Complementary Venue Recommendation Model for Yelp

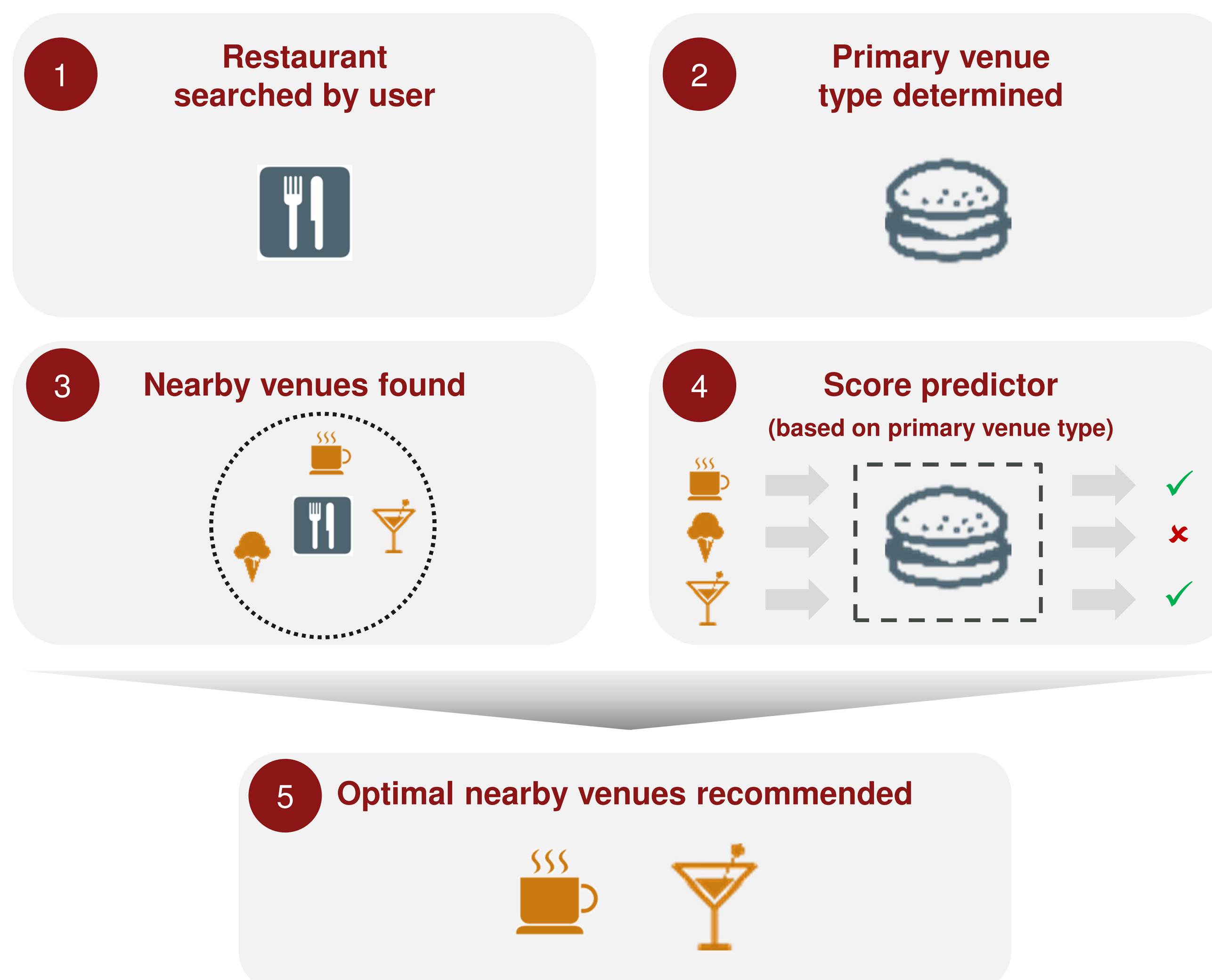
Hyun Sik Kim (hsik@stanford.edu), Ryan Wong (rawong@stanford.edu)

Stanford ENGINEERING
Electrical Engineering

Overview

We developed a **novel machine learning model** that **recommends nearby complementary venues** (e.g. café) that the user is likely to enjoy based on a restaurant searched using Yelp.

