Chengjun Yuan cy3yb@virginia.edu

1.1 Codes for implementation of text normalization module are shown in CODES. 1:

```
token=token.replaceAll("\\p{Punct}+","");  // remove all punctuations.
if(!token.isEmpty()){
        token=token.toLowerCase();          // shift to lower case.
        if(token.matches(".*\\d+.*"))
                token="NUM";                // numbers to NUM.
        else{
                stemmer.setCurrent(token);
                if (stemmer.stem())
                        token=stemmer.getCurrent();
        }
}
```

CODES. 1

1.2 The relation of TTF and DF to the ranks are plotted as $\log_{10}$ scale in FIG. 1 below:

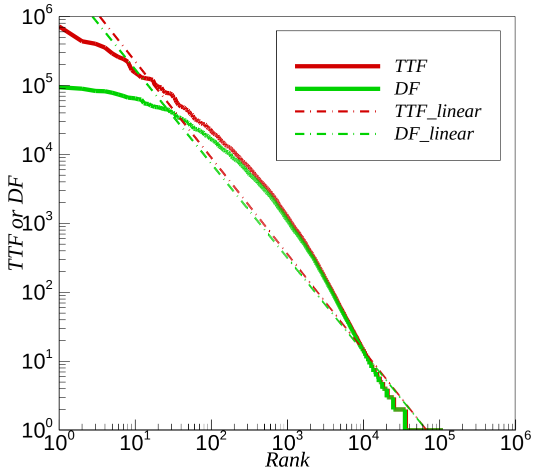

FIG. 1

Their linear regression plots are also drawn in FIG. 1. The equations and parameters are listed below:

$$log_{10}(TTF) = -1.3965 * log_{10}(rank) + 6.7446$$
$$log_{10}(DF) = -1.3667 * log_{10}(rank) + 6.5984$$

1.3 From the upper results, it seems that the rank-DF relation fits Zipf's law (where the slope is about -1) better than rank-TTF does in overall dataset. The main reason is that the volume of words' pool in our current data set is far from enough to satisfy the Zipf's distribution. In addition, in the region of rank from 1 to 100, the absolute slope of the rank-DF line is much smaller than that of rank-TTF line. That is why the overall linear regression of rank-DF relation is closer to Zipf's law than that of rank-TTF does. **However**, if we focus on the region of rank from 1 to 100, as shown in FIG. 2. Their equations and parameters are shown below:

$$log_{10}(TTF) = \ -0.89 * log_{10}(rank) + 6.11$$
$$log_{10}(DF) = \ -0.62 * log_{10}(rank) + 5.43$$

It can be seen that the TTF performs better to fit Zipf's law than the DF does in the head region of words rank.

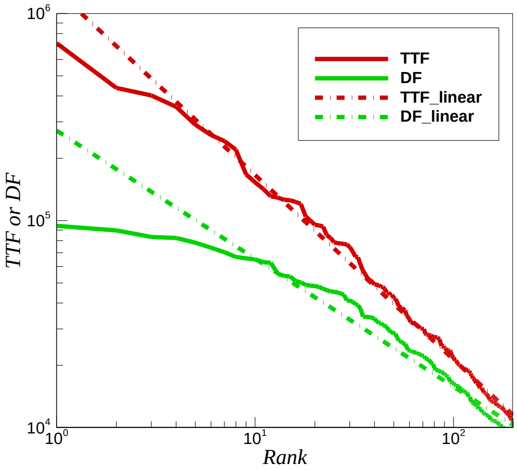

FIG. 2

2.1 The list of new stopwords specific to restaurant reviews are shown below:

NUM
nt
good
food
great
it-was
and-the
of-the
order
time
wait
this-place
it-s
servic
back
in-the
friend
the-food
love
on-the
and-i
ve
delici
do-nt
i-was
restaur
if-you
for-the
fri
for-a
eat

2.2 The size of the resulting controlled vocabulary is **34570**.

2.3 Top 50 and bottom 50 N-grams according to DF and their corresponding IDFs:

Top 50 N-grams								
Rank	**Token**	**DF**	**IDF**		**Rank**	**Token**	**DF**	**IDF**
1	tabl	18457	1.743		26	nice	14543	1.847
2	becaus	18300	1.747		27	Is-a	14323	1.853

