?

Big Data - The Abundance of Options



Paradox of Choice

- Modern consumers are inundated with choices.
- Retailers and content providers offer a huge selection of products, with unprecedented opportunities to meet a variety of special needs and tastes.
- Matching consumers with the most appropriate products is key to enhancing user satisfaction and loyalty.
- Personalized recommendation systems have become a keen interest for companies to enhance their market share.
- Good personalized recommendations add another dimension to a user's experience.
- Many companies combine collaborative & content based methods.

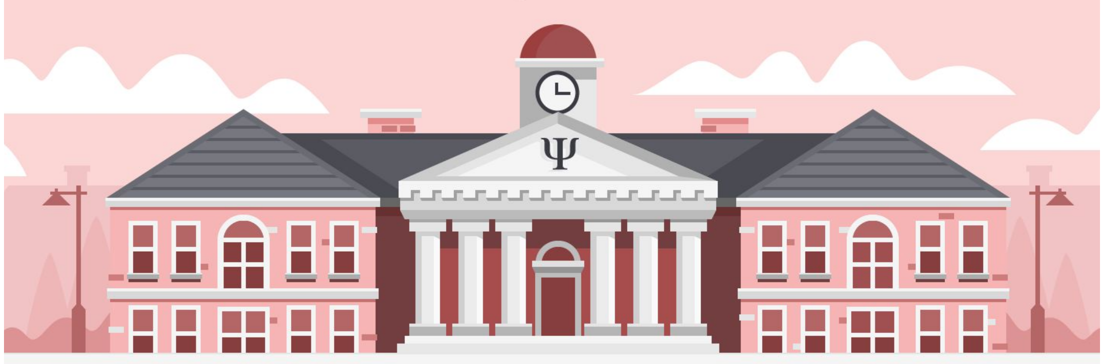
Dataset Challenge

Round 12

Competition to use data in innovative ways

Users, Businesses, Reviews, Tips, Check-ins, Photos

From 10 metro areas in 2 countries



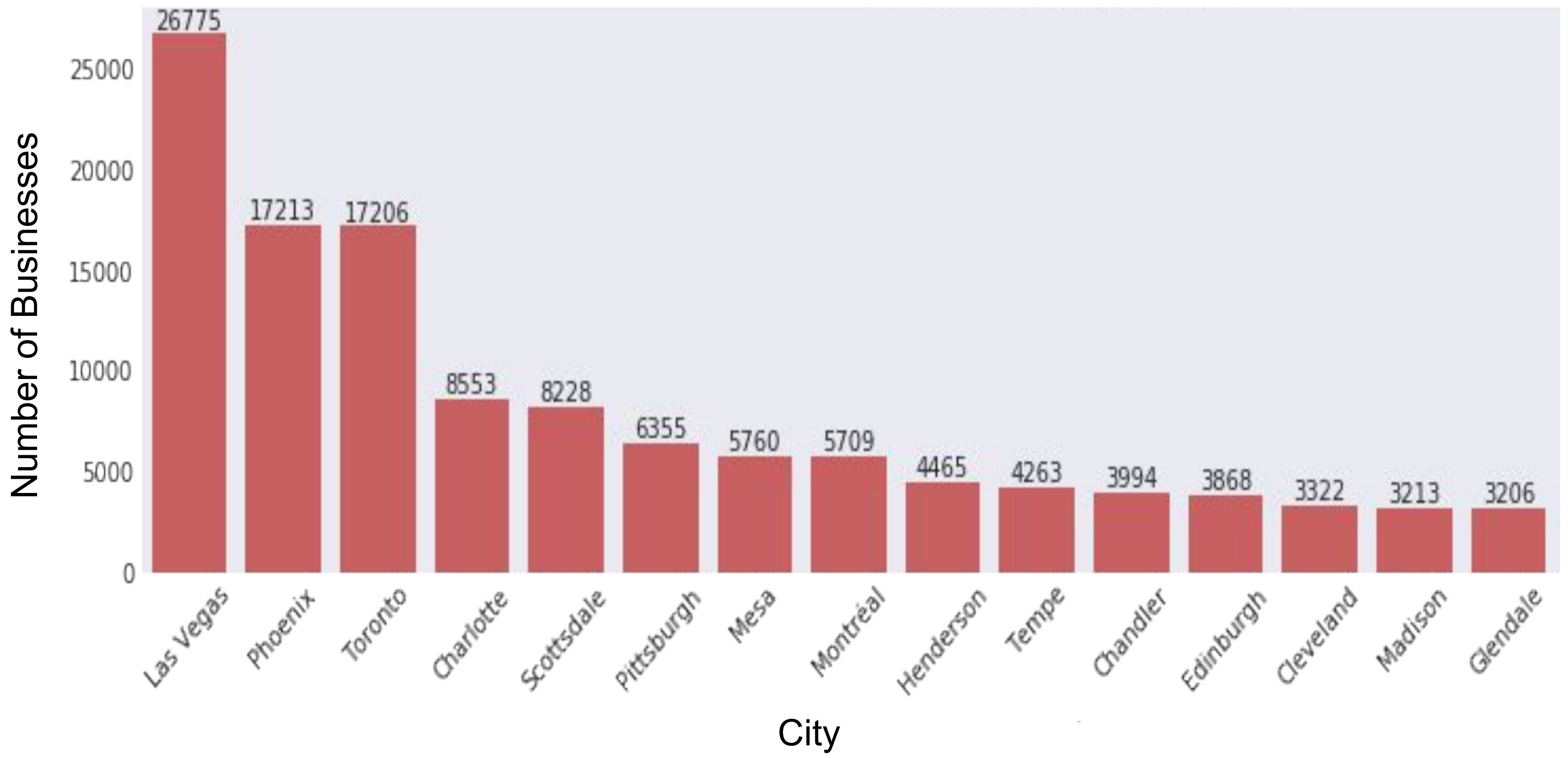
"Unsupervised Foodies"

