Classification of Restaurants from Customer Reviews

Navit Gaur
Raghavendra Guru
Shubham Vadhera

Motivation

- Currently, classification of restaurants is done by what restaurants say
- We should be able to classify a restaurant based on customer reviews
- A restaurant may be American by classification but people mostly like it for its Noodles - thus its Chinese as per the customer reviews
- Cheesecake factory serves amazing Pasta but classified as Desserts - should be classified as Italian as well

The Cheesecake Fac ×

Shubham

https://www.yelp.com/biz/the-cheesecake-factory-santa-clara-2

Apps Archive SJSU Jobs Personal Social Study Cloud9 WIP SleepIQ Work 239 297

yelp

Find tacos, cheap dinner, Max's | Near San Francisco, CA | Sign Up | Log In

Home About Me Write a Review Find Friends Messages Talk Events

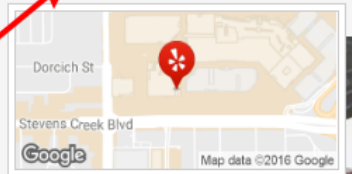
The Cheesecake Factory

Claimed

1399 reviews Details

$$ · Desserts, American (Traditional)

Write a Review | Add Photo | Share | Bookmark



3041 Stevens Creek Blvd
Santa Clara, CA 95050
West San Jose

Edit

Get Directions

(408) 246-0092

thecheesecakefactory.com

See all 1694

Buy a $25 gift card & get a slice of cheesecake. Order Online

"The only bad part is the usual hour long wait :(Ordered the **Chicken Madeira** which always is delicious." in 46 reviews

"The spinach artichoke dip, luau salad, pesto pasta, burgers, and **miso salmon** are just a few of my go-to selections." in 42 reviews

"My favorite options so far have been the: **Pasta Da Vinci**, Chipotle Chicken Pasta and Factory Meatloaf." in 26 reviews

Show more review highlights

Today **11:00 am - 12:30 am**
Open now

Menu

$$$$ Price range **$11-30**

Hours

Mon	11:00 am - 11:00 pm
Tue	11:00 am - 11:00 pm
Wed	11:00 am - 11:00 pm
Thu	11:00 am - 11:00 pm
Fri	11:00 am - 12:30 am Open now
Sat	11:00 am - 12:30 am
Sun	10:00 am - 11:00 pm

Edit business info

Dataset

Yelp.com academic dataset

- 2.7M reviews
- by 687,000 users
- for 86,000 businesses

Allrecipes.com

- Scrapped over 300,000 URLs to extract
- 218,924 recipes - 77,830 unique recipes
- 11,033 unique food words

Wordfrequency.info, Insightin.com

- About 7000 unique commonly used words in English language - to be removed from reviews