- To simplify the search process for selecting multiple venues for a single outing using Yelp
- Our model **recommends complementary venues** using binary classifiers that predict the venue score using review text:
  - **Great venue (1):** 4-5 star rating
  - **Mediocre / poor venue (0):** 1-3 star rating
- Optimal recommendation/s based on:
  - **Highest predicted score**
  - **Proximity** (within 1 mile of restaurant)
- Model trained based on review text written by a **common reviewer** of the restaurant and complementary venue



Data and features

Feature vector comprised the count of **unigrams and bi-grams** in **Yelp reviews** of **complementary venues** from the **Yelp academic dataset**.

1 Review text pre-processing

Converted to lower case

Punctuation removed

Stop words removed

Porter stemming algorithm

2 Extract features and scores of reviews

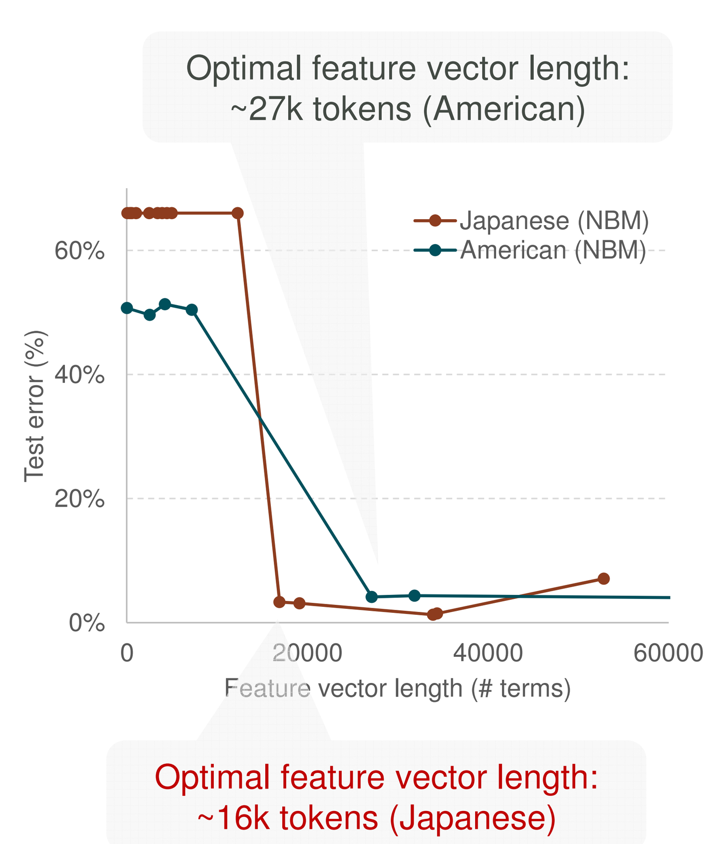
yelp Academic Dataset

Comp. Venues	Unigrams: good	...	bad	Bigrams: very good	...	dont go	Score: 1-3 stars (0) / 4-5 stars (1)
Review 1	1	...	1	0	...	2	0
Review 2	2	...	0	3	...	0	1
...	...	...	...	...	...	...	...
Review m	0	...	1	0	...	1	0

Feature Vector — Labels

TF-IDF and normalization applied

3 Feature selection using MI



Model performance and evaluation

Our **Naive Bayes with TF-IDF and normalization** implementation exhibited **~3-4% test error performance**. This was comparable to SVM performance.