Project Focus:

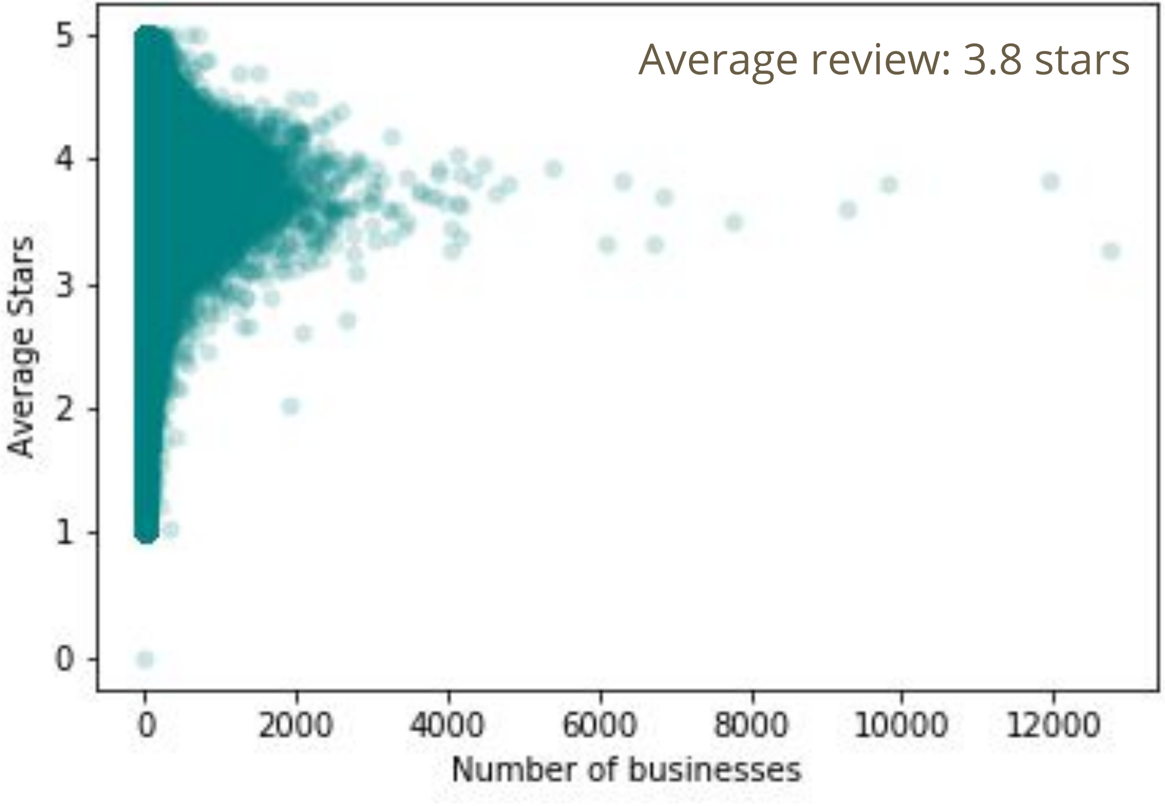
Las Vegas Restaurants

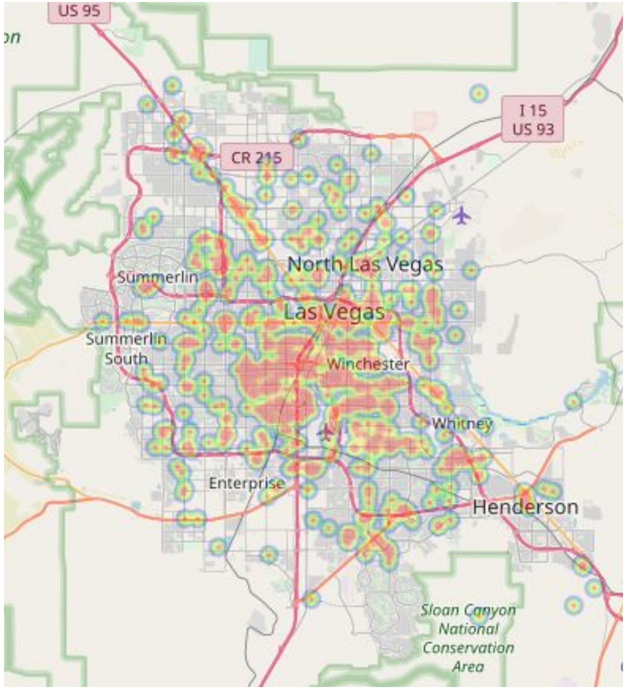
https://www.yelp.com/dataset/challenge

Why Las Vegas?