Nltk corpus for stop words

Tools and technologies

Natural Language
Analyses with NLTK

python™

PC

SciPy

NumPy

matplotlib

IP[y]: IPython
Interactive Computing

pandas

scikit learn

jupyter

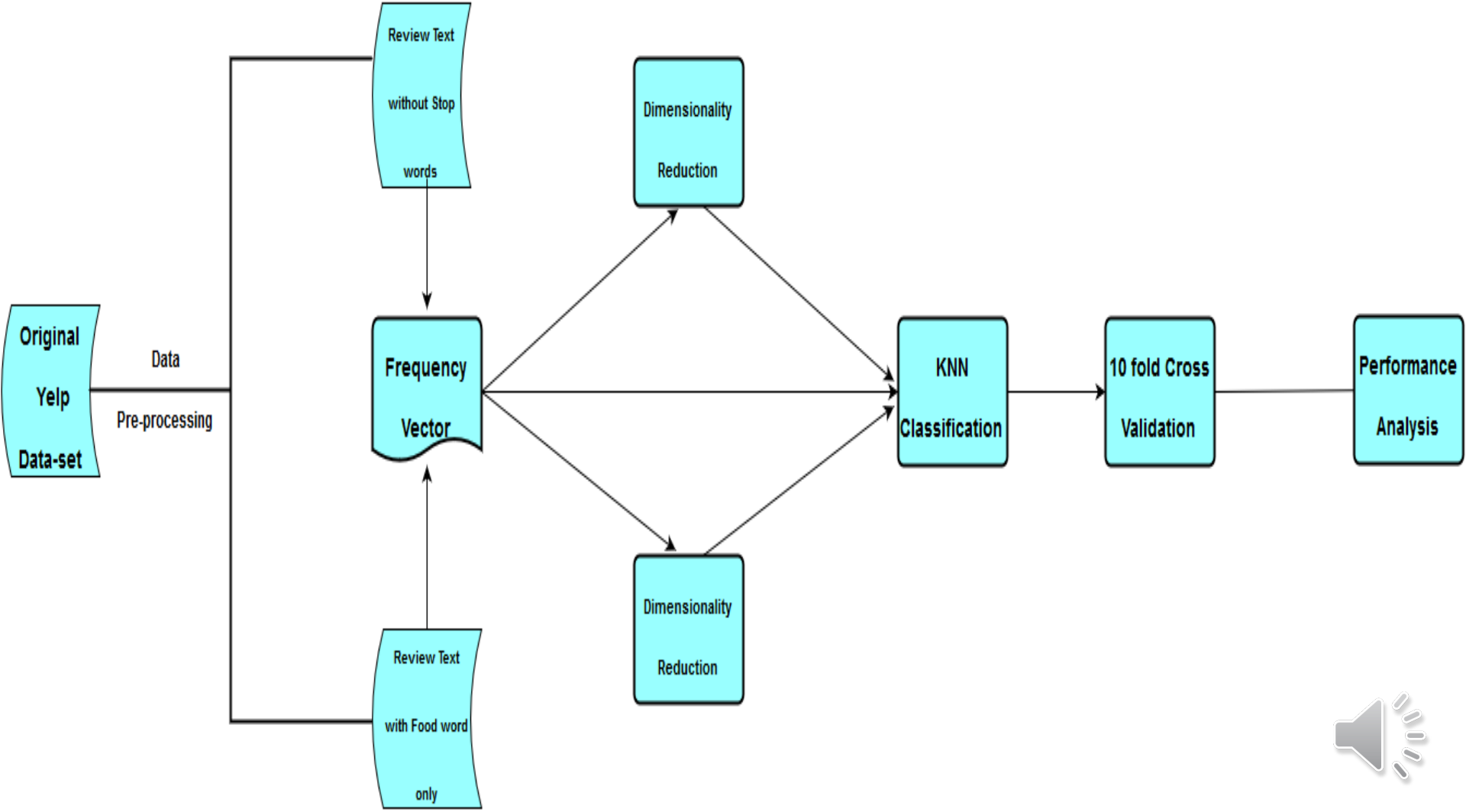
Methodology

We created four different datasets:

1. Term Frequency Vector of all reviews without stop words, common english words
2. Term Frequency Vector of all reviews with only food related words
3. Dimensionally Reduced dataset from Step 1
4. Dimensionally Reduced dataset from Step 2

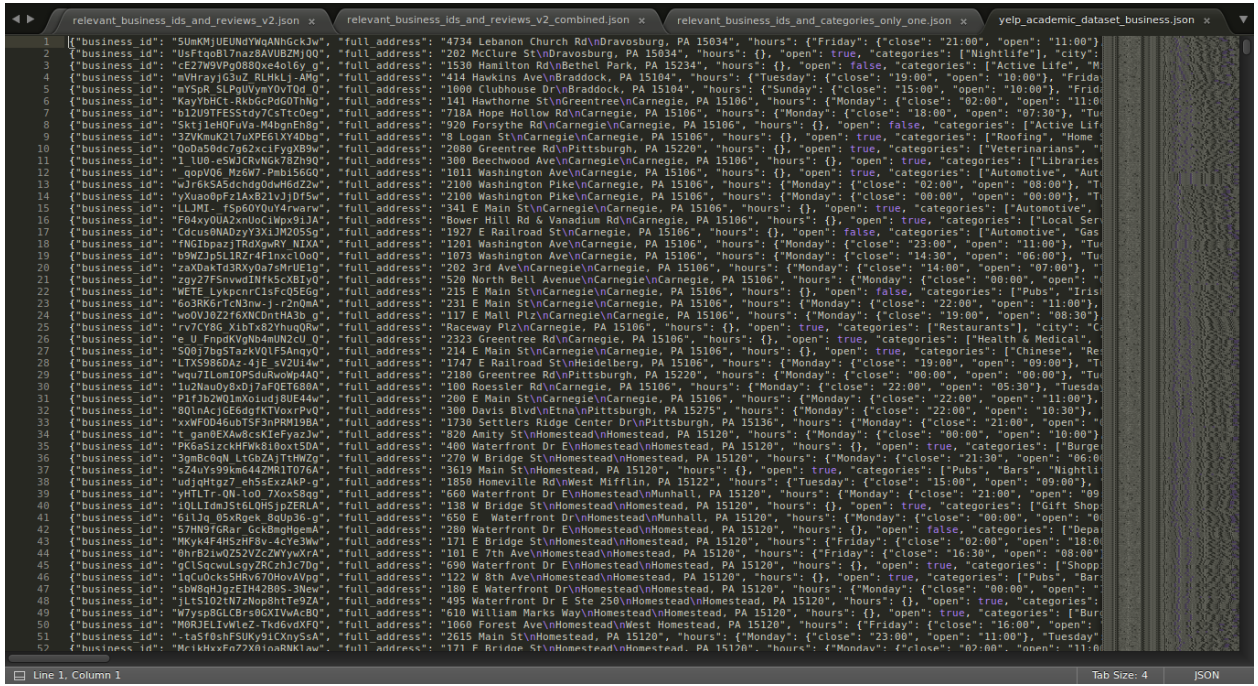
Build classification model using training dataset and test it using test dataset

Perform 10 fold cross validation to validate the model


Review Text
without Stop
words
Dimensionality
Reduction
Original
Yelp
Data-set
Data
Pre-processing
Frequency
Vector
KNN
Classification
10 fold Cross
Validation
Performance
Analysis
Dimensionality
Reduction
Review Text
with Food word
only