Naive Bayes comparable to SVM – linear was optimal SVM kernel

American Restaurants

Average Error rate	Training	Test[1]
Naive Bayes (TF-IDFN[2])	2.2%	4.1%
SVM with linear kernel	2.9%	5.5%
SVM with linear kernel (TF-IDFN[2])	1.4%	5.1%
SVM with RBF kernel	1.8%	6.2%

Sample	Training	Test[1]
No. of complementary venues (#)	51	51
No. of reviews (#)	5,652	565
Average review length (words)	130	130

Naive Bayes robust despite conditional independence assumption

Linear outperformance over other kernels suggest **text classification** problem **linearly separable**

Both models benefited from **TF-IDF and normalization**

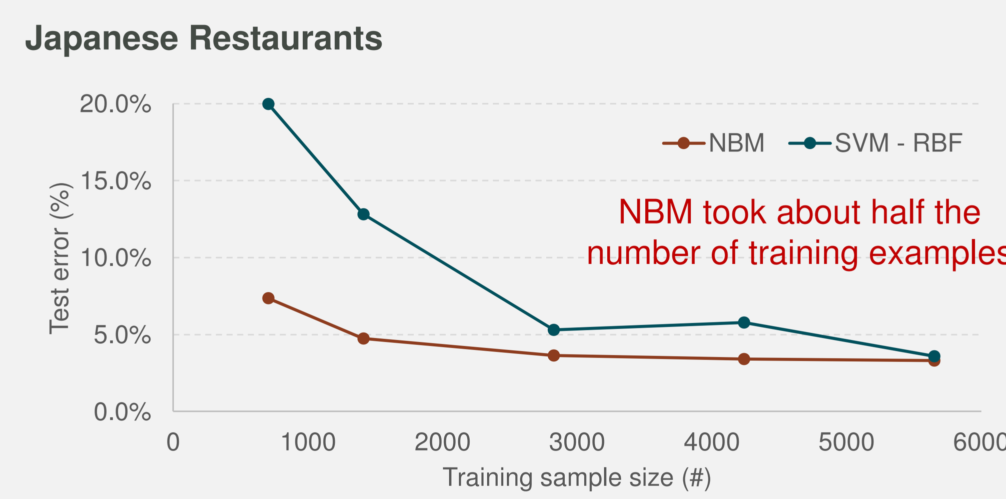
Japanese Restaurants

Average Error rate	Training	Test[1]
Naive Bayes (TF-IDFN[2])	1.8%	3.3%
SVM with linear kernel	2.5%	3.8%
SVM with linear kernel (TF-IDFN[2])	0.5%	2.4%
SVM with RBF kernel	1.9%	3.6%

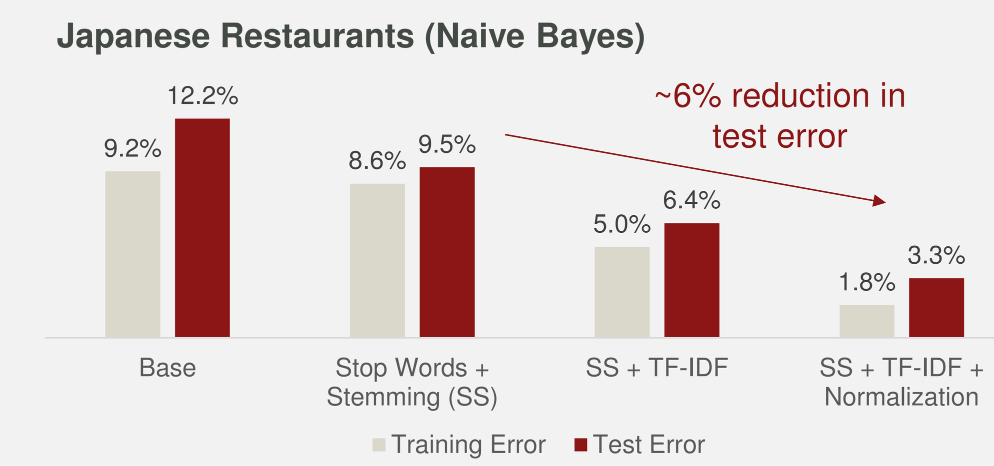
Sample	Training	Test[1]
No. of complementary venues (#)	20	20
No. of reviews (#)	5,652	565
Average review length (words)	136	136

(1) Test error based on k-fold cross validation with k = 10 over the training sample. Hence, test sample essentially refers to the randomly selected folds, each one-tenth the size of the training sample.
(2) TF-IDF and normalization.