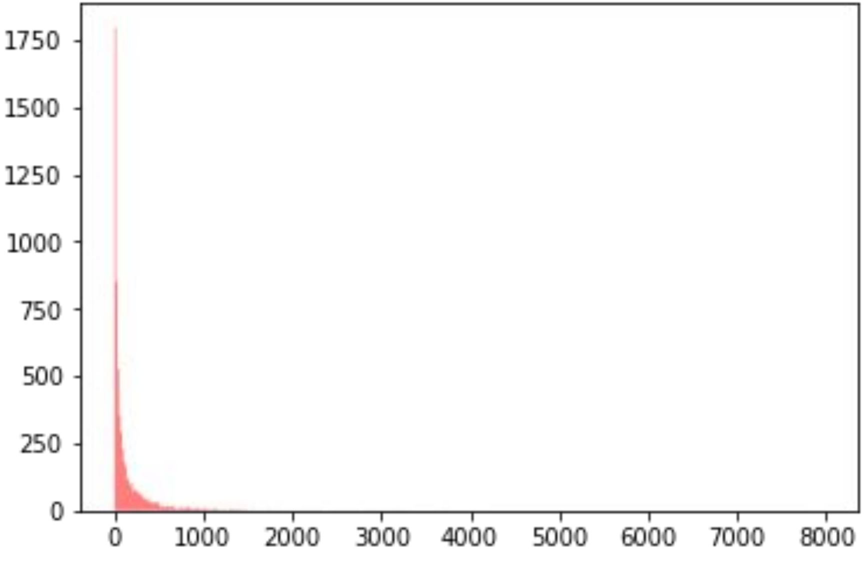
Exploratory Data Analysis



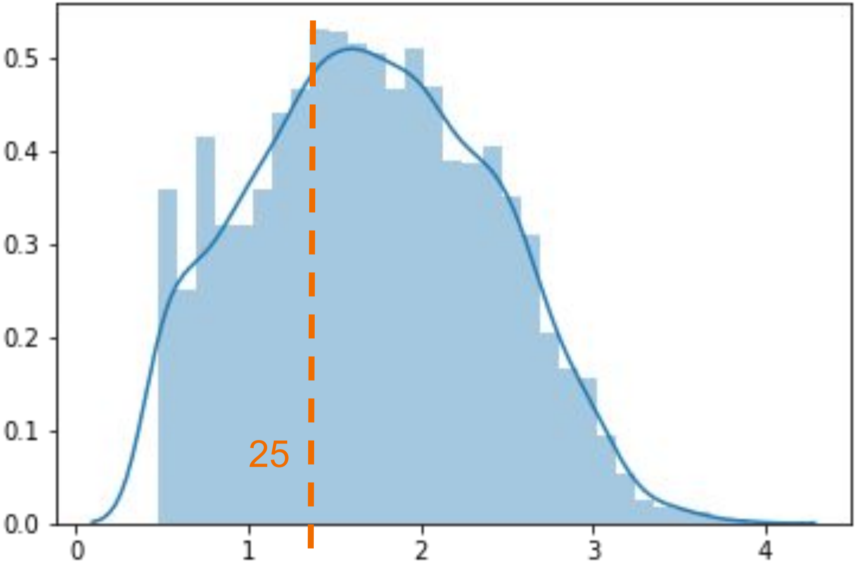


Exploratory Data Analysis

6153 restaurants in Las Vegas, 4064 are currently