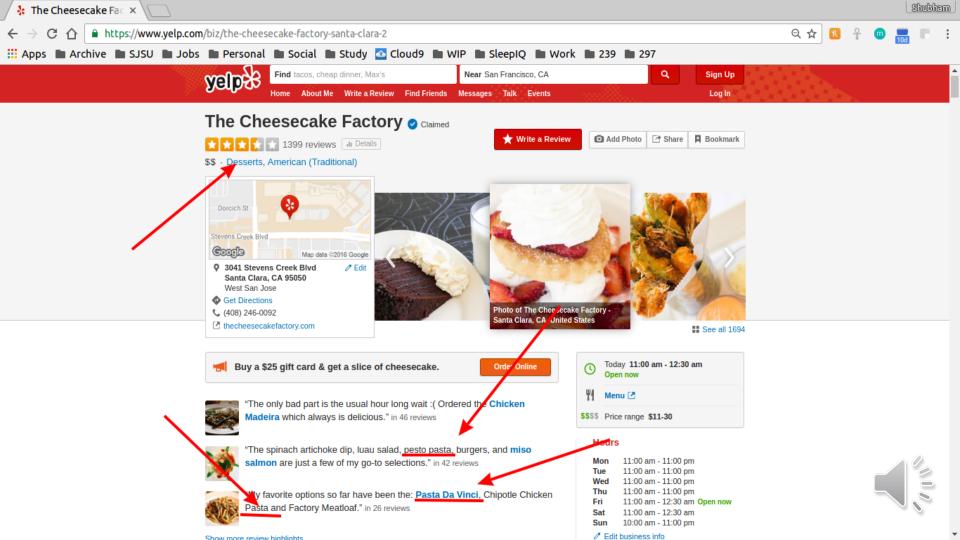
Classification of Restaurants from Customer Reviews

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





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















NumPy







Natural Language Analyses with NLTK









Methodology

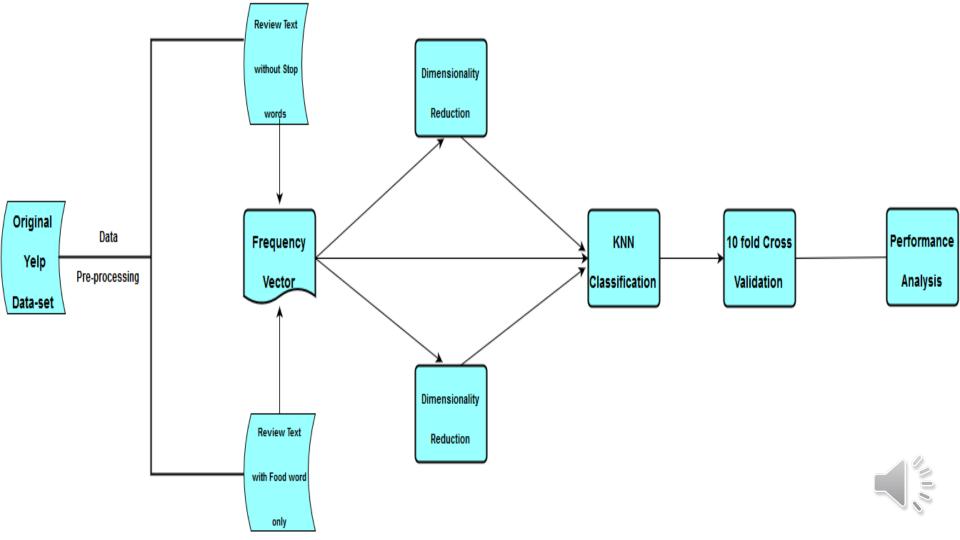
We created four different datasets:

- 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





Raw input data – Business related data

> //	relevant_business_ids_and_reviews_v2.json ×	relevant_business	_ids_and_reviews_v2_combined.json x velevant_business_ids_and_categories_only_one.json x velp_academic_dataset_business.json x
1	{"business id": "5UmKMjUEUNdYWqANhGckJw",	"full address":	"4734 Lebanon Church Rd\nDravosburg, PA 15034", "hours": {"Friday": {"close": "21:00", "open": "11:00"}
	Tusiness id": "UsFtqoBl7naz8AVUBZMjQQ",	"full address":	"202 McClure St\nDravosburg, PA 15034", "hours": {}, "open": true, "categories": ["Nightlife"], "city":
	{"business id": "cE27W9VPq0880xe4ol6y g",	"full address":	"1530 Hamilton Rd\nBethel Park, PA 15234", "hours": {}, "open": false, "categories": ["Active Life", "M
	{"business_id": "mVHrayjG3uZ_RLHkLj-AMg",	"full address":	"414 Hawkins Ave\nBraddock, PA 15104", "hours": {"Tuesday": {"close": "19:00", "open": "10:00"}, "Frida
			"1000 Clubhouse Dr\nBraddock, PA 15104", "hours": {"Sunday": {"close": "15:00", "open": "10:00"}, "Frid
	{"business_id": "KayYbHCt-RkbGcPdG0ThNg",	"full address":	"141 Hawthorne St\nGreentree\nCarnegie, PA 15106", "hours": {"Monday": {"close": "02:00", "open": "11:0
	{"business_id": "b12U9TFESStdy7CsTtcOeg",	"full_address":	"718A Hope Hollow Rd\nCarnegie, PA 15106", "hours": {"Monday": {"close": "18:00", "open": "07:30"}, "Tu
			"920 Forsythe Rd\nCarnegie\nCarnegie, PA 15106", "hours": {}, "open": false, "categories": ["Active Lif
			"8 Logan St\nCarnegie\nCarnegie, PA 15106", "hours": {}, "open": true, "categories": ["Roofing", "Home
			"2080 Greentree Rd\nPittsburgh, PA 15220", "hours": {}, "open": true, "categories": ["Veterinarians", "
			"300 Beechwood Ave\nCarnegie\nCarnegie, PA 15106", "hours": {}, "open": true, "categories": ["Libraries
	{"business_id": "_qopVQ6_Mz6W7-Pmbi56GQ",	"full_address":	"1011 Washington Ave\nCarnegie, PA 15106", "hours": {}, "open": true, "categories": ["Automotive", "Aut
	{"business_id": "wJr6kSA5dchdg0dwH6dZ2w",	"full address":	"2100 Washington Pike\nCarnegie, PA 15106", "hours": {"Monday": {"close": "02:00", "open": "08:00"}, "T
			"2100 Washington Pike\nCarnegie, PA 15106", "hours": {"Monday": {"close": "00:00", "open": "00:00"}, "T
			"341 E Main St\nCarnegie\nCarnegie, PA 15106", "hours": {}, "open": true, "categories": ["Automotive",
			"Bower Hill Rd & Vanadium Rd\nCarnegie, PA 15106", "hours": {}, "open": true, "categories": ["Local Ser
			"1927 E Railroad St\nCarnegie, PA 15106", "hours": {}, "open": false, "categories": ["Automotive", "Gas
			"1201 Washington Ave\nCarnegie, PA 15106", "hours": {"Monday": {"close": "23:00", "open": "11:00"}, "Tu
			"1073 Washington Ave\nCarnegie, PA 15106", "hours": {"Monday": {"close": "14:30", "open": "06:00"}, "Tu
			"202 3rd Ave\nCarnegie\nCarnegie, PA 15106", "hours": {"Monday": {"close": "14:00", "open": "07:00"}, "
			"520 North Bell Avenue\nCarnegie\nCarnegie, PA 15106", "hours": {"Monday": {"close": "00:00", "open": "
			"215 E Main St\nCarnegie\nCarnegie, PA 15106", "hours": {}, "open": false, "categories": ["Pubs", "Iris
			"231 E Main St\nCarnegie\nCarnegie, PA 15106", "hours": {"Monday": {"close": "22:00", "open": "11:00"},
			"117 E Mall Plz\nCarnegie\nCarnegie, PA 15106", "hours": {"Monday": {"close": "19:00", "open": "08:30"}
			"Raceway Plz\nCarnegie, PA 15106", "hours": {}, "open": true, "categories": ["Restaurants"], "city": "C
			"2323 Greentree Rd\nCarnegie, PA 15106", "hours": {}, "open": true, "categories": ["Health & Medical",
			"214 E Main St\nCarnegie\nCarnegie, PA 15106", "hours": {}, "open": true, "categories": ["Chinese", "Re
			"1747 E Railroad St\nHeidelberg, PA 15106", "hours": {"Monday": {"close": "19:00", "open": "09:00"}, "T
			"2180 Greentree Rd\nPittsburgh, PA 15220", "hours": {"Monday": {"close": "00:00", "open": "00:00"}, "Tu
			"100 Roessler Rd\nCarnegie, PA 15106", "hours": {"Monday": {"close": "22:00", "open": "05:30"}, "Tuesda
			"200 E Main St\nCarnegie\nCarnegie, PA 15106", "hours": {"Monday": {"close": "22:00", "open": "11:00"},
			"300 Davis Blvd\nEtna\nPittsburgh, PA 15275", "hours": {"Monday": {"close": "22:00", "open": "10:30"},
			"1730 Settlers Ridge Center Dr\nPittsburgh, PA 15136", "hours": {"Monday": {"close": "21:00", "open": "()
			"820 Amity St\nHomestead\nHomestead, PA 15120", "hours": {"Monday": {"close": "00:00", "open": "10:00"}
			"400 Waterfront Dr E\nHomestead\nHomestead, PA 15120", "hours": {}, "open": true, "categories": ["Burge
36			"270 W Bridge St\nHomestead\nHomestead, PA 15120", "hours": {"Monday": {"close": "21:30", "open": "86:0
			"3619 Main St\nHomestead, PA 15120", "hours": {}, "open": true, "categories": ["Pubs", "Bars", "Nightli
38			"1850 Homeville Rd\nWest Mifflin, PA 15122", "hours": {"Tuesday": {"close": "15:00", "open": "09:00"},
			"660 Waterfront Dr E\nHomestead\nMunhall, PA 15120", "hours": {"Monday": {"close": "21:00", "ppen": "09
40			"138 W Bridge St\nHomestead\nHomestead, PA 15120", "hours": {}, "open": true, "categories": ["Gift Shop
			"650 E Waterfront Dr\nHomestead\nMunhall, PA 15120", "hours": {"Monday": {"close": "00:00", "open": "0
			"280 Waterfront Dr E\nHomestead\nHomestead, PA 15120", "hours": {}, "open": false, "categories": ["Depail
			"I71 E Bridge St\nHomestead\nHomestead, PA 15120", "hours": {"Friday": {"close": "02:00", "open": "18:0
44			"101 E 7th Ave\nHomestead\nHomestead\nHomestead, PA 15120", "hours": {"Friday": {"close": "16:30", "open": "08:00"
45			"690 Waterfront Dr Einhomestead/Homestead, PA 15120", "hours: {}, "open": true, "categories": ["Shopp:
46			"122 W 8th Ave\nHomestead\nHomest
47			"180 E Waterfront Dr\nHomestead\Undomestead\PA 15120", "hours": {"Monday": {"close": "00:00", "open": "
48			"495 Waterfront Dr E Ste 250\nHomestead\Homestead, PA 15120", "hours": {}, "open": true, "categories":
49			"610 William Marks Way\Homestead\nHomestead, PA 15120", "hours": {}, "open": true, "categories": ["Bur:
50			"1969 Forest Ave\nhomestead\nwest Homestead, PA 15120", "hours": {"riday": {"close": "16:00", "open":
			"2615 Main St\nHomestead, PA 15120", "hours": {"Monday": ("close": "23:00", "open": "11:00"), "Tuesday" "171 F Ridder St\nHomestead\
	TOUSTHESS IG": "MC1KHXXFG//X016ARNKIAW".	inti address":	