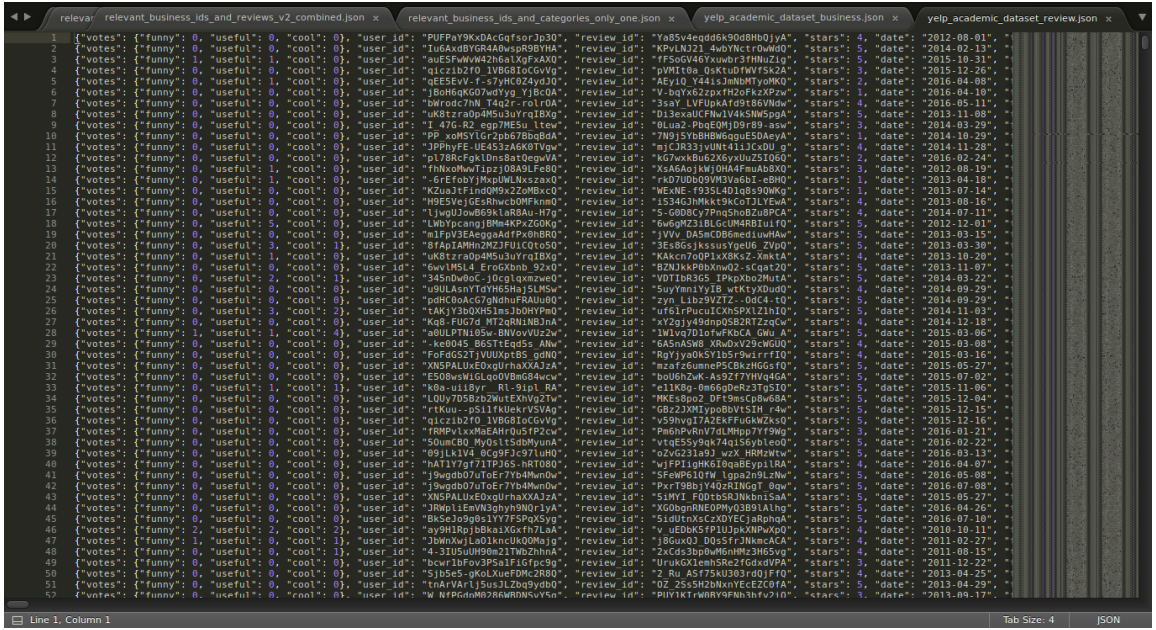
Data Pre-processing

Raw input data – Business related data



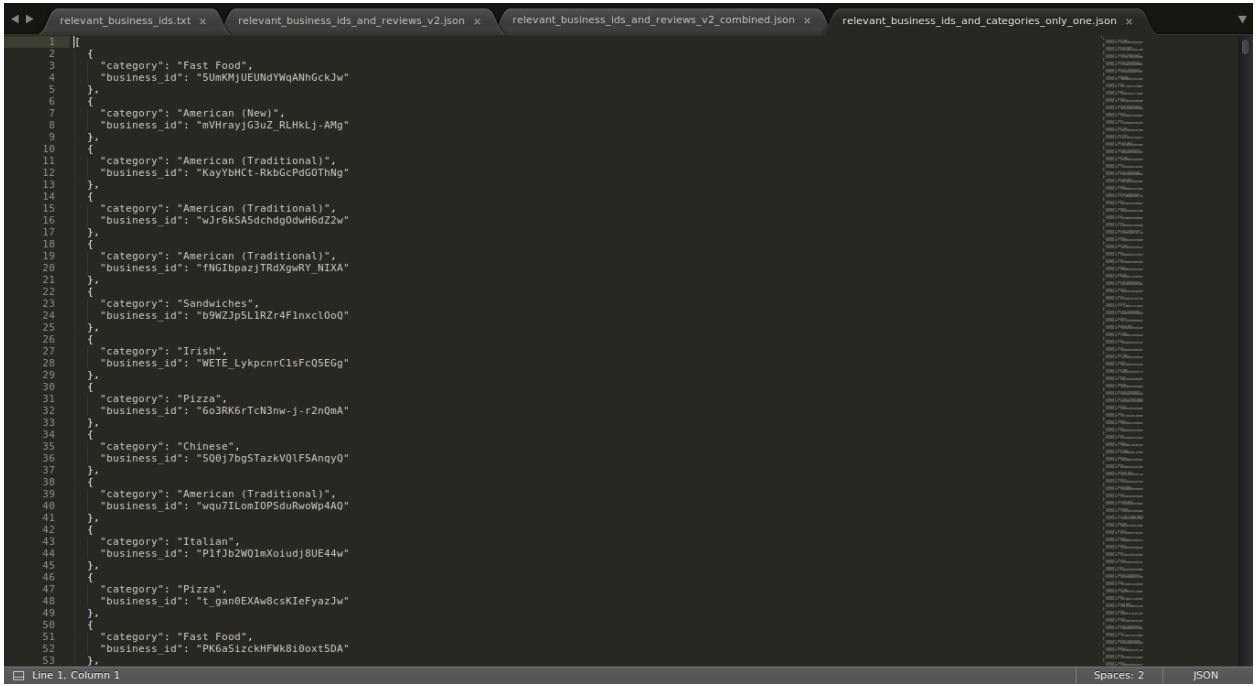
Data Pre-processing

Raw input data – Review related data



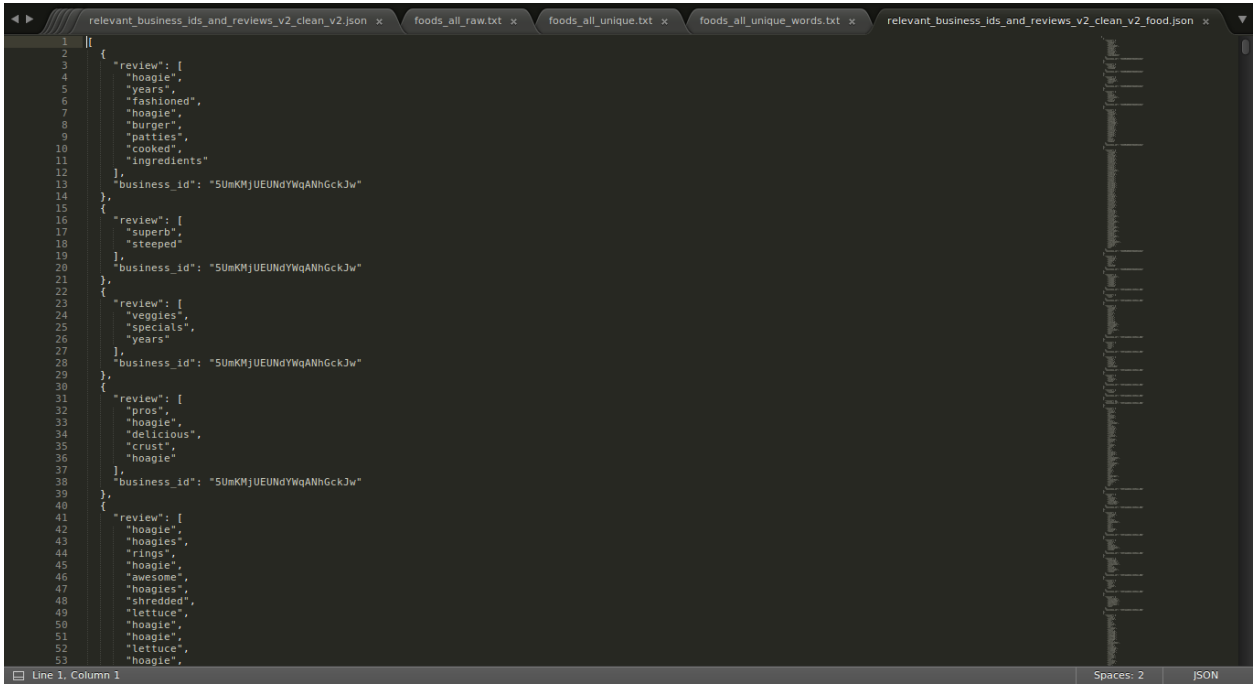
Data Pre-processing

Extract Single Category



Data Pre-processing

Extract only Food related words

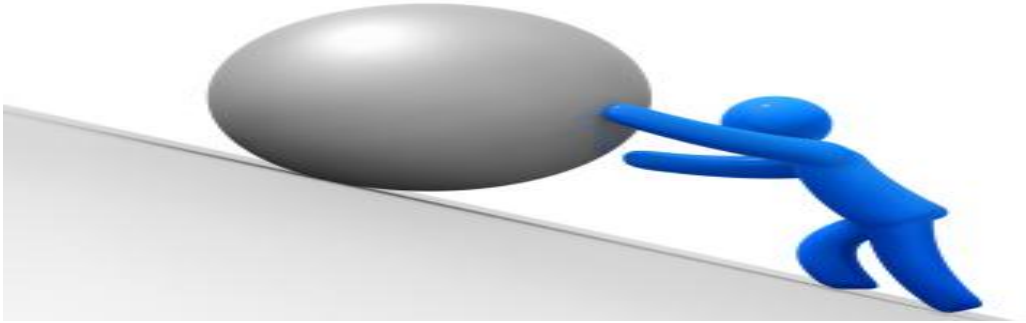


Data Pre-processing

Prepare Bag Of Words

```
4  Italian sherbert fab chic spaghetti meatballs picky spumoni spumoni chic rustic spumoni flavorful tuscan
   satisfying spumoni spumoni fabulous fabulous macaroni spaghetti years circa tastes meatballs balsamic balsamic
   years spaghetti spaghetti years tomatoes tomatoes breadsticks spaghetti spaghetti pirates pineapple scrumptious
   spaghetti diego tastes spaghetti spaghetti spaghetti tastes spaghetti childrens spaghetti spaghetti spaghetti
   years spaghetti rave nostalgic spumoni spaghetti spaghetti spaghetti years years spaghetti meatballs spinach
   spaghetti delicious spaghetti years spaghetti spaghetti balsamic spaghetti years spaghetti meatballs spaghetti
   spumoni spaghetti spumoni years homey spaghetti meatballs spaghetti stuffed spumoni spumoni spaghetti spaghetti
   ho spinach ravioli spaghetti delicious spaghetti satisfying years delicious spumoni spaghetti spaghetti bacon
   spumoni spaghetti spaghetti spaghetti spaghetti meatballs years years minestrone minestrone requested specials
   years spaghetti grandma theyre carrots spaghetti spaghetti spumoni spaghetti spaghetti spaghetti outback
   steakhouse spaghetti spaghetti flavorful sherbert tastes spaghetti spaghetti spumoni sherbert spaghetti tastes