open; 3020 have > 25 reviews



# of reviews per restaurant

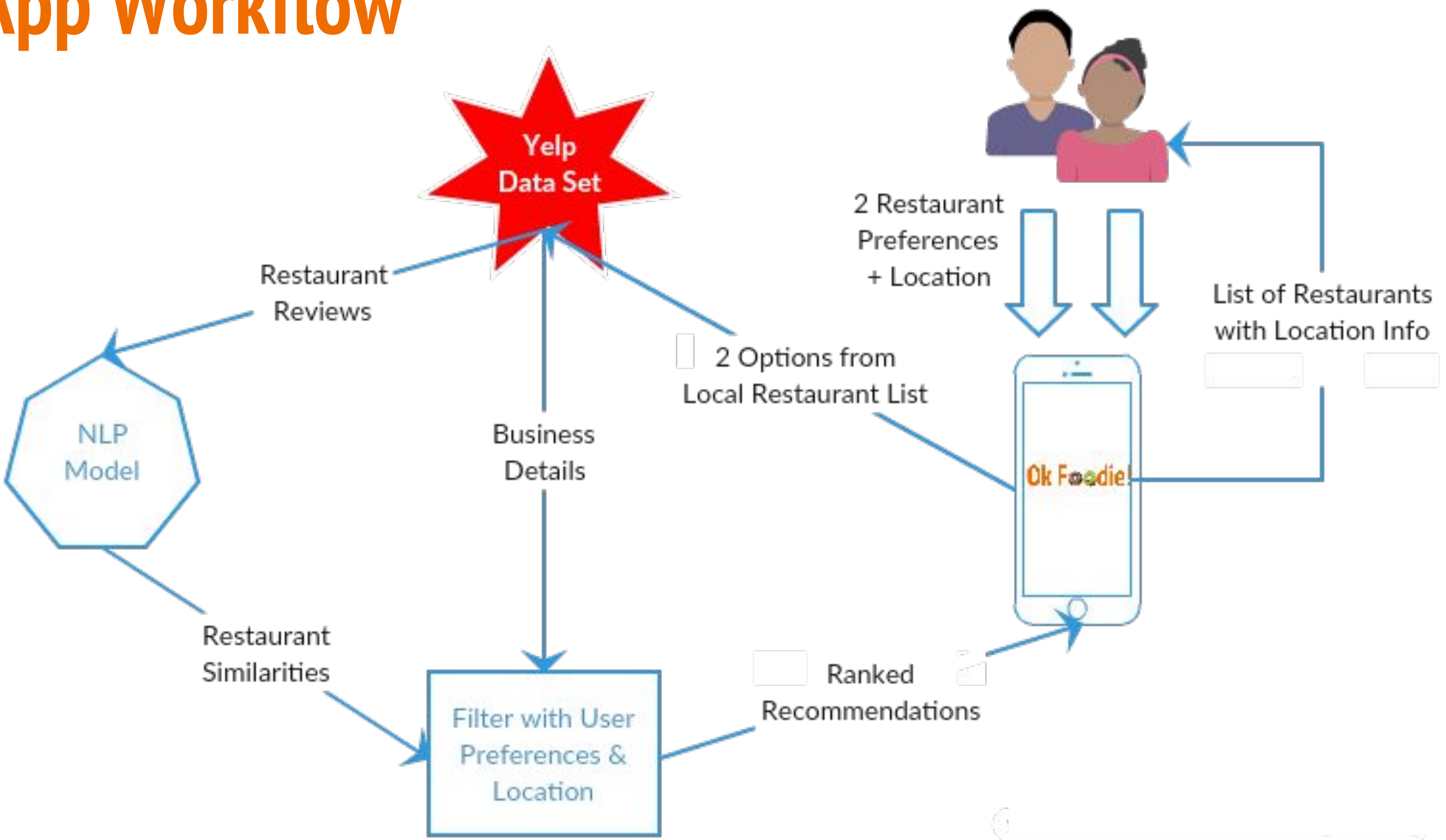


$\log_{10}$(reviews)

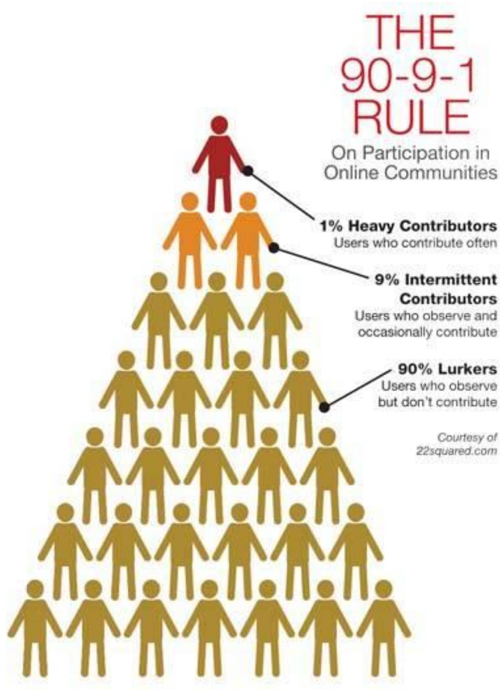
Who are the Users?