Raw input data – Review related data

Type	re		ant_business_id	s_and_reviews_		on × re	elevant_business_ids_and_categ		on x yelp_academic_dat	taset_business.json ×	yelp_academic	_dataset_revie	
Total Content Total Conten	″ ₁	f"votes":	{"funny": 0.	"useful": 0.	"cool": 01.	"user id":	"PUFPaY9KxDAcGafsorJp30".	"review id":	"Ya85v4eqdd6k90d8Hb0jvA"	, "stars": 4, "date"	"2012-08-01".	- (0.001001311111	MARKE SELECT
4 Tyotes Tummy 6 Userluis 1 Cocol 0 Userluis Coc													253 II S
Transport Tran		{"votes":	{"funny": 1,			"user id":	"auESFwWvW42h6alXgFxAXQ",	"review id":	"fFSoGV46Yxuwbr3fHNuZig"	, "stars": 5, "date":	"2015-10-31",	-611111	
Cyotes Timmy 0. *useriul 0. *coll 0. *user id 1 Selfenkoll 7 Selfen		{"votes":	{"funny": 0,	"useful": 0,			"qiczib2f0 1VBG8IoCGvVg",	"review id":	"pVMIt0a QsKtuDfWVfSk2A"	, "stars": 3, "date":	"2015-12-26",	- GENERAL	
Transprint Control C		{"votes":	{"funny": 0,	"useful": 1,			"qEE5EvV-1-s7yHC0Z4ydJQ",	"review id":	"AEyiQ Y44isJmNbMTyoMKQ"	, "stars": 2, "date":	"2016-04-08",	-6111111	(60X
		{"votes":											200 H S
Votes												• EEEEEEE	
Total		{"votes":										• 6 8 8 8 8 8	
Transparence Trumny 0		{"votes":										•	
Cyotes Cyunny 0			{"funny": 0,	"useful": 0,	"cool": θ},	"user_id":	"PP_xoMSYlGr2pb67BbqBdA",	"review_id":	"7N9j5YbBHBW6qguE5DAeyA"		"2014-10-29",	•	
Transport Trumpy 0,												**************************************	
Transprint Tra													
Tumpy												* THE HILL	
													100 H ST
Tumpy													655 H S
Tunny													
Contest Cont													
Transport Trumy 0,													531 I F
22 ("votes" ("tunny" 6, "useful" 1, "cool" 1), "user_id" 3-Shibbool 29.00 "review_id" "VOITBASS (PhybloDMIAL" 1, "tars" 5, "date" 2013-11-07", "2017-01-01-01-01-01-01-01-01-01-01-01-01-01-													
Transport Tunny 6													
Transparence Trumny 6													
Transparent Trumpy 6													
Cotes Tunny 0, useful 1, cot 2, user d 1, kky 390MySlas]BOHYPRO' review id useful 1, cot 2, user d 1, ksy 1, cot 2, user d													
Transparence Tumny 6 Serial 9 Control 9 Serial 1 Control 9 Serial 1 Control 9 Serial 1 Control 9 Serial 1													
Procest Company 1.													
Procest of Control o													
("votes: ["tumny" 6, "useful": 0, "cool": 0], "user_id": "Faf6GSJT)VULKyRES gdUO. "review_id": "RaffylyookSYlbSrbutFTIO", "stars: 4, "date": 2015-03-16", "] ("votes: ["tumny" 6, "useful": 0, "cool": 0], "user_id": "MSSPALUEDGWJTHOWSDAYA". "review_id": "PasfSummerSCBscH65GFO," "stars: 5, "date": 2015-03-16", "] ("votes: ["tumny" 6, "useful": 0, "cool": 0], "user_id": "ESGOSSHIGGOVENGSAWCA". "review_id": "Bolinchak-RSSZTTHMEGAS, "stars: 5, "date": 2015-03-02", "] ("votes: ["tumny" 6, "useful": 0, "cool": 0], "user_id": "Louely "User_												.255111	
Transport Trumpy 0, "usefult 0, "cool" 0), "user ind "Edoswithicapowers ("rumpy 0, "usefult 0," cool" 0), "user ind "Edoswithicapowers ("rumpy 0," usefult 0," cool" 0), "user ind "Edoswithicapowers ("rumpy 0," usefult 0," cool" 0), "user ind "Edoswithicapowers 0," cool 0," co													
Transport Trumpy 6												-2000	
Transport Trumy 0, "usefult 1, "cool" 1), "user 31 140 150 1												-22211	
4 (votes: ['tunny' 0, 'useful': 0, 'cool': 0], 'user_id': 'LUDY/DSB2/20/UEN/V27W', 'review_id': 'MECSB0/20 PF:09sc08688*, 'stars: 5, 'date': '2015-12-64', 'date': '2015-12-64', 'date': 'date												-200	
1												-2222	
36 {"votes"; ("runny"; 0, "useful"; 0, "cool"; 0), "user_id": "qiczib2fo 1W86810C6VV0, "review_id": "V95Nvq17A2EKFFGAKYX.00", "stars"; 5, "date": "2015-12-16", " 37 {"votes"; ("runny"; 0, "useful"; 0, "cool"; 0), "user_id": "RPMPNLXXBEAHTUGHEPPTY0MP," "stars"; 3, "date": "2016-02-21", " 38 {"votes"; ("runny"; 0, "useful"; 0, "cool"; 0), "user_id": "SplitVid GCQ975VLH0", "review_id": "DvatESSyVqkYAq156VHeO0", "stars"; 5, "date": "2016-02-22", " 40 {"votes"; ("runny"; 0, "useful"; 0, "cool"; 0), "user_id": "SplitVid GCQ975VLH0", "review_id": "DvatESSyVqkYAq156VHeO0", "stars"; 5, "date": "2016-03-13", " 41 {"votes"; ("runny"; 0, "useful"; 0, "cool"; 0), "user_id": "SplitVid GCQ975VLH0", "review_id": "SplitVid GQ975VLH0", "stars": 5, "date": "2016-05-08", " 40 {"votes"; ("runny"; 0, "useful"; 0, "cool": 0), "user_id": "SplitVid GQ975VLH0", "review_id": "SplitVid GQ975VLH0", "stars": 5, "date": "2016-07-08", " 41 {"votes"; ("runny"; 0, "useful"; 0, "cool": 0), "user_id": "SplitVid GQ975VLH0", "review_id": "SplitVid GQ975VLH0", "stars": 5, "date": "2016-07-08", " 42 {"votes"; ("runny"; 0, "useful"; 0, "cool": 0), "user_id": "SplitVid GQ975VLH0", "review_id": "SplitVid GQ975VLH0", "stars": 5, "date": "2016-07-08", " 43 {"votes"; ("runny"; 0, "useful"; 0, "cool": 0), "user_id": "SplitVid GQ975VLH0", "review_id": "SplitVid GQ975VLH0", "stars": 5, "date": "2016-07-08", " 44 {"votes"; ("runny"; 0, "useful"; 0, "cool": 0), "user_id": "SplitVid GQ975VLH0", "review_id": "SplitVid GQ975VLH0", "stars": 5, "date": "2016-07-08", " 45 {"votes"; ("runny"; 0, "useful"; 0, "cool": 0), "user_id": "SplitVid GQ975VLH0", "review_id": "SplitVid GQ975VLH0", "stars": 5, "date": "2016-07-08", " 46 {"votes"; ("runny"; 0, "useful"; 0, "cool": 0), "user_id": "SplitVid GQ975VLH0", "review_id": "SplitVid GQ975VLH0", "stars": 5, "date": "20												- HEEFER	
