Michael Gyarmathy
CSCE 470
11/4/14

Homework 4 Report

Task 1.1: Basic HITS

```
$ python task1-1.py

Top Hubs
1.   User ID: 61814506    Score: 7605671766190
2.   User ID: 34452333    Score: 6980972614142
3.   User ID: 20594561    Score: 6781091196613
4.   User ID: 34964450    Score: 6595112985949
5.   User ID: 66865122    Score: 6522349266194
6.   User ID: 21845494    Score: 6445797228944
7.   User ID: 15675786    Score: 6371705034878
8.   User ID: 24092738    Score: 6370518944370
9.   User ID: 21482564    Score: 6347365570633
10.  User ID: 14464925    Score: 6216964845684

Top Authorities
1.   User ID: 26257166    Score: 9391314819021
2.   User ID: 32135704    Score: 7559416821751
3.   User ID: 40519997    Score: 7504968586687
4.   User ID: 34701524    Score: 5978541795709
5.   User ID: 21254264    Score: 4878407602304
6.   User ID: 52529345    Score: 4870135708124
7.   User ID: 22159122    Score: 4663121295370
8.   User ID: 43139414    Score: 4368781242280
9.   User ID: 55307193    Score: 4321762099168
10.  User ID: 23778898    Score: 3931286405057
```

Task 1.2: Manipulating HITS

a) Added `["spam_hub", {"in": [], "auth": 0, "hub": 0, "out": ["26257166", "32135704", "40519997", "34701524", "21254264", "52529345", "22159122", "43139414", "55307193", "23778898"]}]` to sports_list.json. After rerunning HITS, the new "spam_hub" has an authority score of 0 and a hub score of 2482831616662. This ranks "spam_hub" at number 489 in the list of Top Hubs and tied for last in the Top Authorities.

```
Top Hubs
1.   User ID: 61814506    Score: 7605671766190
2.   User ID: 34452333    Score: 6980972614142
3.   User ID: 20594561    Score: 6781091196613
```

```
...
497. User ID: 44564856    Score: 2488731626407
498. User ID: spam_hub    Score: 2482831616662
499. User ID: 70488241    Score: 2477992691669
500. User ID: 16655109    Score: 2477954456396
```

b) Renamed “spam_hub” to “spam_hub1” and then added “spam_hub0-9” clones. Also added `["spam_auth", {"in": ["spam_hub1", "spam_hub2", "spam_hub3", "spam_hub4", "spam_hub5", "spam_hub6", "spam_hub7", "spam_hub8", "spam_hub9", "spam_hub0" ], "auth": 0, "hub": 0, "out": []}]` to sports_list.json. After rerunning HITS, the new “spam_auth” has an authority score of 49929413780 and a hub score of 0. This ranks “spam auth” at number 1469 in the list of Top Authorities and tied for last in Top Hubs.

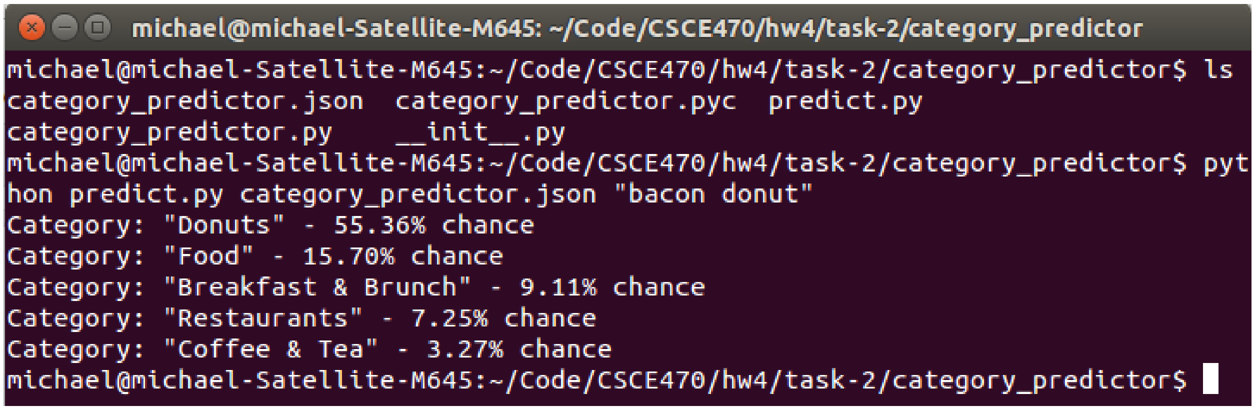
```
Top Authorities
1.    User ID: 26257166    Score: 9391314819021
2.    User ID: 32135704    Score: 7559416821751
3.    User ID: 40519997    Score: 7504968586687
...
1468. User ID: 22163953    Score: 50040887926
1469. User ID: spam_auth   Score: 49929413780
1470. User ID: 33324885    Score: 49898707328
1471. User ID: 14835313    Score: 49852260574
```