- Average rating is 3.8 stars
- 80% of the reviewers only write 5 reviews
- Yelp Elite users are more influential and considered extremely active prolific users, therefore their reviews should have more weight.
- Yelp users with friends are more likely to trust their friend’s opinions on a restaurant.

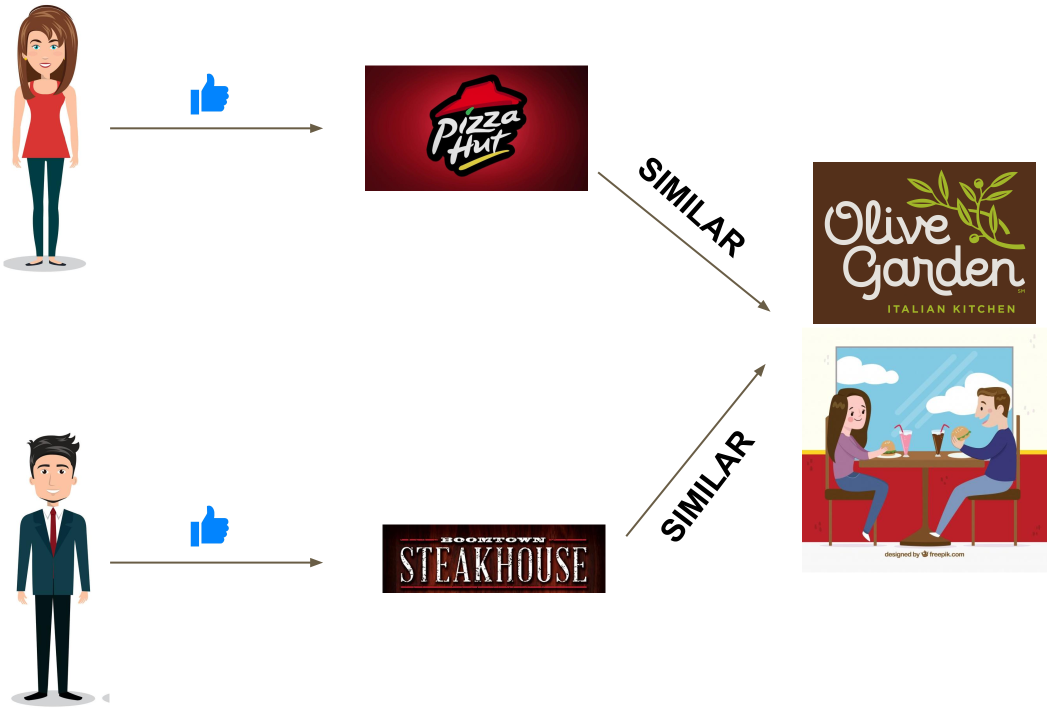
App Workflow



Yelp and the 1/9/90 rule: Approach (unknown users)



Content (item) -based recommendation



Similarity based on text vectorization

Pre-processing

- Tokenize
- Lemmatize
- Filter stop words
- Filter infrequent words

Document-to-vector models (Word Embedding)

Frequency based Embedding:

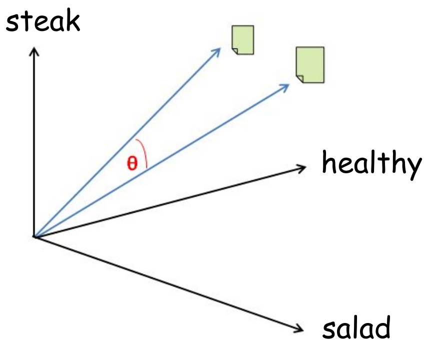
- Count vector
- TF-IDF (NLTK, Gensim)
- Co-occurrence vector (GloVe)

Prediction based Embeddings:

- Word2Vec & Doc2Vec (Neural networks, Gensim)

Document Similarity

1. **Cosine similarity**

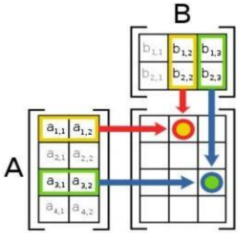


2. **Rank restaurants**

A

Term Frequency x Inverse Document Frequency

$\underbrace{\text{Term Frequency}}_{\text{bag of words}} \times \underbrace{\text{Inverse Document Frequency}}_{\text{importance of the word in the document}}$