37 (*votes: ['tunny: 6, "useful: 0, "cool: 0], "user_id: "FRPVLxMSEAHTQUSTEZOU", "review_id: "PROPYGNYTONG", "stars: 3, "date: "2016-01-21", " 38 ("votes: ['tunny: 6, "useful: 0, "cool: 0], "user_id: "SounGO Mybqls StarbyNynA", "review_id: "voteStyPyA-YalsSyNleO", "stars: 5, "date: "2016-02-22", " 39 ("votes: ['tunny: 6, "useful: 0, "cool: 0], "user_id: "SounGO Mybqls StarbyNynA", "review_id: "Soung StarbyNynA", "stars: 5, "date: "2016-02-22", " 40 ("votes: ['tunny: 6, "useful: 0, "cool: 0], "user_id: "SpageDOT/IGET/YNHAMON", "review_id: "SFGMYSIGHTQUBLONGUB												- 22 2 2 2 2 2	
38 ("votes": ("tunny": 0, "user[ul": 0, "cool: 0], "user_id": "SQUIREOD MyQGALSOMYUNA", "review_id": "votesSydycAqLSSSyDiecO", "stars: 5, "date": "2016-02-22", 40 ("votes": ("tunny": 0, "user[ul": 0, "cool: 0], "user_id": "SQUIREOD MyGGALSOMYUNA", "review_id": "votesSydycAqLSSSyDiecO", "stars: 5, "date": "2016-03-13", 41 ("votes": ("tunny": 0, "user[ul": 0, "cool: 0], "user_id": "SQUIREOD MyGGALSOMYUNA", "stars: 1, "date": "2016-08-13", 41 ("votes": ("tunny": 0, "user[ul": 0, "cool: 0], "user_id": "SQUIREOD MyGGALSOMYUNA", "stars: 1, "date": "2016-08-08", 42 ("votes: "("tunny": 0, "user[ul": 0, "cool: 0], "user_id": "SQUIREOD MyGGALSOMYUNA", "stars: 1, "date": "2016-08-08", 43 ("votes: "("tunny": 0, "user[ul": 0, "cool: 0], "user_id": "SQUIREOD MyGGALSOMYUNA", "review_id": "SQUIREOD MyGGALSOMY", "stars: 5, "date": "2016-08-25", 43 ("votes: "("unny": 0, "user[ul": 0, "cool: 0], "user_id": "SQUIREOD MyGGALSOMY", "review_id": "SQUIREOD MyGGALSOMY", "stars: 5, "date": "2016-08-25", 44 ("votes: "("unny": 0, "user[ul": 0, "cool: 0], "user_id": "SQUIREOD MyGGALSOMY", "review_id": "SQUIREOD MyGGALSOMY", "stars: 5, "date": "2016-08-25", 45 ("votes: "("unny": 0, "user[ul": 0, "cool: 0], "user_id": "SQUIREOD MyGGALSOMY", "review_id": "SQUIREON MyGGALSOM", "stars: 5, "date": "2016-08-25", 46 ("votes: "("unny": 1, "user[ul": 0, "cool: 0], "user_id": "SQUIREOD MyGGALSOMY", "review_id": "SQUIREON MyGGALSOM", "stars: 4, "date": "2016-08-25", 47 ("votes: "("unny": 0, "user[ul": 0, "cool: 0], "user_id": "SQUIREOD MyGGALSOMY", "review_id": "SQUIREON MyGGALSOM", "stars: 4, "date": "2011-08-27", 48 ("votes: "("unny": 0, "user[ul": 0, "cool: 0], "user_id": "SQUIREOD MyGGALSOM", "review_id": "SQUIREON MyGGALSOM", "stars: 4, "date": "2011-08-27", 49 ("votes: "("unny": 0, "user[ul": 0, "cool: 0], "user_id": "SQUIREOD MyGGALSOM", "review_id": "SQUIREON MyGGALSOM", "stars: 4, "date": "2011-08-27", 40 ("votes: "("unny": 0, "user[ul": 0, "cool: 0], "user_id": "SQUIREOD MyGGALSOM", "review_id": "SQUIREON MyGGALSOM", "stars: 4,												-1111111	
39 ("votes"; ("unny"; 0, "useful"; 0, "cool"; 0), "user_id"; 09]Lk1V4 &Gq#91c97(ibH0), "review_id"; "ufF97(ibK6)Gq86Epp1RAY, "stars"; 5, "date": 2016.03-13", " 40 ("votes"; ("unny"; 0, "useful"; 0, "cool"; 0), "user_id"; "hypedgoViole7/YokMnobu", "review_id"; "MF97(ibK6)Gq8Epp1RAY, "stars"; 4, "date": 2016.03-13", " 41 ("votes"; ("unny"; 0, "useful"; 0, "cool"; 0), "user_id"; "MyaqdoViole7/YokMnobu", "review_id"; "Pk77(BB)YOkQRING[0] que", "stars"; 5, "date": 2016.03-08", " 42 ("votes"; ("unny"; 0, "useful"; 0, "cool"; 0), "user_id"; "MyaqdoViole7/YokMnobu", "review_id"; "Pk77(BB)YOkQRING[0] que", "stars"; 5, "date": 2016.03-08", " 43 ("votes"; ("unny"; 0, "useful"; 0, "cool"; 0), "user_id"; "MRANDLEN/SQRINGNOLIAN, "review_id"; "XGOBQRINGHONGSAN, "stars"; 5, "date": 2016.03-08", " 44 ("votes"; ("unny"; 0, "useful"; 0, "cool"; 0), "user_id"; "MRANDLEN/SQRINGNOLIAN, "review_id"; "XGOBQRINGHONGSAN, "stars"; 5, "date": 2016.03-08", " 45 ("votes"; ("unny"; 0, "useful"; 0, "cool"; 0), "user_id"; "MRANDLEN/SQRINGNOLIAN, "review_id"; "XGOBQRINGHONGSAN, "stars"; 5, "date"; 2016.03-10", " 46 ("votes"; ("unny"; 0, "useful"; 0, "cool"; 0), "user_id"; "MRANDLEN/SQRINGNOLIAN, "review_id"; "XGOBQRINGHONGSAN, "stars"; 5, "date"; 2016.03-10", " 47 ("votes"; ("unny"; 0, "useful"; 0, "cool"; 0), "user_id"; "MRANDLEN/SQRINGNOLIAN, "review_id"; "XGOBQRINGHONGSAN, "stars"; 4, "date"; "2011-03-15", " 48 ("votes"; ("unny"; 0, "useful"; 0, "cool"; 0), "user_id"; "MranDlen/SQRINGNOLIAN, "review_id"; "XGOSQRINGNOLIAN, "stars"; 4, "date"; "2011-03-15", " 49 ("votes"; ("unny"; 0, "useful"; 0, "cool"; 0), "user_id"; "SDSG-OgoColueFPNCRSOL", "review_id"; "XGOSQRINGNOCECCOR", "stars"; 4, "date"; "2011-03-15", " 40 ("votes"; ("unny"; 0, "useful"; 0, "cool"; 0), "user_id"; "SDSG-OgoColueFPNCRSOL", "review_id"; "XGOSQRINGNOCECCOR", "stars"; 4, "date"; "2011-03-15", " 40 ("votes"; ("unny"; 0, "useful"; 0, "cool"; 0), "user_id"; "SDSG-OgoColueFPNCRSOL", "review_id"; "XGOSQRINGNOCECCOR", "stars"; 4, "date"; "2011-03-15", " 41 ("votes"; ("unny"; 0,												-45	1000
40 (*votes: [*tiumy* 6. *useful* 6. *cool* 6), "user_id*: hallY76717DiSS.NRT0807. "review_id*: "wiFPlightK6T08E8ppilRA", "stars: 4. "date": 2286-04-077. " (*votes*: [*tiumy* 6. *useful* 6. *cool* 6), "user_id* 1. *plagedDVIncF77NHMnDv*, "review_id* 1. *SepReJDVIN [paga291LN*, "stars: 5. "date": 2286-05-08, " (*votes*: [*tiumy* 6. *useful* 6. *cool* 6), "user_id* 1. *plagedDVIncF77NHMnDv*, "review_id* 1. *plagedDVIncF77NHMnDv*, "stars: 5. *date* 1. *plagedDVIncF77NHMnDv*, "review_id* 1. *plagedDVIncF77NHMnDv*, "stars: 5. *date* 1. *plagedDVIncF77NHMnDv*, "review_id* 1. *plagedDVIncF7NHMnDv*, "stars: 5. *date* 1. *plagedDVIncF7NHMnDv*, "stars: 6. *plagedD													Sept 11 87