c) By adding the Top 200 authorities to “spam_hub1”’s out network, I was easily able to catapult this hub into the top spot. Boosting the rankings of an authority is a little more difficult if you cannot modify the pre-existing network. In order to get around this, you could flood the network with top- spam hubs and have them exclusively point to your spam authority.

```
Top Hubs
1.    User ID: spam_hub1   Score: 10764874702391
2.    User ID: 61814506    Score: 7605671766190
3.    User ID: 34452333    Score: 6980972614142
4.    User ID: 20594561    Score: 6781091196613
```

Task 2.1: Getting Started with Yelp

1. Successfully ran the EMR job on AWS that produced category_predictor.json as output.
2. `python predict.py category_predictor.json “bacon donut”`


michael@michael-Satellite-M645: ~/Code/CSCE470/hw4/task-2/category_predictor

```
michael@michael-Satellite-M645:~/Code/CSCE470/hw4/task-2/category_predictor$ ls
category_predictor.json  category_predictor.pyc  predict.py
category_predictor.py    __init__.py
michael@michael-Satellite-M645:~/Code/CSCE470/hw4/task-2/category_predictor$ pyt
hon predict.py category_predictor.json "bacon donut"
Category: "Donuts" - 55.36% chance
Category: "Food" - 15.70% chance
Category: "Breakfast & Brunch" - 9.11% chance
Category: "Restaurants" - 7.25% chance
Category: "Coffee & Tea" - 3.27% chance
michael@michael-Satellite-M645:~/Code/CSCE470/hw4/task-2/category_predictor$ 
```

Task 2.2: Statistical Language Models

The Process

In order to optimize the probability computations using the mixed model, I used the wordCount.py MapReduce program from hw3, (which I renamed to generalCollectionFrequency.py), to calculate term frequencies of the entire corpus of Yelp reviews. This EMR job outputted a file called generalCollectionFrequencyOutput, which is used as part of the input in all of my Statistical Language Models EMR jobs:

```
$ python generalCollectionFrequencyOutput.py
s3://mgyarmathy-yelp-data/yelp_academic_dataset_review.json >
generalCollectionFrequencyOutput -r emr -c emr.conf
```

Next, I wrote an EMR job, called task2-2.py, that calculates the P(q|d) for each Yelp Review. The query for the EMR job is set in the `self.query_terms` variable in the `setup()` function.

```
$ python task2-2.py
s3://mgyarmathy-yelp-data/yelp_academic_dataset_review.json >
reputation_output -r emr -c emr.conf --file generalCollectionFrequencyOutput
```

Then, I wrote a quick python script, called task2-2-results.py, to read the results from the task2-2.py EMR job and output the top 5 reviews with their P(q|d) scores:

```
> python task2-2-results.py reputation_output

[('zqHax4v_OdBtL6Ry-3bTeA', 0.09999999999999999), ('nLTnwqtWNb9eco4B583P0Q',
0.05833333333333333), ('_ZycwzJazjED6rNgYo9_wg', 0.049999999999999996),
('AQ9KkpTnNJDfQFKOdD54hA', 0.04666666666666666), ('KRo1uHs4xaTdv0YdYxg2Iw',
0.04375)]
```