   spaghetti diego spaghetti spaghetti spaghetti nostalgic homey spumoni spaghetti spicy meatballs spicy flavorful
   tastes ravioli ravioli ravioli flavors spumoni spaghetti squash mock spaghetti bacon spaghetti spaghetti tastes
   spaghetti meatballs spumoni youll youll youll greens tastes breadsticks spaghetti spumoni satisfying spaghetti
   potpourri spaghetti years spaghetti meatballs creepy balsalmic spaghetti tomatoes marinated nostalgic spaghetti
   mid specials years years spaghetti spaghetti spaghetti spaghetti years spaghetti spumoni sandwiches spumoni
   spaghetti sublime steamed crisp flavors meatballs penne minestrone penne spaghetti meatballs diego spaghetti
   spaghetti ricotta delicious raggedy steamy delicious spumoni penne womens spaghetti spaghetti spumoni rustic
   stump spaghetti spinach years requested creepy ravioli spaghetti spaghetti minestrone spaghetti spaghetti
   spaghetti spaghetti years rave reviews years tastes artichoke sliders spumoni pops spaghetti spaghetti spaghetti
   delicious spaghetti spaghetti years sensational youll spaghetti spaghetti spaghetti popsicle carrots packaged
   chopped spaghetti flavorful daniel spaghetti sampler corona boiled spaghetti years spaghetti rustic spinach
   artichoke delicious years spaghetti spaghetti spaghetti satisfying spaghetti delicious sliders sliders sliders
   meatballs artichoke specials years jason bacon bacon bacon spaghetti meatballs spaghetti meatballs spaghetti
   tomatoes spaghetti meatballs spaghetti delicious flavors reviews flavorful spumoni mid spaghetti meatballs
```