41 ('votes': ['tunny': 0, 'userlul': 0, 'cool: 0], 'user id':]9egeb07u10fe7Y04Mendow', 'review' id': SfeMP610fW [dpa2n9LNN-, "stars: 5, "date': 2016.05.08", 42 ('votes': ['tunny': 0, 'userlul': 0, 'cool: 0], 'user id':]9egeb07u10fe7Y04Mendow', 'review' id': 'SFMP610fMR010fMg07u10fmg10fmg', "stars: 5, "date': 2016.05.08", 43 ('votes': ['tunny': 0, 'userlul': 0, 'cool: 0], 'user id': 'XMSPALLMEXQU17mAXXAJ2A', 'review' id': 'XGODgARNEWPMG0810HAN; 'stars: 5, "date': 2016.06.26", 43 ('votes': ['tunny': 0, 'userlul': 0, 'cool: 0], 'user id': 'BRSeJ03g05117YFSPGSSgg', 'review' id': 'XGODgARNEWPMG0810HAN; 'stars: 5, 'date': '2016.06.26", 44 ('votes': ['tunny': 0, 'userlul': 0, 'cool: 0], 'user id': 'BRSeJ03g05117YFSPGSSgg', 'review' id': 'XGODgARNEWPMG0810HAN; 'stars: 5, 'date': '2016.06.10", 46 ('votes': ['tunny': 0, 'userlul': 0, 'cool: 0], 'user id': 'XHINSHINGB621TMG7HDAN, 'review' id': 'XHINSHINGB621TMG7HG7HDAN, 'review' id': 'XHINSHINGB621TMG7HG7HG7HG7HG7HG7HG7HG7HG7HG7HG7HG7HG7HG													
42 ("votes": ("tunny": 0, "userlu": 9, "cool: 9), "user_id": "jwjedb0/l/lofe/YbM-molw", "review_id": "Pxr/98b)YdyñzHingf Qww", "stars": 5, "date": 22816-07-08", "date": 22816-0													SCR ST
43 ('votes': ['funny': 0, 'useful': 0, 'cool: 0), 'user id': 'MMSPALLMEONgUITHACKALLAN', 'review' id': 'SSIMIT FÜÜTLÖSAJMKENISAA', 'stars': 5, 'date': '2015-05-27', 'd' ('votes': ['funny': 0, 'useful': 0, 'cool: 0), 'user id': 'BMSPALLMEONGUITHACKALLAN', 'review' id': 'MOODPARREMPHONGUITHACKALLAN', 'stars': 5, 'date': '2016-08-10', 'd' ('votes': ['funny': 0, 'useful': 0, 'cool: 0), 'user id': 'BMSPALLMEONGUITHACKALLAN', 'review' id': 'SWITHACKALLAN', 'stars': 5, 'date': '2016-08-10', 'd' ('votes': ['funny': 1, 'useful': 0, 'cool: 1), 'user id': 'BMSPALLMEONGUITHACKALLAN', 'review' id': 'SWITHACKALLAN', 'stars': 4, 'date': '2016-08-10', 'd' ('votes': ['funny': 1, 'useful': 0, 'cool: 1), 'user id': 'DMSPALLMEONGUITHACKALLAN', 'review' id': 'BMSPALLMEONGUITHACKALLAN', 'stars': 4, 'date': '2016-08-17', 'd' ('votes': ['funny': 0, 'useful': 0, 'cool: 0), 'user id': 'DwsPALLMEONGUITHACKALLAN', 'travel': 'd': 'VunNACLMENSEZ/GAGMAP', 'stars': 4, 'date': '2016-08-17', 'd' ('votes': ['funny': 0, 'useful': 0, 'cool: 0), 'user id': 'burlettionguithaCkallan', 'travel': 'd': 'VunNACLMENSEZ/GAGMAP', 'stars': 4, 'date': '2018-08-25', 'd' ('votes': ['funny': 0, 'useful': 0, 'cool: 0), 'user id': 'burlettionguithaCkallan', 'travel': 'd': 'VunNACLMENSEZ/GAGMAP', 'stars': 4, 'date': '2018-08-25', 'd' ('votes': ['funny': 0, 'useful': 0, 'cool: 0), 'user id': 'burlettionguithaCkallan', 'votes': 'd': 'votes': 'funny': 0, 'useful': 'd': 'votes': 'votes': 'votes': 'd': 'votes': 'd': 'votes': 'd': 'votes': 'd': 'vot												-41	BE 18
44 ("votes: ["funny": 0, "useful": 0, "cool: 0], "user id": "NRPLIENNUSJHYNDNO(1)A", "review id": "XGODD;RNEDPHQG3891Alhp", "stars: 5, "date": 2016-04-26", " { "votes: ["funny": 2, "useful": 2, "cool: 2], "user id": "Skelendogs11/77F9PQASyg", "review id": "SidUtNX2CVDE(1)ARDA,", "stars: 5, "date": 2016-07-10", " 46 ["votes: ["funny": 2, "useful": 2, "cool: 2], "user id": "syPHIRpliBRAINGAFTLAA", "review id": "suDUNGSFTUNDACAN", "stars: 4, "date": 2010-10-11", " 47 ["votes: ["funny": 0, "useful": 0, "cool: 1], "user id": "ShinwAjiaOlanuChOMajg", "review id": "SROUAD 1005FT/NHACACA", "stars: 4, "date": 2011-00-2-77, " 48 ["votes: ["funny": 0, "useful": 0, "cool: 1], "user id": "SHINDFORDERSHOA", "review id": "UruxXCLBABNEGVg", "stars: 4, "date": 2011-00-2-77, " 49 ["votes: ["funny": 0, "useful": 0, "cool: 0], "user id": "box1FDox3PShifferGog", "review id": "UruxXCLBABNEGVGCMGAPM", "stars: 3, "date": 2011-12-22", " 50 ["votes: ["funny": 0, "useful": 0, "cool: 0], "user id": "ShOSE-5 vGCNLWeFDMCZRGO", "review id": "OZ ZSSSTENGWGTGCTGOTA", "stars: 4, "date": 2013-04-25", " 51 ["votes: ["funny": 0, "useful": 0, "cool: 0], "user id": "ShOSE-5 vGCNLWeFDMCZRGO", "review id": "OZ ZSSSTENGWGTGCTGOTA", "stars: 5, "date": "2013-04-25", " 51 ["votes: ["funny": 0, "useful": 0, "cool: 0], "user id": "ShOSE-5 vGCNLWeFDMCZRGO", "review id": "OZ ZSSSTENGWGTGCTGOTA", "stars: 5, "date": "2013-04-25", " 51 ["votes: ["funny": 0, "useful": "date": "2013-04-25", " 51 ["votes: ["funny": 0, "useful": "date": "2013-04-25", " 52 ["votes: ["funny": 0, "useful": "date": "2013-04-25", " 53 ["votes: ["funny": 0, "useful": "date": "2013-04-25", " 54 ["votes: ["funny": 0, "useful": "date": "2013-04-25", " 52 ["votes: ["funny": 0, "useful": "date": "2013-04-25", " 53 ["votes: ["funny": 0, "useful": "date": "2013-04-25", " 54 ["votes: ["funny": 0, "useful": "date": "2013-04-25", " 55 ["votes: ["funny": 0, "useful": "date": "2013-04-25", " 56 ["votes: ["funny": 0, "useful": "date": "2013-04-25", " 57 ["votes: ["funny": 0, "useful												-400	
45 ("votes: ["tunny": 0, "user[ul": 0, "cool": 0], "user _id": BkSchodgopl?YFSPQXSyd", "reviewid": "SidUtXx6CXXDECjARphq&", "stars: 5, "date": 2016-07-10", " 46 ("votes": ["tunny": 1, "user[ul": 0, "cool": 1]), "user _id": "DMPMRg]bksNAKAffhlaA", "review_id": "uBDMSFTDINDEXAMPA", "stars: 4, "date": 2016-10-11, " 47 ("votes": ["tunny": 1, "user[ul": 0, "cool": 1]), "user _id": "DMPMRg]bksNAKAffhlaA", "review_id": "BROWD]DQSFTDINDEXAKAF, "stars: 4, "date": 2011-02-27, " 47 ("votes": ["tunny": 0, "user[ul": 1, "user_id": 1, "user_id": "DMPMRg]bksNAKAFflaA", "review_id": "BROWD]DQSFTDINDEXAKAF, "stars: 4, "date": 2011-08-27, " 48 ("votes": ["tunny": 0, "user_id": "DMPMRg]bksNAKAFflaA", "review_id": "BROWD]DQSFTDINDEXAKAFFLAATAFFL												-4111111	
