CS 6501: Text Mining

Homework 1: MP1

Student: Mohammad Al Boni

NetID: ma2sm

Pledge

Part 1.1:

1. Normalization code:

```
    public String Normalization(String token) {
            // convert to lower case
            token = token.toLowerCase();
            // rating by stars
            token = token.replaceAll("\\d+star(s)?", "RATE");
            // Some scales and measures
            token = token.replaceAll("\\d+(oz|lb|lbs|cent|inch|piec)", "SCALE");
            // convert some of the dates/times formats
            // 12 hours format
            token = token.replaceAll("\\d{2}(:\\d{2})?(\\s)?(a|p)m", "TIME");
            // 24 hours format
            token = token.replaceAll("\\d{2}:\\d{2}", "TIME");
            // 1st 2nd 3rd 4th date format
            token = token.replaceAll("\\d{1,2}(th|nd|st|rd)", "DATE");
            // convert numbers
            token = token.replaceAll("\\d+.\\d+", "NUM");
            token = token.replaceAll("\\d+(ish)?", "NUM");
// tested on "a 123 b 3123 c 235.123 d 0 e 0.3 f 213.231.1321 g +123 h -123.123"
            // remove punctuations
            token = token.replaceAll("\\p{Punct}", "");
//tested on this string:  "This., -/ is #! an <>|~!@#$%^&*()_-+={[]\"':;?/>.<, $ % ^
& * example ;: {} of a = -_ string with `~)() punctuation"
            return token;
            }
```

2. Curves:

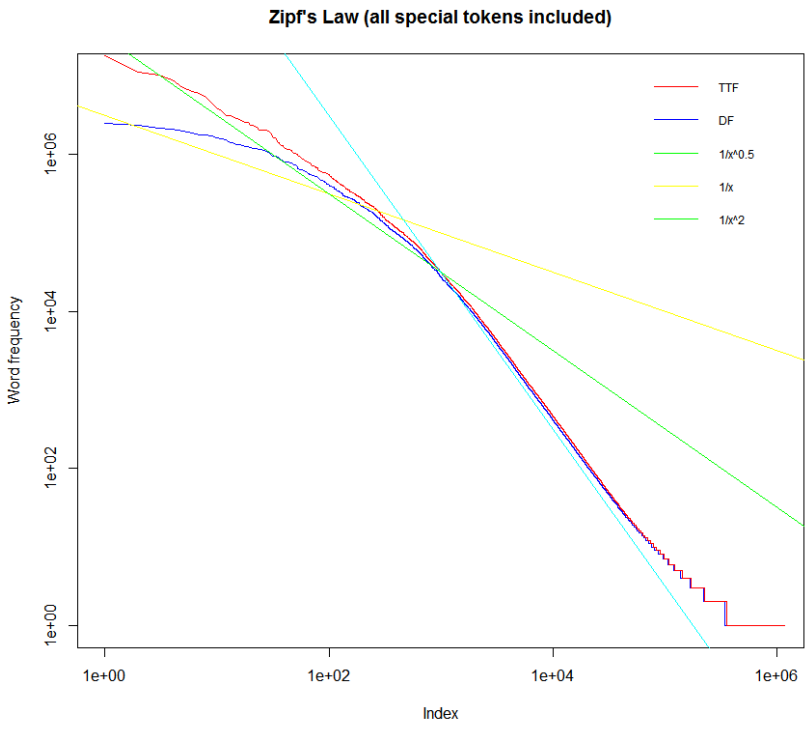

Fig 1. Zipf's Law analysis (case 1: all special tokens are included such as NUM, TIME, SCALE ... etc.).