Challenges

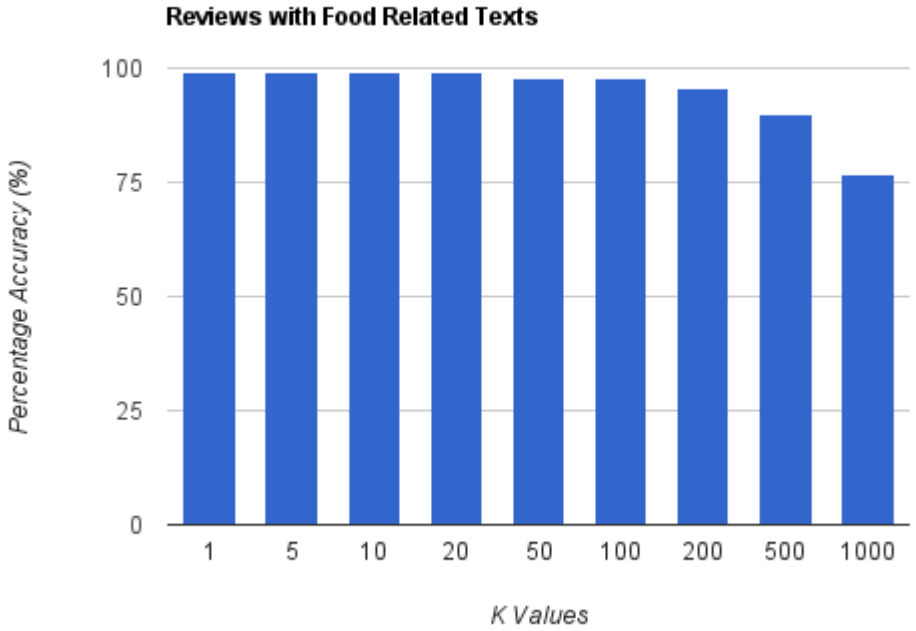


- Extracting all food items from Allrecipes.com was challenging. We had to tackle network outages, system failures while running the script.
- We kept our food scraper script continuously running for almost 3 days on 3 different system in parallel.
- Running the complete project script was time as well memory consuming. Hence we divided the tasks into smaller chunks and kept saving the results after each iteration.

Things that didn’t work

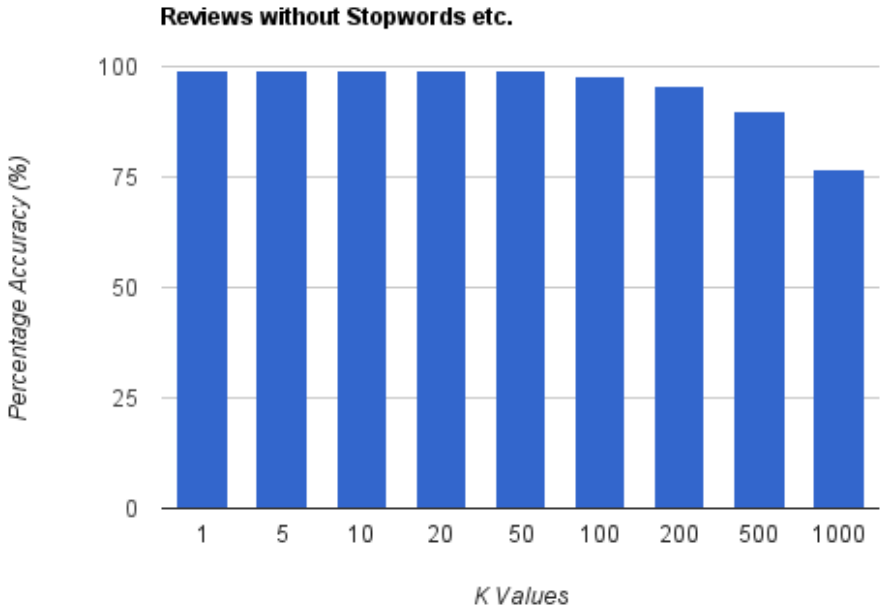
- Word2Vec - pre-trained on Google News
- Sklearn - feature_extraction.text.HashingVectorizer
- Sklearn - naive_bayes import MultinomialNB