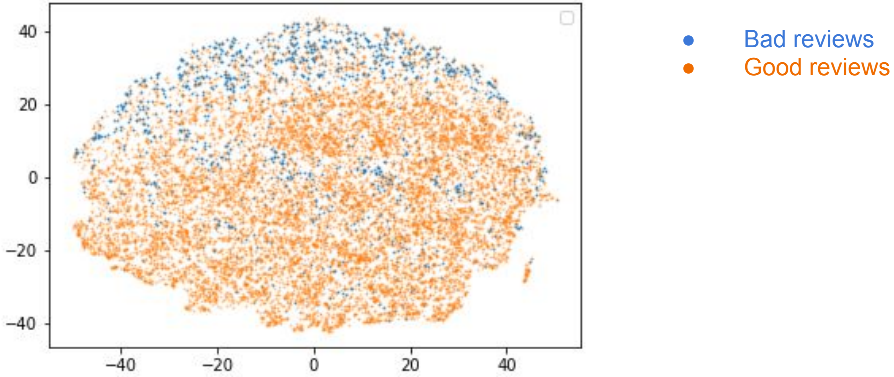


B

$\underbrace{\text{GloVe Embedding vector}}_{\text{Co-occurrence vector}}$

TF-IDF weighted sum of embedding vectors

- ➢ Reduce dimensionality: PCA (75500,300) → (75500, 8)
- ➢ t-SNE (75500, 2)
- ➢ Feed into collaborative filtering model



Natural Language Processing with Neural Networks

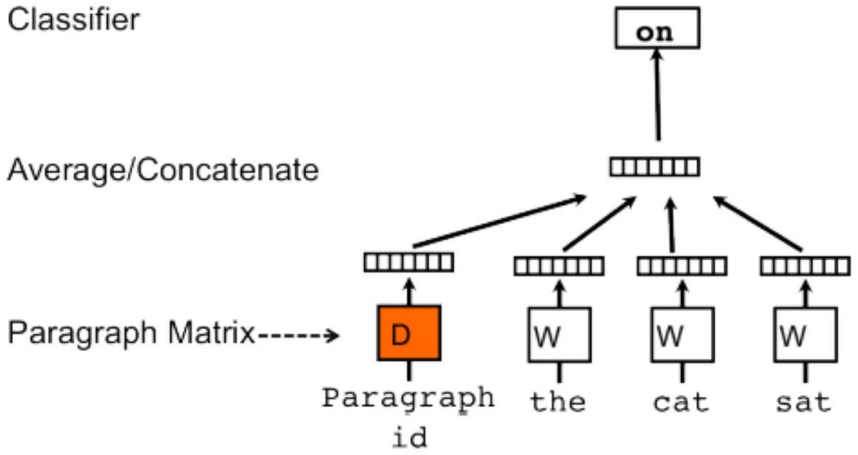
Word Vectors

Hidden layer of neural network

Word representation transformed to probability distribution of nearby words

Paragraph/Document Vectors

Fixed-length feature representations of *contextual meaning* from text of any length



Inferring Similarities From Words

Find similarities from word frequencies:

- menu items
- service quality

Comparable Options

Inferring Similarities Through Context

Find deeper meaning within the reviews:

- menu items
- service quality
- uniqueness
- reputation
- ‘cult’ status
- high ratings
- customer base

Expanded Options

NLP Vectorization of Yelp Reviews

Using GenSim Doc2Vec Module:

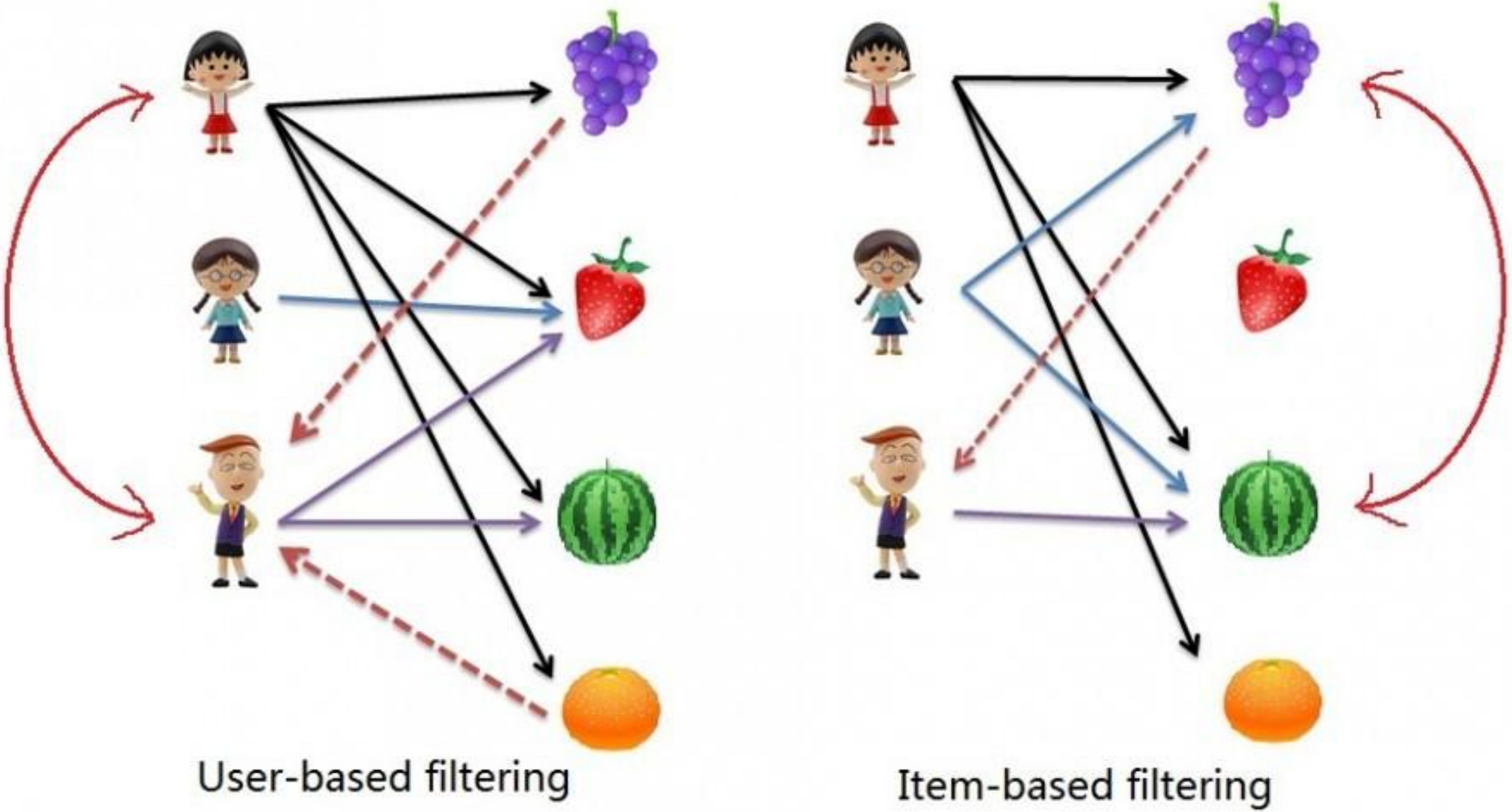
Pre-process	Model	Vectors	Compare
• 25 sampled reviews per restaurant • Tokenize, remove punctuation, lowercase • Tag documents	• Train on vocabulary & context	• 200-dimension document vectors	• Vector similarities of reviews • Median values • Rank results & merge both users' restaurants

TF-IDF Vectors
+
Doc2Vec Vectors

→

Collaborative Filtering Predictive Model

Joining Forces

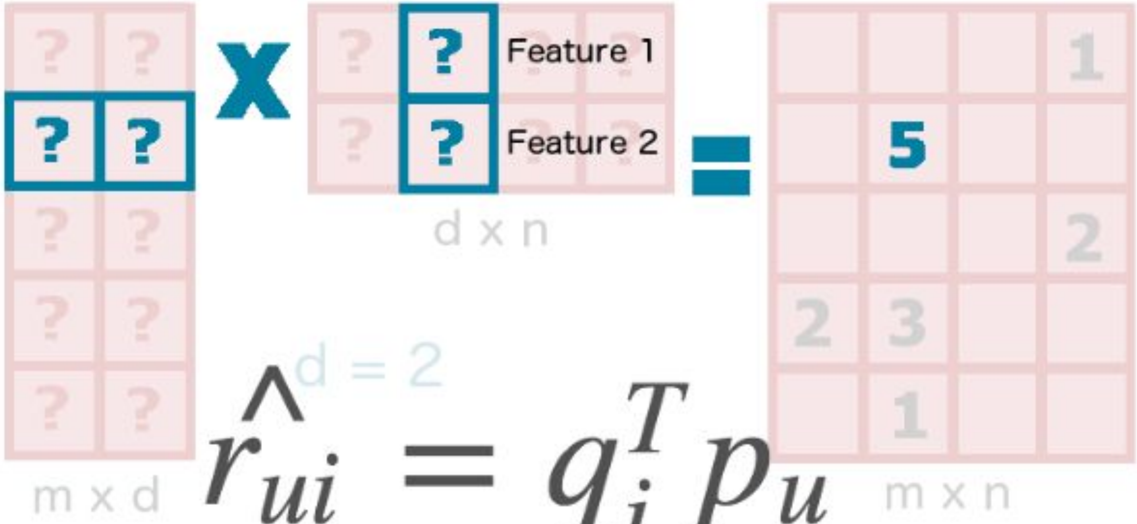


Using Matrix Factorization

- Collaborative filtering analyzes relationships between users and interdependencies among products to identify new user-item associations.
- Matrix Factorization is a mathematical tool used to discover the latent (hidden) interactions between users and restaurants, driving the raw data (which we understand as tastes and preferences).
- Matrix factorization is the breaking down of one matrix into a product of multiple matrices.

Matrix Factorization

m = number of users, n = number of items
choose d, the number of features



$$\hat{r}_{ui} = q_i^T p_u$$

Using Collaborative Filtering for Two

The matrix factorization model provides a rank and predicted score for each user and restaurant.