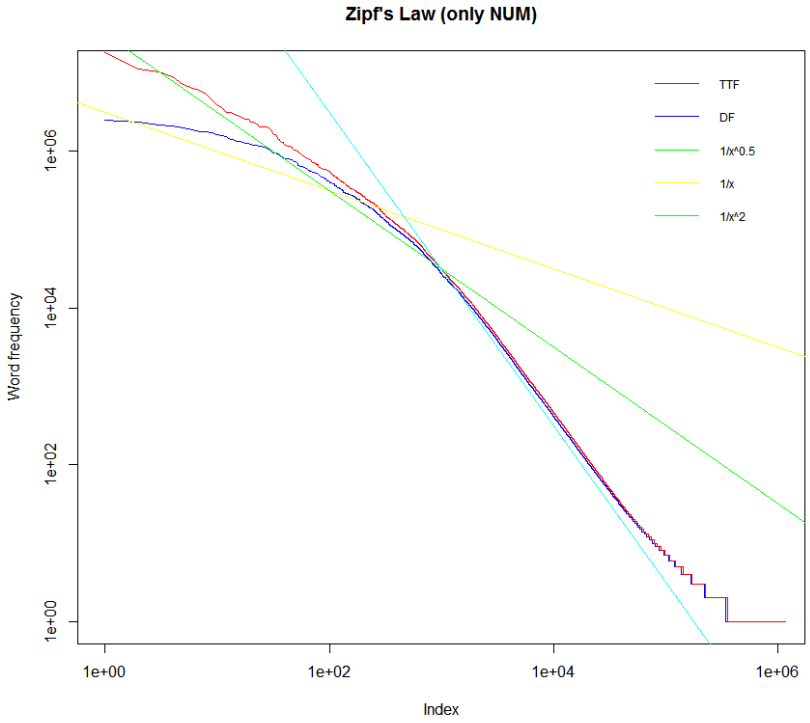

Fig 2. Zipf's Law analysis (case 2: only integers and doubles were normalized to NUM).

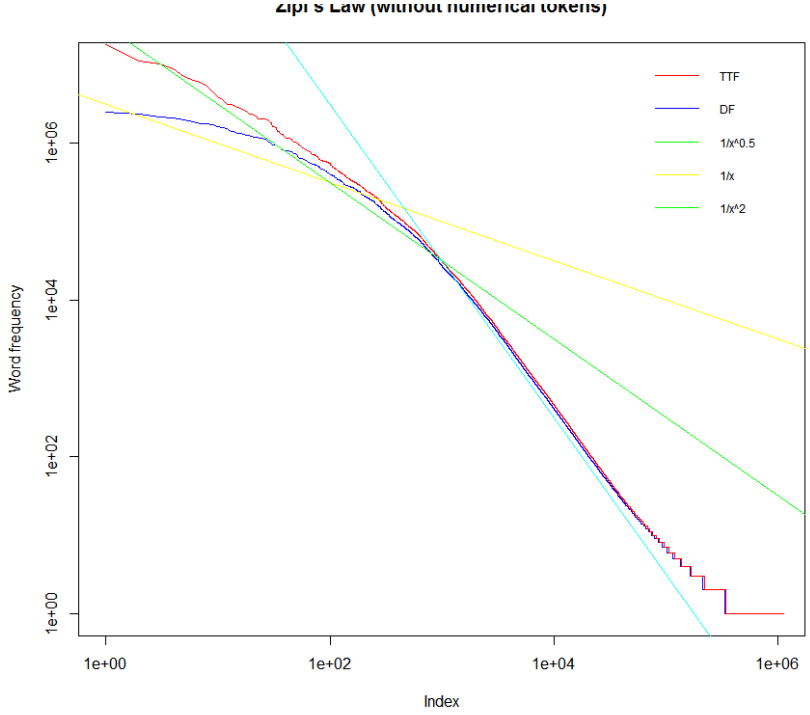

Fig 3. Zipf's Law analysis (case 3: All numerical tokens were excluded from the counting process).

Slops and intercepts:

Curve	Slope	Intercept	Adjusted-R^2
Case 1-DF	-1.099800	6.420667	0.8744
Case 1-TTF	-1.122459	6.555299	0.8785
Case 2-DF	1.099762	6.420244	0.8743
Case 2-TTF	-1.122425	6.554889	0.8785
Case 3-DF	-1.103460	6.428879	0.8737
Case 3-TTF	-1.126050	6.562731	0.8777

3. Discussion:

I preformed three different tests. In the first test, I noticed that the data include some numerical tokens in different formats such as time, date, scales etc. When normalizing only integer and doubles, then we will have separate tokens for each of these values. For example, the normalized vocabulary will include NUMpm, NUMam, NUMst, NUMnd, NUMrd, NUMth, NUMoz, NUMlb … etc. By using special tokens for these values, we can group them and add their counts together. As an instance, NUMpm and NUMam will be grouped and replaced by TIME, NUMst, NUMnd, NUMrd and NUMth will be grouped and replaced by DATE, NUMoz and NUMlb will be grouped and replaced by SCALE … etc. The curves resulted from this test are shown in Figure 1. On the other hand, in the second test, only integer and double values were replaced by the NUM token. As a result, different numerical values will be represented by separate tokens (as the ones mentioned above). The curves resulted from this test are shown in Figure 2. However, Zipf's law describes the language vocabulary which does not typically include special tokens. Therefore, in the third test, all numerical tokens were not included in the counting process. Figure 3 shows the document frequency (DF) and total term frequency (TTF) curves of the third case. After examining the slopes, intercepts and a djusted-R^2 for all of these curves, I found the following: 1) creating special tokens for various numerical values will not have a significant impact on the frequency curves. Although the counts of various tokens will be grouped up, such tokens don't have high frequencies in the various reviews, and thus, will not impact the curves. 2) Although I assumed that Zipf's law does not include special tokens, both TTF and DF curves that include the numerical tokens have better fit than the ones that exclude such tokens (0.8785 vs 0.8777 R^2 and 0.8743 vs 0.8737 R^2 for TTF and DF respectively).