Results



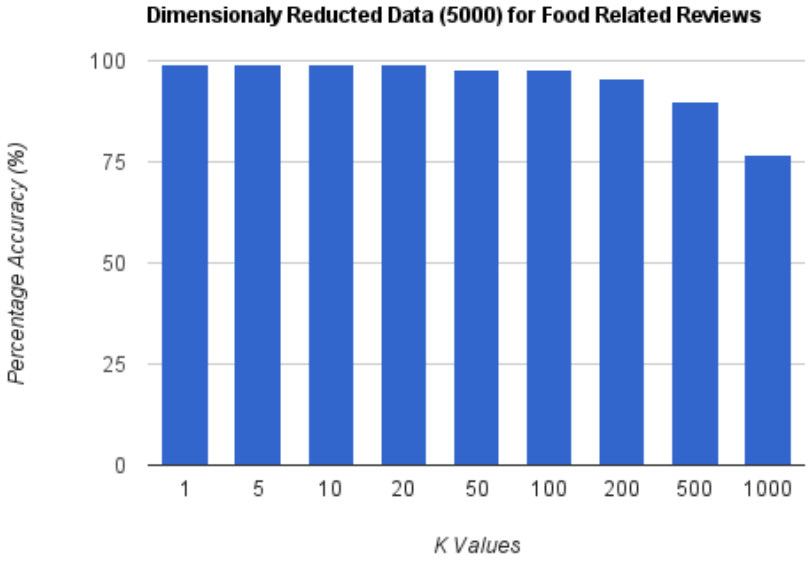
k	Mean Accuracy
1	99.99%
5	99.94%
10	99.91%
20	99.73%
50	98.92%
100	97.77%
200	95.59%
500	90.11%
1000	76.92%

Results..



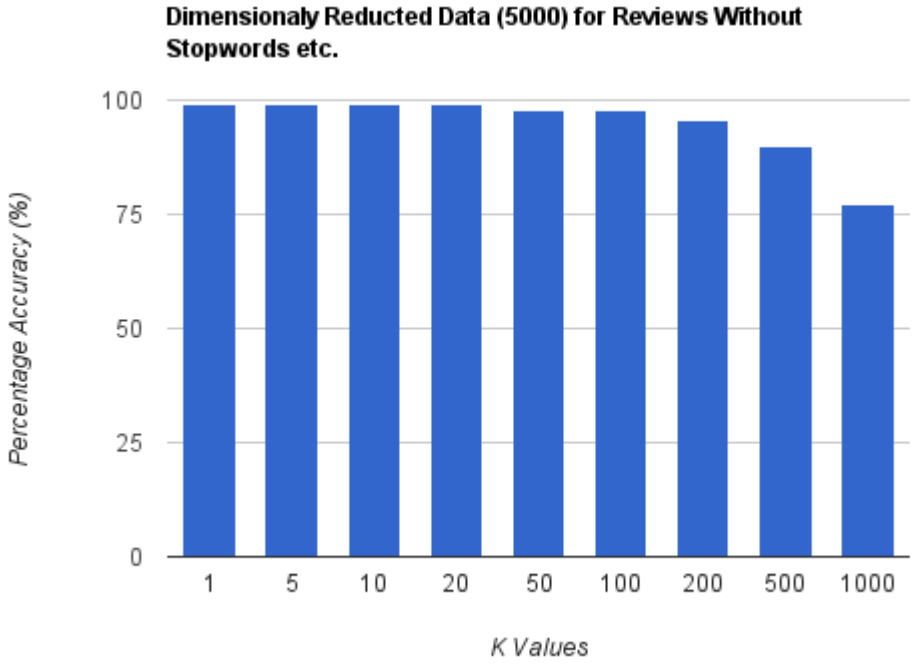
k	Mean Accuracy
1	99.99%
5	99.94%
10	99.91%
20	99.73%
50	99.00%
100	97.75%
200	95.50%
500	90.11%
1000	76.95%

Results..



k	Mean Accuracy
1	99.99%
5	99.98%
10	99.92%
20	99.75%
50	99.91%
100	97.75%
200	95.52%
500	90.11%
1000	76.78%

Results...



k	Mean Accuracy
1	99.99%
5	99.97%
10	99.91%
20	99.74%
50	99.94%
100	97.75%
200	95.53%
500	90.11%
1000	77.15%

Results - Random 1000 rows

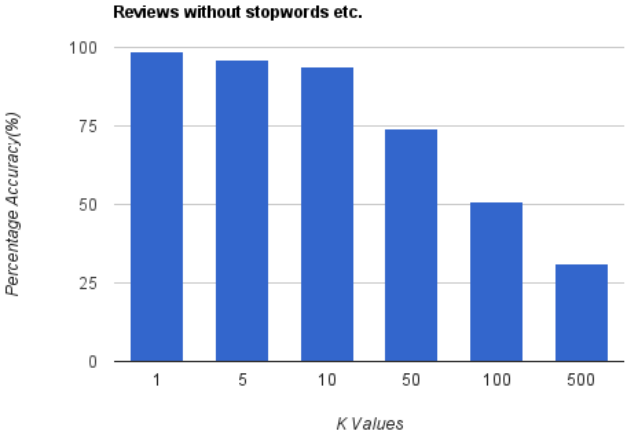
All words	
k	Mean Accuracy
1	98.90%
5	95.90%
10	93.69%
50	74.30%
100	50.90%
500	31.20%