46 ("votes: ["funny": 2, "useful": 2, "cool": 2), "user id": "ayHHRpipBekaiXoxffn[Lai," review id": "u uEDMKSFNUDpixMPwQo", "stars: 4, "date": 2010-10-11," 47 ("votes: ["funny": 0, "useful": 0, "cool": 1), "user id": "DimmylaollantQoolsig", "review id": "BROUND DOSSFTNHRACKAY, "stars: 4, "date": 2011-02-277, " 48 ("votes: ["funny": 0, "useful": 0, "cool": 0), "user id": "bruffborgNEnifferGog", "review id": "ZxXdSbp0wMEnMM2HBSVg", "stars: 4, "date": 2011-08-15", 49 ("votes: ["funny": 0, "useful": 0, "cool": 0), "user id": "Sb565-GotQNLweFDMCRRGO", "review id": "UruxXCLmanReZCGGAVPM", "stars: 3, "date": 2011-12-22", 50 ("votes: ["funny": 0, "useful": 0, "cool": 0), "user id": "Sb565-GotQNLweFDMCRRGO", "review id": "DX ZSSFXBUDGYTGOJFFO", "stars: 4, "date": 2013-08-25", 51 ("votes: ["funny": 0, "useful": 0, "cool": 0), "user id": "Sh565-GotQNLweFDMCRRGO", "review id": "DX ZSSFXBUDGYTGOJFFO", "stars: 4, "date": 2013-08-25", 51 ("votes: ["funny": 0, "useful": "ater: "your jud": "tarnAfv15]SusJL2GygydoO", "review id": "DX ZSSFXBUDGYTGOJFFO", "stars: ", "stars: "S, "date": "2013-08-29", 52 ("votes: "["funny": 0, "useful": "stars: "your jud": "tarnAfv15]SusJL2GygydoO", "review id": "DX ZSSFXBUDGYTGOJFFO", "stars: ", "stars: "S, "date": "2013-08-29", 53 ("votes: "["funny": 0, "useful": "stars: "S, "date": "2013-08-29", 54 ("votes: "["funny": 0, "useful": "stars: "S, "date": "2013-08-29", 54 ("votes: "["funny": 0, "useful": "stars: "S, "date": "2013-08-29", 55 ("votes: "["funny": 0, "useful": "stars: "S, "date": "2013-08-29", 56 ("votes: "["funny": 0, "useful": "stars: "S, "date": "2013-08-29", 57 ("votes: "["funny": 0, "useful": "stars: "S, "date": "2013-08-29", 58 ("votes: "["funny": 0, "useful": "stars: "S, "date": "2013-08-29", 58 ("votes: "["funny": 0, "useful": "stars: "S, "date": "2013-08-29", 58 ("votes: "["funny": 0, "useful": "stars: "S, "date": "2013-08-29", 58 ("votes: "["funny": 0, "useful": "stars: "S, "date": "2013-08-29", 59 ("votes: "["funny": 0, "useful": "stars: "S, "date": "2013-08-2													
47 (*votes: ["funny": 1, "userlu": 0, "cool: 1), "user lu": "JöhnövjiaOlkncikQONajo", "review id: "jökux03 0035/r]NkmackA", "stars: 4, "date: "2011-02-27", " ("votes: ["funny": 0, "userlu": 0, "cool: 1), "user lu": "AllSUMH09021DD2hhA", "review id: "2xKdS3bb0040MM23H65vy, "stars: 4, "date: "2011-03-15; " ("votes: ["funny": 0, "userlu": 0, "cool: 0), "user lu": "Sib56-3goolkuerBHC5800", "review id: "2, "al. A5775xUB31701Ff0", "stars: 4, "date: "2011-04-25", " ("votes: ["funny": 0, "userlu": 0, "cool: 0), "user lu": "Sib56-3goolkuerBHC5800", "review id: "2, "al. A5775xUB31701Ff0", "stars: 4, "date: "2011-04-25", " ("votes: ["funny": 0, "userlu": "user lu": "sib56-3goolkuerBHC5800", "review id: "2, "al. A5775xUB31701Ff0", "stars: 4, "date: "2011-04-25", " ("votes: ["funny": 0, "userlu": "user lu": "sib56-3goolkuerBHC5800", "review id: "2, "al. A5775xUB31701Ff0", "stars: "4, "date: "2011-04-25", " ("votes: ["funny": 0, "userlu": "ater: "2011-04-25", " ("votes: ["funny": 0, "userlu": 0, "cool: "0), "user lu": "sib56-3goolkuerBHC5800", "review id: "2, "al. A5775xUB31701Ff0", "stars: "4, "date: "2011-04-25", " ("votes: ["funny": 0, "userlu": "ater: "2011-04-25", " ("votes: ["funny: 0, "userlu": 0, "userlu": "ater: "2011-04-25", " ("votes: ["funny: 0, "userlu": 0, "userlu": "ater: "2011-04-25", " ("votes: ["funny: 0, "userlu": 0, "userlu": "ater: "2011-04-25", " ("votes: ["funny: 0, "userlu": 0, "userlu": "ater: "2011-04-25", " ("votes: ["funny: 0, "userlu": 0, "userlu": "ater: "2011-04-25", " ("votes: ["funny: 0, "userlu": 0, "use		{"votes":										-41111111	600 H S
48 ("votes: ["funny": 0, "useful": 0, "cool": 1), "user id": "4-31U5JUH9902ITMZhhañ, "review id": "2xCd32bp04M6HMY2H65vy", "stars: 4, "date": 2011-08-15", " {"votes: ["funny": 0, "useful": 0, "cool": 0), "user id": "broxForeSportsforeS		{"votes":											900 II W
50 {"votes"; {funny": 0, "useful": 0, "cool": 0), "user_id": "Sjb5e5-pKoLXuerPMC2880", "review_id": 72 Ru_AST75KU303rd0]ff0", "stars": 4, "date": "2013-04-25", " 51 {"votes": {funny": 0, "useful": 0, "cool": 0), "user_id": "thArVArlij5bsJlZbd9ydb0", "review_id": "22_25SSHZbMxNPCEZCOFA", "stars": 5, "date": "2013-04-23", "		{"votes":										-4111111	100 P
50 {"votes"; {funny": 0, "useful": 0, "cool": 0), "user_id": "Sjb5e5-pKoLXuerPMC2880", "review_id": 72 Ru_AST75KU303rd0]ff0", "stars": 4, "date": "2013-04-25", " 51 {"votes": {funny": 0, "useful": 0, "cool": 0), "user_id": "thArVArlij5bsJlZbd9ydb0", "review_id": "22_25SSHZbMxNPCEZCOFA", "stars": 5, "date": "2013-04-23", "		{"votes":	{"funny": 0,				"bcwrlbFov3PSalFiGfpc9g",	"review id":	"UrukGX1emhSRe2fGdxdVPA"		"2011-12-22",	-65	
			{"funny": 0,				"Sjb5e5-gKoLXueFDMc2R8Q",	"review_id":	"2_Ru_ASf75kU303rdQjFfQ"			-111111	
52 {"votes": ("funnov": 0. "useful": 0. "cool": 0). "user id": "W NfPGdoM6286/RDNSvYSo". "review id": "PUYIKTrW6RY9FNh3hfv210". "stars": 3. "date": "2013-09-17". "											"2013-04-29",	* GENERAL	MESS 100
							"W NfPGdpM0286WBDNSvY5a".		"PUY1KTrW8BY9FNb3bfv2i0"			- 4 m () [] []	