3	make	17983	1.755		28	On-a	14240	1.856
4	i-m	17711	1.761		29	At-the	14135	1.859
5	sauc	17560	1.765		30	drink	14111	1.86
6	i-ve	17458	1.767		31	And-it	14104	1.86
7	The-best	17131	1.776		32	chees	14074	1.861
8	i-had	17113	1.776		33	i-would	13985	1.864
9	To-the	16855	1.783		34	i-have	13977	1.864
10	But-i	16483	1.792		35	Did-nt	13741	1.871
11	dish	16482	1.792		36	worth	13336	1.884
12	tast	16405	1.794		37	We- were	13196	1.889
13	littl	16126	1.802		38	meal	13075	1.893
14	definit	15904	1.808		39	flavor	12874	1.9
15	Was-a	15878	1.809		40	bar	12815	1.902
16	menu	15857	1.809		41	perfect	12707	1.905
17	With-a	15794	1.811		42	But- the	12397	1.916
18	amaz	15784	1.811		43	Had-the	12395	1.916
19	With-the	15645	1.815		44	price	12358	1.918
20	peopl	15494	1.819		45	recommend	12346	1.918
21	thing	15388	1.822		46	One-of	12294	1.92
22	pretti	14939	1.835		47	a-littl	12171	1.924
23	To-be	14876	1.837		48	But-it	12053	1.928
24	seat	14822	1.839		49	This- is	11952	1.932
25	night	14757	1.84		50	star	11922	1.933

Bottom 50 N-grams

Rank	Token	DF	IDF		Rank	Token	DF	IDF
34521	the-canter	50	3.310485		34546	onto-my	50	3.310485
34522	hate-when	50	3.310485		34547	hostess-came	50	3.310485
34523	bayless-restaur	50	3.310485		34548	to-downtown	50	3.310485
34524	made-at	50	3.310485		34549	wall-that	50	3.310485
34525	lobster-is	50	3.310485		34550	good-dessert	50	3.310485
34526	strawberri-jam	50	3.310485		34551	arriv-my	50	3.310485
34527	dinner-around	50	3.310485		34552	she-apolog	50	3.310485
34528	slider-are	50	3.310485		34553	pleas-note	50	3.310485
34529	bump-up	50	3.310485		34554	long-list	50	3.310485
34530	went-perfect	50	3.310485		34555	burger-there	50	3.310485
34531	was-curious	50	3.310485		34556	millenium-park	50	3.310485
34532	for-convers	50	3.310485		34557	took-him	50	3.310485
34533	spot-as	50	3.310485		34558	area-they	50	3.310485

34534	here-can	50	3.310485		34559	of-surpris	50	3.310485
34535	folk-i	50	3.310485		34560	light-of	50	3.310485
34536	averag-it	50	3.310485		34561	word-the	50	3.310485
34537	of-say	50	3.310485		34562	corn-here	50	3.310485
34538	beehiv-for	50	3.310485		34563	dinela	50	3.310485
34539	NUM-have	50	3.310485		34564	your-entre	50	3.310485
34540	primari	50	3.310485		34565	NUM-right	50	3.310485
34541	thirst	50	3.310485		34566	troubl-get	50	3.310485
34542	crowd-this	50	3.310485		34567	becaus-one	50	3.310485
34543	stick-and	50	3.310485		34568	way-as	50	3.310485
34544	tabl-though	50	3.310485		34569	quaint-and	50	3.310485
34545	rememb-is	50	3.310485		34570	s-spici	50	3.310485

3.1 Codes for cosine similarity computation are listed below:

```
public double cosineSimilarity(double[] reviewVector1,double[] reviewVector2){
                 double magtitude1=0.0;
                 double magtitude2=0.0;
                 double product=0.0;
                int len=reviewVector1.length;
                for(int i=0;i<len;i++){
                        magtitude1+=reviewVector1[i]*reviewVector1[i];
                        magtitude2+=reviewVector2[i]*reviewVector2[i];
                        product+=reviewVector1[i]*reviewVector2[i];
                }
                if(magtitude1==0.0||magtitude2==0.0)return 0.0;
                else return product/(Math.sqrt(magtitude1)*Math.sqrt(magtitude2));
        }
```

3.2 For the first review :

Author	Date	Content
David W.	2011-11-12	I took a group of 12 friends there based on the recommendation of a friend.... It was some of the blandest Thai food I have ever had. The Panag Nua was tough and lacking in flavor. It was a true Panag curry but it just seemed watered down. The meat had been just sliced and tossed in. Normally good Thai places have nice chunks of beef that they let stew in it all day so the sauce permeates it and the meat just falls apart. The Yum Yai salad had almost no sauce and what litte sauce it did have lacked in any flavor.The Tom Ka Guy soup was watery and heavily lacking in taste. The chicken in it was dry and huge chunks that were just not right for a soup. It also lacked a decent amount of mushrooms.The Kow Pat Moo was totally flavorless for a Thai fried rice. It was more like over priced panda express Chinese fast food rice. This is a huge failure. I mean how hard is it to toss nam plaud and sugar into some rice? The Phad Thai was soggy and had more the consistency of spagetti with marinara then a thai fried noodle dish. The saving grace was the moo sautee. It was nice and