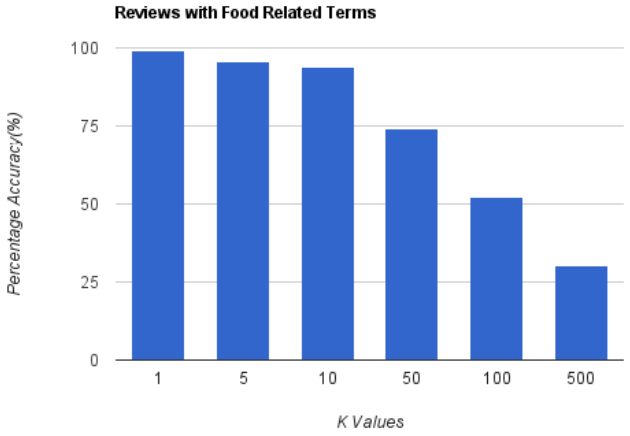
Food words	
k	Mean Accuracy
1	99.20%
5	95.80%
10	93.89%
50	74.00%
100	52.20%
500	30.40%

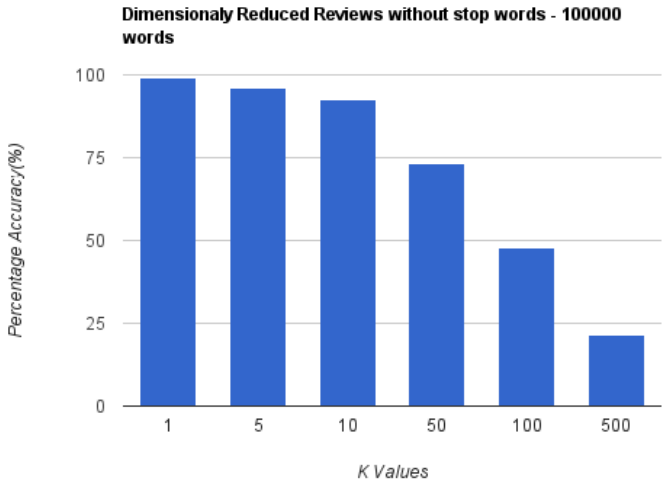
DR All words	
k	Mean Accuracy
1	99.30%
5	96.29%
10	92.69
50	73.30%
100	47.79%
500	21.70%

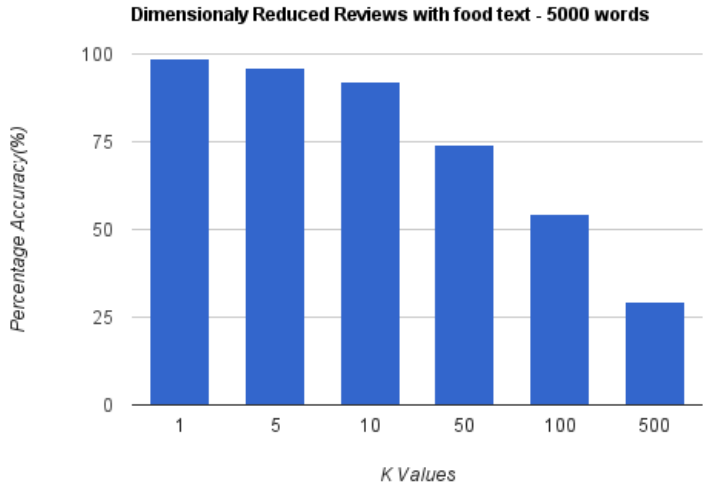
DR Food words	
k	Mean Accuracy
1	98.90%
5	96.00%
10	92.30%
50	74.20%
100	54.40%
500	29.30%

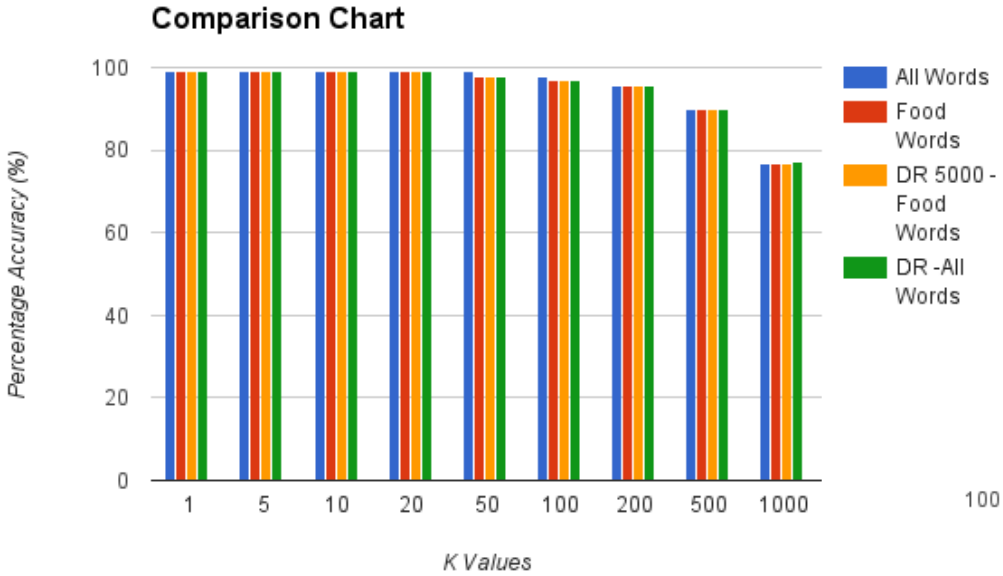
DR 2000 Food Words	
k	Mean Accuracy
1	98.50%
5	95.09%
20	91.79%
50	73.00%
100	50.30%
500	19.89%