Extract Single Category

```
relevant business ids.txt x relevant business ids and reviews v2.json x relevant business ids and reviews v2.combined.json x relevant business ids and categories only one.json x
       "category": "Fast Food",
"business_id": "5UmKMjUEUNdYWqANhGckJw"
       "business_id": "mVHrayjG3uZ_RLHkLj-AMg"
       "business id": "KayYbHCt-RkbGcPdGOThNg'
       "business id": "wJr6kSA5dchdg0dwH6dZ2w"
       "business id": "fNGIbpazjTRdXgwRY NIXA"
       "category": "Sandwiches",
"business_id": "b9WZJp5L1RZr4Flnxcl0oQ"
       "category": "Chinese",
"business id": "SQ0j7bgSTazkVQlF5AnqyQ"
       "business id": "wqu7ILomI0PSduRwoWp4AQ"
       "business_id": "PlfJb2WQ1mXoiudj8UE44w"
       "business_id": "t_gan0EXAw8csKIeFyazJw"
       "business id": "PK6aSizckHFWk8i0oxt5DA"
```

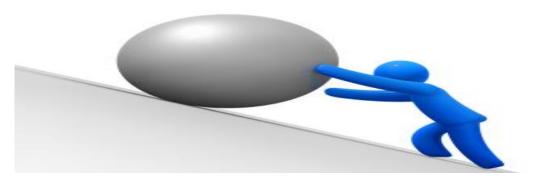
Extract only Food related words

```
relevant business ids and reviews v2 clean v2.json x foods all raw,txt x foods all unique.txt x foods all unique.txt x relevant business ids and reviews v2 clean v2 food.json x
   "business id": "5UmKMjUEUNdYWqANhGckJw"
   "business_id": "5UmKMjUEUNdYWqANhGckJw"
   "business id": "5UmKMjUEUNdYWqANhGckJw"
     "hoagie",
     "hoagie".
```