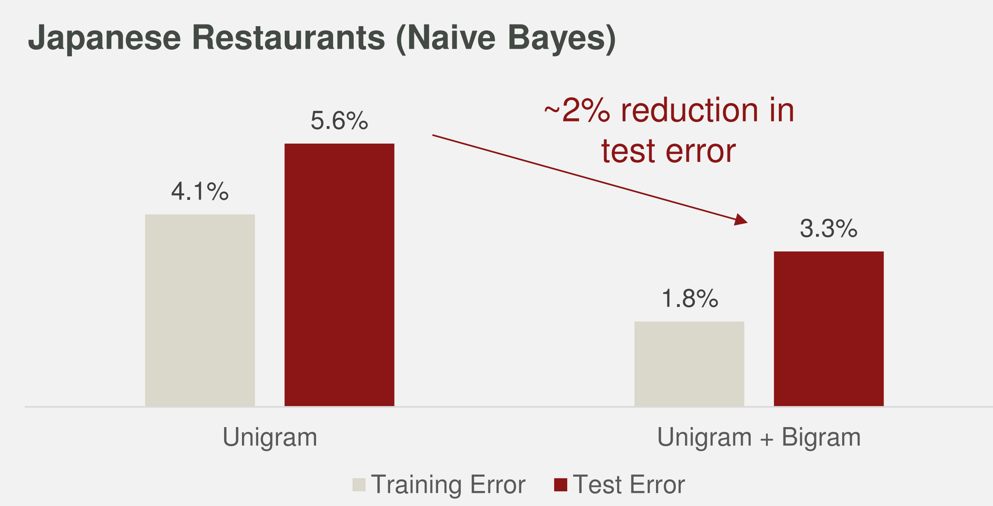
Naive Bayes was a faster learner



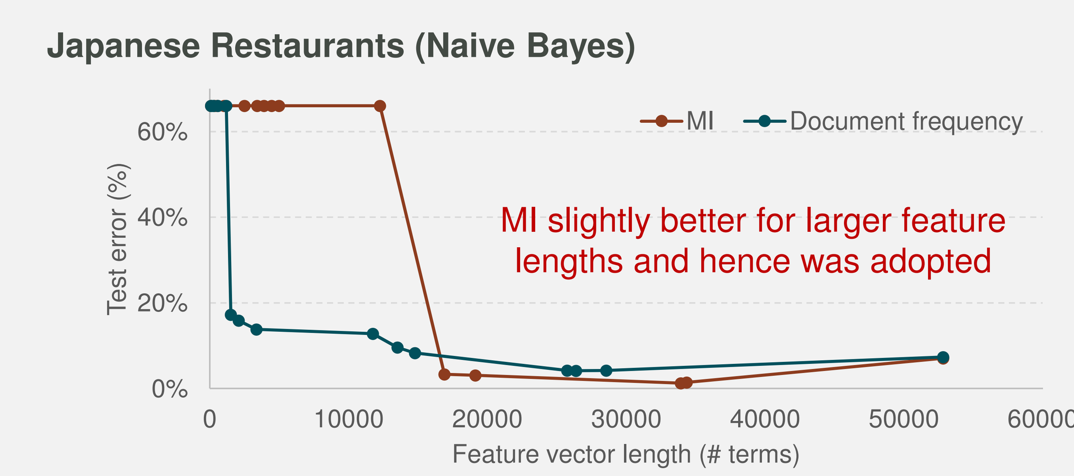
TF-IDF and normalisation key to Naive Bayes accuracy



Bigrams improved performance



Document frequency was a simple and effective filter but MI was better



Future work

- Second classifier that distinguishes between 4 and 5 stars to provide finer granularity of venue scores
- Vector representations of words (word2vec)
- Exploration of n-grams to explore predictive power of phrases and idioms