User A

Name	Score	Rank
Incognito Wraps	4.413172	1
Taco Naco	4.348425	2
Kame Omakase	4.327258	3
Chef @ Your Home	4.319013	4

\+

User B

Name	Score	Ranks
Taco Naco	4.327258	1
Pollos El Dorado	4.32112	2
Incognito Wraps	4.319013	3
Kame Omakase	4.310336	4

=

Recommendation

Name	Rank
Taco Nacho	3
Incognito Wraps	4
Kame Omakase	7

Model comparison - collaborative filtering

Model	RMSE
Baseline	1.01601
+ Feature Engineering	0.9556
+ NLP	0.82792

Comparison of Approaches

- Collaborative filtering prediction easy to evaluate
- Unsupervised learning lacks direct evaluation metric
- Use collaborative filtering to evaluate new user recommendations
- Intuition: match new user to existing user(s) that rated choice similarly
- Predict all ratings using collaborative filtering
- Compute average ratings among similar users
- Estimate quality of recommendations across models

Model evaluation: overview

- Use selected restaurants as starting point
- Sample set of restaurant pairs to simulate input from app
- For any given restaurant combination:
  - Identify existing Yelp users that also like the input choices
  - Select top 25 users with highest *demeaned* rating for each
- Identify restaurants that comparison group likes
- Compute probability both “unknown” users would like recommendations
- Iterate over sample

Model evaluation: parameters

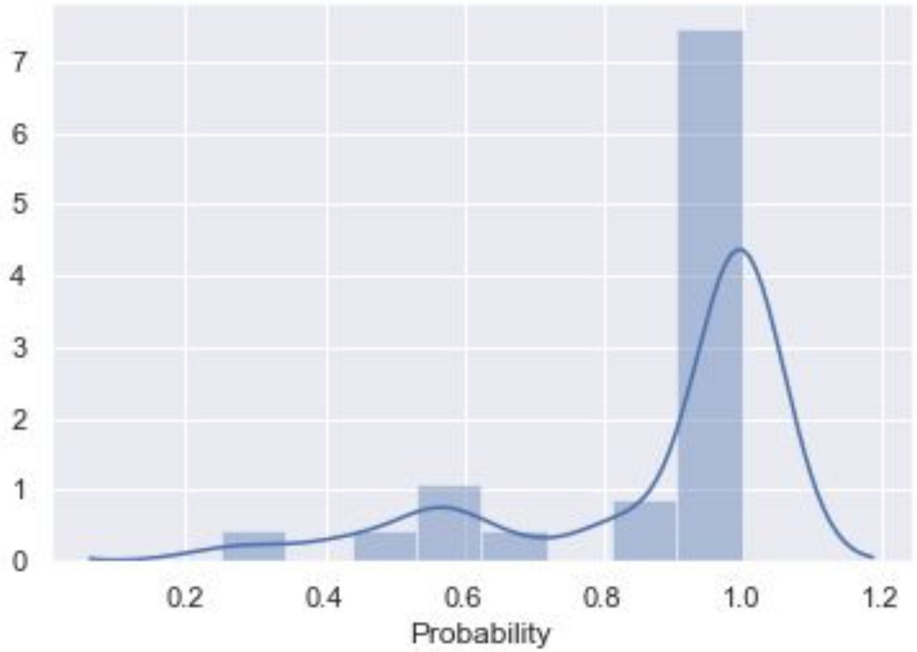
- We explored several parameters
- Comparison group should also like input choice
  - Star rating threshold for comparison group
- Tokenized review distribution can vary a lot across reviews
  - Mean versus median aggregation of individual reviews
  - Weighting tokenized reviews differently

Model evaluation: tf-idf based on nltk

What is the probability that both individuals like the recommendation?

	Review filters			
	None	3+	4+	5 only
Median, no weights	0.034062	0.034062	0.034062	0.020609
Median, weighted	0.059977	0.059977	0.059977	0.020340
Mean, no weights	0.032545	0.032545	0.032545	0.019433
Mean, weighted	**0.200828**	**0.200828**	**0.200828**	**0.058580**

Density of nltk recommendation quality



Product Demo

Conclusions & Next Steps

- Continue Model Refinement:
  - Incorporate larger samples of dataset to improve accuracy and personalization
  - Add text from user tips to NLP models
- Scale Up:
  - DataBricks/PySpark
  - Include more cities from the Yelp Challenge DataSet
  - Link to Yelp API
- Next Generation:
  - Mobile app
  - Personalized recommendations for registered Yelp users
  - Optimizing models based on user selections

Acknowledgements

- Yelp
- NYC DS Academy Teaching Staff
- Teaching Assistants (Dragos)
- Our Patient Friends & Families

Questions and Suggestions

Final Product