Last, I used egrep to search the yelp_academic_dataset_review.json file for the review_ids of the top 5 reviews to retrieve the text content of the review. (NOTE: due to its enormous file size, yelp_academic_dataset_review.json is not included in the homework submission)

```
> egrep
'zqHax4v_OdBtL6Ry-3bTeA|nLTnwqtWNb9eco4B583P0Q|_ZycwzJazjED6rNgYo9_wg|AQ9KkpTn
NJDfQFKOdD54hA|KRo1uHs4xaTdv0YdYxg2Iw' data/yelp_academic_dataset_review.json
```

```
{"votes": {"funny": 1, "useful": 0, "cool": 1}, "user_id":
"tFyQbNbBQEyEc9oCr1pJUg", "review_id": "zqHax4v_OdBtL6Ry-3bTeA", "stars": 5,
"date": "2012-11-05", "text": "Starbucks carries a great reputation. LOVE
STARBUCKS!", "type": "review", "business_id": "2jCyo55dFgGzYpoIX8TwgA"}
{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id":
"vPw8ZwVq9oB7-MCIR9UPFQ", "review_id": "AQ9KkpTnNJDfQFKOdD54hA", "stars": 4,
"date": "2014-04-30", "text": "This place lives up to its reputation!  Jerk
chicken and pork fried rice were great!", "type": "review", "business_id":
"Zx8_4zKdDBSO3qGrkukBIA"}
{"votes": {"funny": 0, "useful": 1, "cool": 1}, "user_id":
"5FG3YcvOngpZ1Yl-xFMH_A", "review_id": "nLTnwqtWNb9eco4B583P0Q", "stars": 5,
"date": "2014-07-14", "text": "This restaurant definitely lives up to its
reputation, great meal and service.", "type": "review", "business_id":
"gVYju3XRcO1R4aNk7SZJcA"}
{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id":
"4eBoF2Dh8cXIW99_a-EBOA", "review_id": "KRo1uHs4xaTdv0YdYxg2Iw", "stars": 2,
"date": "2007-11-03", "text": "food is not bad, but for all the reputation of
Puck, it's way below expectation.", "type": "review", "business_id":
"BPRMEhCf8ugvQrNnRM1xUw"}
{"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id":
"nsZruFN0TBW06ixhBcNokg", "review_id": "_ZycwzJazjED6rNgYo9_wg", "stars": 4,
"date": "2013-09-29", "text": "For the price and reputation, Scottsdale
location can do a little better, service-wise", "type": "review",
"business_id": "OE5nAmaSVaopeRS1Cs9Kuw"}
```

The Results

The queries for this assignment are fairly short, so I decided to calculate P(q|d) rather than log(P(q|d)) because the overhead of computing the log for each term probability and then adding them all together is greater than multiplying out all of the fractions. Also, computing the sum of logs is difficult if you have zero probabilities for some terms. I've included the code for computing P(q|d) and log(P(q|d)) in the comments of task2-2.py

- q = "reputation"

Rank	Review ID	Text	P(q|d)
1.	zqHax4v_OdBtL6Ry-3bTeA	"Starbucks carries a great reputation. LOVE	0.09999999999999999

		STARBUCKS!"	
2.	nLTnwqtWNb9eco4B583P0Q	"This restaurant definitely lives up to its reputation, great meal and service."	0.058333333333333333
3.	_ZycwzJazjED6rNgYo9_wg	"For the price and reputation, Scottsdale location can do a little better, service-wise"	0.049999999999999996
4.	AQ9KkpTnNJDfQFKOdD54hA	"This place lives up to its reputation! Jerk chicken and pork fried rice were great!"	0.046666666666666666
5.	KRo1uHs4xaTdv0YdYxg2Iw	"food is not bad, but for all the reputation of Puck, it's way below expectation."	0.04375

- q = “vegetarian”

Rank	Review ID	Text	P(q|d)
1.	ubW5K3BdHSx4IV75mlJ9mw	"Vegetarian Curry!"	0.35
2.	WO71GRBY67y5MC8m25peWg	"Vegetarian delights found here!"	0.175
3.	sIF5aD4N8vsBbZSEZOVMEQ	"Best vegan, vegetarian and non-vegetarian cafe in Las Vegas period!!"	0.12727272727272726
4.	dRwZA6x3MiYLryqjrq7Jlw	"Best vegetarian restaurant in Phoenix area!"	0.11666666666666665
5.	nM9QndpLpoYrAP2PG2ycfw	"Update: They are also anti-vegetarian!"	0.11666666666666665

- q = “car dealer”

Rank	Review ID	Text	P(q|d)
1.	Cr2ecTLiw1PqOtczoK5L3g	"the salesmen have gotten scuzzier ands lot more pushy. when I leased my car 3 years ago they were wonderful. now they're just like every other car dealer. shame."	0.0010888888888888886
2.	jwOqs3HiBzSzAlgoiohaqw	"Honest and professional people and service. \n\nNoticed the car was still under warranty for the part in question, didn't charge me any thing and got the car fixed by the dealer for no additional costs."	0.0007561728395061726
3.	kukvvd1eo5GEMXRonUcN1Q	"Stay away from client advisor Alexander Smith (especially when you come ONLY to test drive). He would pressure you to buy the car. Very non-BMW dealer characteristic."	0.0006249999999999999
4.	E6PJgkCEAvnHMgUruZLsUQ	"Overcharged and overfilled.\nThey insisted that my car required synthetic oil which I found out after was not the case. They overfilled my car with oil resulting in my car leaking oil in the garage for many days in a row. I am going to the dealer next time."	0.0006122448979591835