In regard to the questions for this part, we **can** find a strong linear relationship between the x and y axes in the log-log scale curves. This is supported by the fact that all curves have high adjusted-R^2 values (>0.87) which means that they have good fit to a line. Also, I conclude **TTF** fits Zipf's law better on this data set that **DF** does. This is supported by the fact that TTF curves have higher adjusted-R^2 values that DF. This can be explained as

follows: the tail words have very low frequencies regardless to the counting method because they rarely occur in the reviews and if they occur once within a review, both TTF and DF will be increased by 1. Therefore, both TTF and DF will roughly have the same counts for the tail words. As we go towards the head of the curve, TTF will differ from the DF because for any word that occur more than once in any review, TTF will add its count to the overall count while DF will only increase the count by 1 for each document. Using DF, words very close to the head (specially the top ranked words) will saturate to a fixed number which is the number of the documents. Therefore, TTF will fits Zipf's law better since it will have different counts even for the top ranked words.

Part 1.2:

1. Restaurant-specific stop words:
   food,good,nt,NUM,great,order,time,servic,the-food,back,friend,do-nt,love,i-ve,restaur,ve,delici,eat,nice,wait,make,menu,did-nt,tast,price,pretti,fri,tabl,chicken,thing,sauc,flavor,peopl,drink,dish,fresh,night,amaz,meal,i-love,order-the,perfect,side,the-servic,made,bit,chees,lot,salad,dinner,serv,recommend,bar,a-bit,was-nt,seat,enjoy,the-menu,star,a-great,small,day,favorit,experi,food-is,lunch,staff,ll,meat,worth,food-was,tasti,review,servic-was,top,a-good,feel,special,i-order,ca-nt,hour,bad,long,sweet,sandwich,ca,my-friend,find,cook,lot-of,NUM-star,awesom,big,atmospher,expect,the-restaur,bread,server,portion,area.
2. Size controlled vocabulary: 216294 N-grams.
3. Top 50 N-grams:

Rank	N-gram	IDF	Rank	N-gram	IDF	Rank	N-gram	IDF
1	visit	2.125276	18	locat	2.161298	35	check	2.213581
2	plate	2.127093	19	excel	2.162405	36	huge	2.217435
3	minut	2.12778	20	dessert	2.168133	37	select	2.217816
4	beer	2.132289	21	park	2.170488	38	potato	2.22023
5	pork	2.132921	22	end	2.180854	39	kind-of	2.220624
6	NUM-minut	2.135661	23	start	2.183187	40	egg	2.227791
7	rice	2.137601	24	super	2.183499	41	year	2.236507
8	burger	2.138877	25	spici	2.183705	42	dine	2.23774
9	high	2.139513	26	kind	2.185104	43	decid	2.238426
10	spot	2.141319	27	sit	2.196264	44	live	2.243927
11	busi	2.148544	28	quick	2.20064	45	home	2.244092
12	hot	2.149994	29	soup	2.20091	46	appet	2.250607
13	happi	2.152487	30	qualiti	2.202696	47	wine	2.257545
14	walk	2.155357	31	beef	2.202803	48	line	2.258911
15	work	2.156486	32	thought	2.206745	49	roll	2.259722

16	disappoint	2.157866	33	pizza	2.209553	50	larg	2.267461
17	open	2.158308	34	full	2.210673			

Bottom 50 N-grams:

Rank	N-gram	IDF	Rank	N-gram	IDF	Rank	N-gram	IDF
1	upon-bite	5.60062	18	benign	5.60062	35	lifestyl-and	5.60062
2	the-turbot	5.60062	19	mustardi-sauc	5.60062	36	been-compar	5.60062
3	mingon	5.60062	20	these-comment	5.60062	37	lollipop-with	5.60062
4	mind-s	5.60062	21	orenchi-and	5.60062	38	as-refil	5.60062
5	tall-as	5.60062	22	heaven-my	5.60062	39	look-decor	5.60062
6	is-renown	5.60062	23	at-eye	5.60062	40	wtf-pancak	5.60062
7	again-excel	5.60062	24	s-excus	5.60062	41	unlimit-sangria	5.60062
8	cute-coffe	5.60062	25	season-here	5.60062	42	bangkok-and	5.60062
9	guy-soon	5.60062	26	at-dba	5.60062	43	mix-by	5.60062
10	drain-my	5.60062	27	to-spectacular	5.60062	44	shape-pizza	5.60062
11	an-architectur	5.60062	28	resurfac	5.60062	45	no-bill	5.60062
12	them-free	5.60062	29	raspberri-martini	5.60062	46	life-save	5.60062
13	guy-seat	5.60062	30	park-until	5.60062	47	kielbasa-with	5.60062
14	underdeliv	5.60062	31	arm-is	5.60062	48	resist-but	5.60062
15	cheap-yummi	5.60062	32	aphrodit	5.60062	49	lol-there	5.60062
16	even-coat	5.60062	33	got-dress	5.60062	50	simpli-ignor	5.60062
17	fusion-item	5.60062	34	at-opa	5.60062			

Part 1.3:

1. Cosine Similarity Code:

```
    public double ConsineDistance(HashMap<String,Double> d_i,HashMap<String,Double> d_j){
                // since the VSM are stored as sparse vectors, then in the nominator I will go
through only the shared N-grams since d_i[k]*d_j[k] equals to zero if anyone of them is zero
                double dot=0;
                double d_iNorm=0;
                double d_jNorm=0;
                Set<String> set = d_i.keySet();
                Iterator<String> itr = set.iterator();
                // Calculate the nominator and d_i norm
                while (itr.hasNext())
                {
                        String key = itr.next();
                        if(d_j.containsKey(key))
                                dot+=d_i.get(key)*d_j.get(key);
                        d_iNorm+=d_i.get(key)*d_i.get(key);
                }
                // Calculate the d_j norm
                set = d_j.keySet();
                itr = set.iterator();
                while (itr.hasNext())
                {
                        String key = itr.next();
```

```
                d_jNorm+=d_j.get(key)*d_j.get(key);
            }
            // return the cosine distance
            // if one of the VSM has zeros for all terms, we will have zero in the
denominator. Return -1 for this case
            return (d_iNorm==0||d_jNorm==0)?-1:dot/(Math.sqrt(d_iNorm)*Math.sqrt(d_jNorm));
            }
```

2. Similar to "query.json"

a) 1234567890:

#	Author	Date	Content	Cosine
1	Ellie R.	5/7/2014	Great chinese food. I went here for lunch and got the chicken lo mein with egg drop soup and an egg roll. Everything was delicious. Will definitely come back again.	0.198141
2	Jennifer P.	6/14/2012	I ordered lunch from Victoria's two days ago and I feel torn between the portion and value versus the quality of the meal. I purchased a lunch special that included: Egg Drop Soup, 2 Crab Rangoons, General Gau's Chicken (a spicy entree), and Scallion Lo Mein. Considering that I only paid about $5.50 for the meal, I think it was a good value. The only issue that I had with the lunch special, in regards to portions, was that I would have liked to have more chicken. But overall, the meal was really inexpensive.The quality of the food was extremely subpar. The Egg Drop Soup had giant chunks of egg. The Crab Rangoons were tasty, but much more soggy dough than filling. The "spicy" General Gau was extremely bland, the chicken was dry and the stringiness of the meat meant that it was cooked inappropriately or it was old. Either way, it was not pleasant. The Lo Mein was very bland and literally was only noodles and one scallion; no sauce, no other veggies, no flavor. I am willing to try this place again, because the food is cheap. And I really hope it is better next time.	0.165061
3	Marissa P.	11/20/2010	Not impressed. Came here with friends who were raving about this place. Major let down. For Chinese food especially, this was about as bland as it gets. My egg drop soup had no taste. Flavorless broth and bad texture. I usually love egg drop soup. Then, the crab rangoons were also not good... they were cold in the middle. I asked, "anybody else get a cold crab rangoon?" and my friend replied, "I was just thinking that." So, it wasn't just me-- pretty sure they just microwaved them real quick. Beyond being cold, they weren't very tasty in comparison to crab rangoons I've had elsewhere. In addition, the spare ribs were flavorful but too greasy to eat. I will say though that service was prompt and friendly.	0.155209

b) 12345678901:

#	Author	Date	Content	Cosine
1	William C.	1/25/2014	we were there Jan. 16. for an early dinner. a bit slow and took 2.5 hours to finish. The bread is not as good as the old pastry chef, Jamie. the panna cotta was different but i enjoyed my hake. Used to be my favorite but not sure now	0.205083

2	K B.	5/26/2013	3.5 star dinner review. Based on yelp reviews we ordered the grilled calamari, lobster ravioli, the vege lasagna, and the panna cotta. The calamari had good enough flavor but was really chewy. The lasagna was ehh ok- the pasta was a little tough and it was barely sauced - and a bland sauce at that. The lobster ravioli was good and the sauce delicious. And the panna cotta was decadent. It's a pleasant dining experience overall, though a little warm inside and our waiter didn't check on us at all. And there wasn't any balsamic on the table to mix for the bread. I've had better Italian in the city.	0.167941
3	Jerry B.	9/5/2009	Lobster Ravioli AMAZINGGGGGGGG!!!	0.158787

c) 123456789012:

#	Author	Date	Content	Cosine
1	Jay L.	7/12/2011	Their pork tacos are fantastic. They only serve them for dinner but they are worth the wait.	0.250641
2	Jonathan P.	10/27/2013	I love this place the food is beyond amazing. My mouth is watering just writing this review!!!! the shrimp tacos are to die for.....	0.200149
3	Nina B.	10/5/2008	I could eat just their pico de gallo and be happy. But I wouldn't, because everything else is awesome too.	0.198513

d) 1234567890123:

#	Author	Date	Content	Cosine
1	David G.	3/25/2011	I'm kind of shocked by all the 1 and 2-star reviews. Steve and I had an absolutely perfect meal and experience, dining here for lunch. We started off getting the Asian Lettuce Wraps. We both thought they were the best lettuce wraps we've ever eaten. Delicious grilled chicken, thin cold noodles with black sesame seeds, shredded carrots, thinly sliced cucumbers, bean sprouts and three kinds of sauces. They were unbelievably delicious... and oh-so sticky. So I asked our server if they had wet-naps for our hands. She came back in a few minutes with a plate that was holding steaming hot fabric napkins with lemon slices. We were both very impressed. The appetizer and service were both 5-stars.Steve's entree was Louisiana Chicken and Pasta. Which consisted of bow-tie pasta, assorted peppers, mushrooms, parmesan-encrusted chicken in a creamy pepper sauce. Steve gave his pasta dish 4 out of 5-stars. My entree was a Spicy Crispy Chicken Sandwich (lettuce, tomato, chipotle mayo, homemade bun) served with sweet potato fries. Perfect sandwich. Perfect fries. Perfect 5-star rating.It's been a long time since we've both been so impressed with the service and food at a restaurant -- regardless of the fact that Cheesecake Factory is a chain. It really was one the best meals we've had in eons**eons = one or two months at most ;-)	0.097985
2	Aaron L.	2/22/2011	The Chubby3.5* The Roe	0.095295
3	David E.	5/18/2012	Try the Roe Burger! Staff is very friendly.	0.090298

good, you found it!!

e) 1234567890124:

#	Author	Date	Content	Cosine
1	Patricia S.	6/18/2014	Pretty great so far. They use Amaya Coffee beans which I (gasp!) prefer to Greenway. I've been three times now and on each visit, my drink was perf. The siphon coffee tasted very clean and smooth. I like the interior's layout and there is plenty of seating, inside and out. However, you're out of luck if you don't manage to get a seat at one of those comfy leather chairs or the long bench along the back wall. The stools they use were really uncomfortable for myself and everyone in our party of 4. Maybe it's a clever way to keep people from lingering there too long.I've tried a few of their pastries but I'm mostly into their beef empanadas. I will always choose something savory over a sweet in the morning. I hope they always have those available.All in all, a pretty winsome place.	1
2	Susan C.	11/24/2012	Very clean and nice restaurant	0.165444
3	Bob L.	11/16/2011	Good, and love the interior.	0.137448

3. Type of restaurants:
   a) 1234567890: Chinese ("egg drop soup", "crab rangoons").
   b) 12345678901: Italian ("Lobster Ravioli").
   c) 123456789012: Mexican ("shrimp tacos","pork tacos","pico de gallo").
   d) 1234567890123: Asian ("Asian Lettuce Wraps", "the Roe").
   e) 12345678901234: coffee shop.

It seems your solution prefers shorter reviews.

Please double check your implementation.