		small like you would get on the streets of bangkok. It did taste good but then again that's the staple of Thai places and you would have to go out of your way to screw this one up.All this being said my friends liked it BUT none of them have ever tried Thai food before much less GOOD Thai food. I think this must be the case for a lot of the positive reviews here. An other negative reviewer pointed out that there simply no Thai people in there this does tend to be a good indicator as to if a Thai place is good.
Ginny M.	2011-12-22	Yes, it's expensive considering how cheap pizza is generally. However, I have never had such a perfect meal in every detail as I had at Serious Pie. The appetizer (a crab salad) was astoundingly good, loaded with lots of crab meat (not a few token shreds), on a delicious yogurt sauce. The "basic" pizza was as good as you would get in New York. I usually find pizzas to consist of passable red sauce, plus melted mozzarella on a thick piece of bread - tasty, yes, but not real pizza. At Serious Pie, however, it was cheesy, the crust thin and crispy (but not too) and the sauce was flavored with the right spices. Then we went overboard and ordered the coconut cream pie, and I have to say it was unlike any I'd ever had - with crunchy fresh flaked coconut, and a not-too-sweet creamy filling. Amazing. And I wouldn't want to forget the Cafe Americano (which came with a macaroon). It was true that neither of us could eat the next day after our over-indulgence, but it was worth it for the perfect meal at Serious Pie.

For the second review:

Author	Date	Content
Jaron H.	2014-06-15	Had the lobster and pasta. Pretty good, though the lobster was over cooked. Definitely a lot of food for the price. My friend had the butternut squash ravioli and it was quite tasty.
Kathryn B.	2011-08-13	I highly recommend the butternut squash ravioli and pumpkin tortellini. Both amazing, cooked perfectly, wish I lived in Boston so I could eat these dishes more often. Worth the 45min wait for sure.
Amy L.	2009-09-04	I definitely love this place, Giacomo's.We've waited outside in line for about an hour on Friday night but absolutely worth it :) We started with their popular calamari appetizer and

		the garlic bread then we had the butternut squash ravioli and veal Parmesan dish....The calamari appetizer was fresh and good and the garlic bread had a great flavor but unfortunately resembled the texture of hardtack.....The butternut squash ravioli was rich and delicious... I personally thought the sweetness over took the dish but overall I enjoyed the dish as well as the veal parmesan..I would go back here next time to try some more dishes!! maybe...lobster ravioli or special seafood platter :D

For the third review:

Author	Date	Content
Jacob K.	2011-08-05	I went here at the insistence of a friend to find out what i assumed: so-so food that was over priced and over hyped. There was a line but there is a pick up window where you can get your food fast at a take out window. Tacos are $3 a piece as is the horchada (rice water). I was hungry so I ordered 3 tacos (2 fish and 1 pork shoulder with pineapple)My order was ready and as its from the pick up window your option is a communal park bench outside. Holy crap the tacos were the size of dollar pancakes... I didn't know they made tortillas that small... about 2 bites to a taco. The fish tacos were nothing special. The pork shoulder with pineapple was good but dry. The hot sauces (1 red and one green

		were good). I wanted guacamole and chips but at $8 I don't think so.
Jean-Paul M.	2014-03-23	Best Mexican food I've had in a long time. My mouth is watering thinking about the lobster tacos. Awesome margarita...Tres Viejos. It's spicy (and how a margarita should be!).

For the fourth review:

Author	Date	Content
Amy W.	2010-02-24	Gosh, my review before was short. It's okay, I can make it up now. They close between lunch and dinner, so don't come here in the middle of the day. Upon entering the small parking lot, you will need to grab a ticket from the machine.
David K.	2011-09-29	It's going to be long. Must be brief to fit the experience in 5001 characters.Alinea is a special occasion kinda place. Before you can dine here you must get a reservation here. The trick is they open up reservations two months in advance so if you want to eat in March you better call Jan 1 to snag that spot up (especially if it's a weekend).
Shelley C.	2011-10-24	gotta love sushi gen!Even though its only walk ins and we waited for an hr.Started with a bottle of hot sake to warm up after standing in the cold. baked clam and salmon skin salad were tasty. Oyster was delicious and had to get the uni sushi w quail eggs. (MUST GET!!!) to die for haha I had 2 orders.main we had assorted sashimi, fatty tuna

		was very very yummy, everything is fresh.

3.3 The first review in query.json is about Chinese food.

The second review is Italy food.

The third one is Mexican food.

The fourth one is Japan food.

The fifth one is USA food.