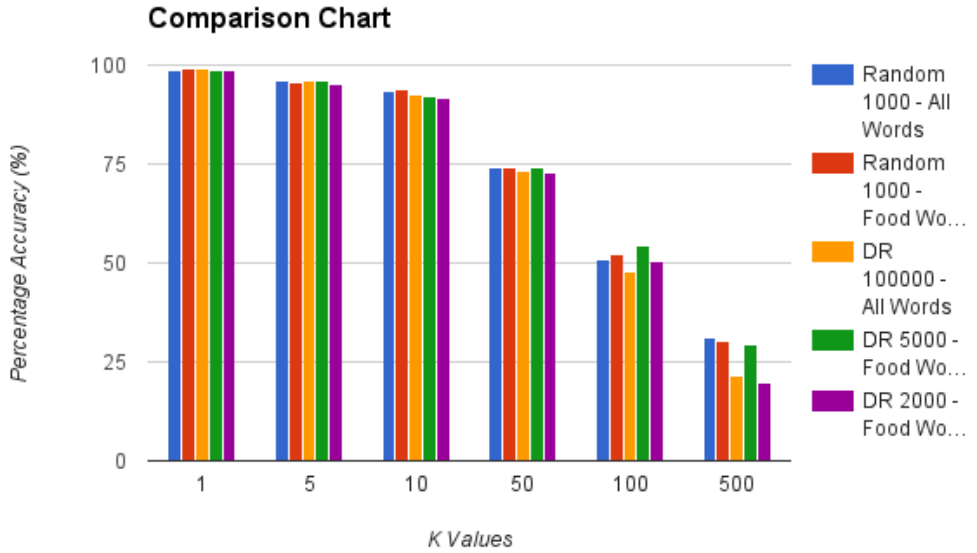

Reviews without stopwords etc.
Percentage Accuracy(%)
100
75
50
25
0
1
5
10
50
100
500
K Values


Reviews with Food Related Terms
Percentage Accuracy(%)
100
75
50
25
0
1
5
10
50
100
500
K Values


Dimensionaly Reduced Reviews without stop words - 100000 words
Percentage Accuracy(%)
100
75
50
25
0
1
5
10
50
100
500
K Values

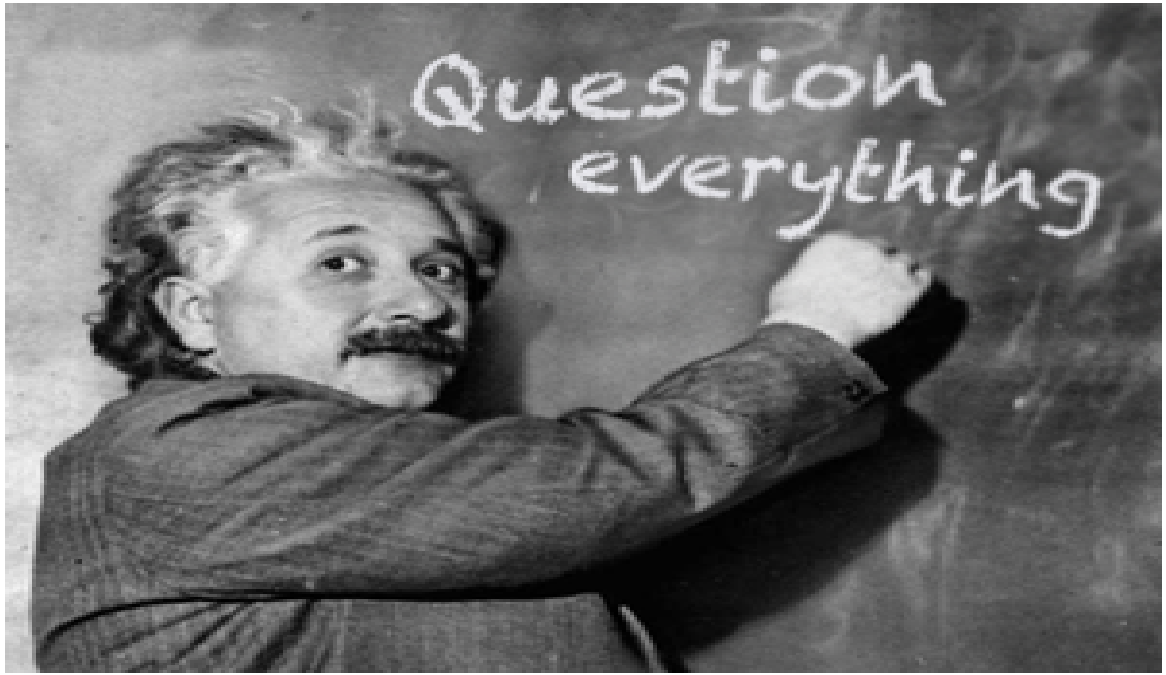

Dimensionaly Reduced Reviews with food text - 5000 words
Percentage Accuracy(%)
100
75
50
25
0
1
5
10
50
100
500
K Values


Comparison Chart
100
80
60
40
20
0
Percentage Accuracy (%)
1
5
10
20
50
100
200
500
1000
K Values
All Words
Food Words
DR 5000 - Food Words
DR -All Words


Comparison Chart
100
75
50
25
0
Percentage Accuracy (%)
1
5
10
50
100
500
K Values
Random 1000 - All Words
Random 1000 - Food Wo...
DR 100000 - All Words
DR 5000 - Food Wo...
DR 2000 - Food Wo...

Conclusion

- All the approaches performed similar with very minor differences in accuracies.
- Reducing the dimensions of the data didn’t change the accuracy of the Classification but it ran much more faster.
- As K increases, accuracy decreases for all approaches.



Questions?