5.	q0Eb7fyQitZAumTSil4pRg	"My car battery died & the dealer sent Customer One Towing to me (it was under warranty if I had the dealer test it and change it out). They sent me a text message to let me know they were on the way and confirm the address and then were here 10 minutes before their promised time. The driver was great, took great care in loading my car onto the flatbed, and got my car to the dealer safely. \n\nI put them in my contacts on my phone in case I ever need another tow!"	0.0004990946129470348

4. q = “cowboy bbq”

Rank	Review ID	Text	P(q|d)
1.	'-2x6p7MEg0qpXA6oLzl9zg'	"Not a huge fan of sushi, but the Cowboy roll is delicious, and the BBQ pork is the best I've ever had!"	0.0009262759924385631
2.	'xpm6TgDiHaQdEDlErFsqvQ'	"One of the best BBQ beef sandwiches I've ever had. No fuss barbecue and quite the experience. Oh, and the cowboy beans are amazing too."	0.0007248520710059172
3.	'Yr1z2cm1oGO5p9-g7SXBSQ'	"Good wholesome BBQ, but they were out of collard greens :-(The cornbread muffins weren't the best either, but the brisket and cowboy beans are worth coming for."	0.0006249999999999999
4.	'_zw06LXbN1NlZlliMK6f3A'	"Love BBQ?? ..NOT SURE WHAT TO GET? ..GET THE RIBLET BASKET!!! It has the most meat for th best price. You will definitely need a box. Waldos is the best BBQ in Mesa. Falls off the bone, sauce is homemade and the sides are all you can eat. Mmmm, cowboy beans, coleslaw and spicy fries! My fave."	0.00030163127116035693
5.	'KJL5zxBoJbrJ8a3c0EtzHg'	"Visiting from Chicago... First, we arrived late, almost closing time. The guys were great-they recommended the brisket and it was amazing. Cowboy beans were excellent and portions huge!! Friendly service, good food, a high character BBQ place. Not for someone looking for a fancy sit down place. But if you are looking for good quality bbq, I Recommended!"	0.00028152829646653255

5. q = “best cookie store”

Rank	Review ID	Text	P(q|d)
1.	M8GLEcssrfFjuY1UfGWeAA	"2nd best breakfast place in Casa Grande. The best is Cookie Jar on 2nd street.\n\nBut I still love this place has the best ambiance, very friendly staff, and always fresh and well-cooked breakfast. Cute little store too."	1.7346887169372375e-05

2.	aZzKBLUDEQ7-nXBPtuJrRQ	"I use the bakery and deli at this store often for food for business meetings and trainings. The bakery makes great cookie trays, not the type you expect from a chain grocery store. Their fresh baked fresh daily rolls and buns are the best. The deli has freshly made fruit and veggie trays that are reasonable priced and nicely done."	3.1759259259259246e-06
3.	7V2InU1bNWWSvMnEhuFezg	"The Cookie Bar had a booth at Best Buddies annual Friendship Walk recently and I purchased the \"virgin\" Pina Colada cookies and their chocolate chip cookies. The Pina Colada cookies were absolutely amazing as they were so different--rich, yet light and truly a delightful creation! Now I have to go to their store!"	2.1782756693593445e-06
4.	OHyy7cjfKkkXNNpWSyoqsg	"I'm beyond exciting to see the first Honolulu Cookie Company store outside of Hawaii. These taste so amazing and with single packed cookies they are perfect for gifts. Coconut dipped in white chocolate is my favorite but I've never had a flavor I didn't enjoy. The best part is that they allow you to sample all the different kinds in the store as well. Here is to a great company with a wonderful product and tremendous staff!"	1.3398437499999995e-06
5.	evwtUVMeajNO09ljFeQxrQ	"This store is a gem! There is an amazing selection of unique and one of kind clothing for the not so \"cookie cutter\" crowd. Lynn the dedicated owner goes above and beyond to help her customers find what they are looking for. It gets better.... The BEST tailoring in Phoenix is available for both men & women of all sizes by the legendary tailor, Kwan. Merchandise does not need to be purchased at the store to be altered. A well kept secret is that many other clothiers in Scottsdale outsource their tailoring to Exclusively Big & Tall. Rub elbows with professional athletes while you shop this wonderful store!"	8.639682422402383e-07