Prepare Bag Of Words

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 <u>spumoni spaghetti spaghetti</u> ho spinach ravioli spaghetti delicious spaghetti satisfying years delicious spumoni spaghetti spaghetti bacon spumoni spaghetti spaghetti spaghetti meatballs vears vears 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 voull voull greens tastes breadsticks spaghetti spumoni satisfying spaghetti potpourri spaghetti vears 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 vears 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

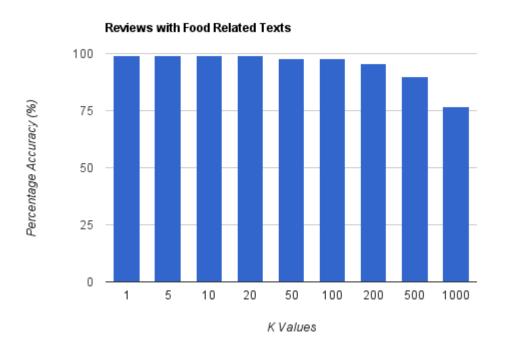


- Extracting all food items from <u>Allrecipes.com</u> was challenging. We had to tackle network outages, system failures while running the script.
- > We kept our food scraper script continuously running for almost 3 days on 3 different system in parallel.
- > Running the complete project script was 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

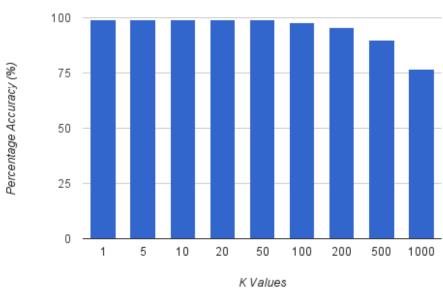


k	Mean Accuracy
1	99.99%
5	99.94%
10	99.91%
20	99.73%
50	98.92%
100	97.77%
200	95.59%
500	90.11%
1000	76.92%



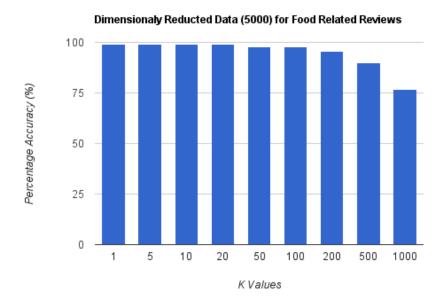
Results..





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%

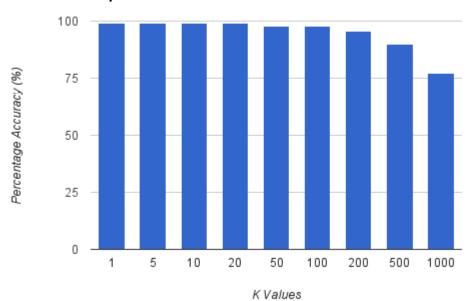
Results..



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...

Dimensionaly Reducted Data (5000) for Reviews Without Stopwords etc.



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

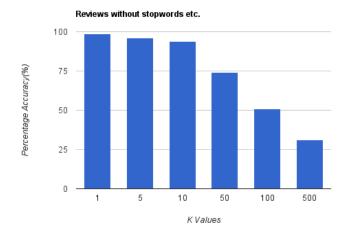
А		Foo	
k	Mean Accuracy		k
1	98.90%		1
5	95.90%		5
10	93.69%		10
50	74.30%		50
100	50.90%		100
500	31.20%		500

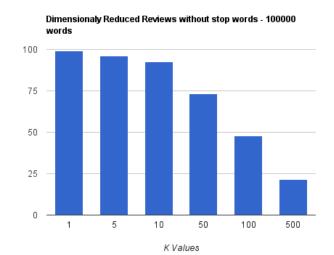
Food words				
k	Mean Accuracy			
1	99.20%			
5	95.80%			
10	93.89%			
50	74.00%			
100	52.20%			
500	30.40%			

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

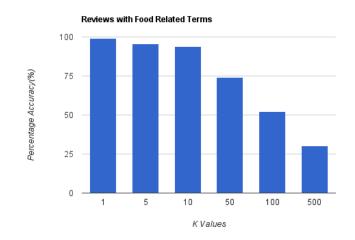
DR Food words				
k	Mean Accuracy			
1	98.90%			
5	96.00%			
10	92.30%			
50	74.20%			
100	54.40%			
500	29.30%			

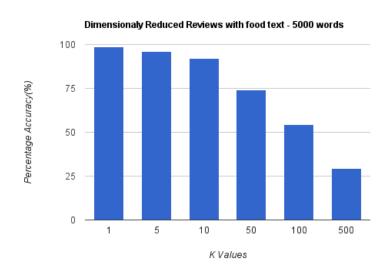
DR 2000 Food Words					
k	Mean Accuracy				
1	98.50%				
5	95.09%				
20	91.79%				
50	73.00%				
100	50.30%				
500	19.89%				





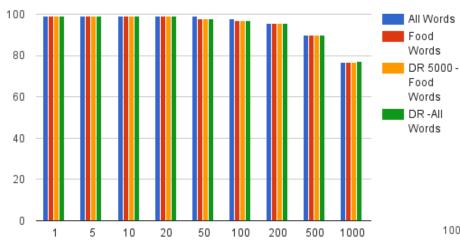
Percentage Accuracy(%)







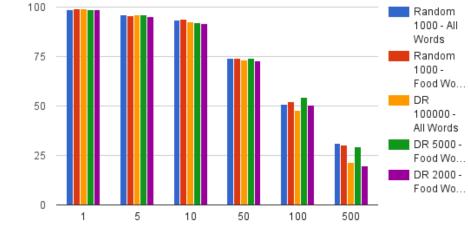
Percentage Accuracy (%)



K Values

Comparison Chart

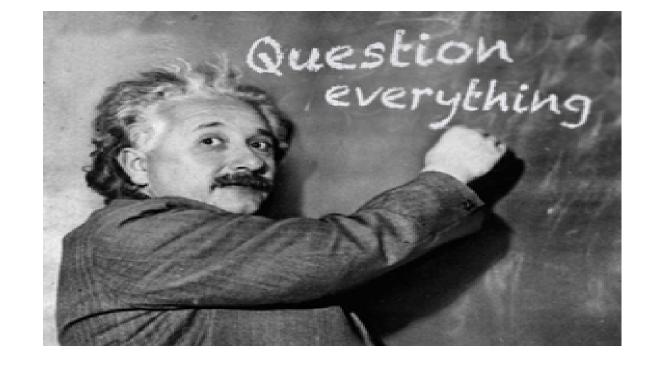
Percentage Accuracy (%)



K Values

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?