ESE 545 Project1

Jiaxiao Cai
Collaborate with Tejas Srivastava

Problem 1

Statement: Pre-processing data to the correct format with (’reviewerID’, ’reviewText’)

Approach:

- Read data using Pandas dataframe
- Discard the other unrelated columns such as 'asin', 'reviewerName', 'helpful', 'overall', 'summary', 'unixReviewTime', 'reviewTime' from the data set
- Remove stop words
- Convert punctuation to empty space
- Convert duplicate empty space to one empty space
- Drop the review with length less than shingles length (k)

Result: Name: ReviewText; Length: 157680

Problem 2

Statement: Convert each review into a set of k-shingles

Approach:

- Define a coo-matrix with shape=(37**4, len(ReviewText))
- Create a function *find_index_in_binary_matrix,* which will take a review as input and calculate the corresponding index for each shingle
- Use Pandas.apply to apply the *find_index_in_binary_matrix* in each review in parallel
- Generate the sparse matrix using row, column and data values

Result: Binary Matrix; Dimensions: 1874161 * 157680

Problem 3

Statement: Pick 10,000 pairs of reviews at random and compute the average Jac-card distance and the lowest distance among all pairs.

Approach:

- Shuffle the index of 142 documents randomly from total documents because we could make 10011 pairs out of 142 documents
- Convert the coo-matrix to numpy array
- Calculate the Jaccard distance for each pair
- Jaccard Distance = 1 – intersection(Pair1, Pair2) / union(Pair1, Pair2)
- Plot the histogram for each pair

Result:
Average value of Jaccard Distance: 0.972
Minimum value of Jaccard Distance: 0.819

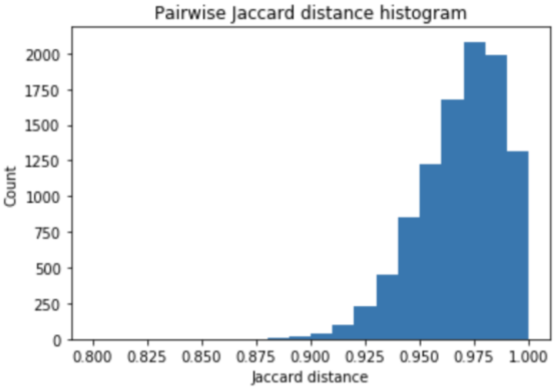


Figure 1 Jaccard Distance

Problem 4

Statement: More effective to store data to find the approximate nearest neighbor of a review

Approach: Min Hashing and Locality Sensitive Hashing is used to reduce the time and space complicity while computing the nearest neighbour

1. Min Hashing:
   - For each document, find the indexes where there is one and store in a numpy array as *binary_matrix_index*
   - Create m (m=120) hash function; select the corresponding a and b for these m hash function in random from 0 to 2000
   - Permute the indexes of each document to new index with respect to the function: permuted_index = (a*x + b) % (37^4)
   - Hence, we get m Hash value for each document of size [120, 157680]
2. Local Sensitivity Hashing:
   - Create b (b=20) bands each consisting of r (r=6) rows; select the corresponding a and b for these functions from 0 to 982451653
   - Permute the hash value of each document to new index with respect to the function: permuted_index = (a*x + b) % p
   - Calculate the sum of each vector band, and the size of the Hash table with the size of [20, 157680]

Result:
1. Original Dataset Sample (sparse matrix)

Index	Value
(455083, 0)	1.0
(1844800, 0)	1.0
……	1.0
(779631, 157679)	1.0
(733950, 157679)	1.0

2. Hash Table Sample

```
[[ 7472 38742  4912 ...   818   412  5894]
 [13935 21572   759 ...  2686  7653   840]
 [ 2177  8327 24461 ...  2302 11211  4704]
 ...
 [  454  5149  8626 ...  3082 10588  4072]
 [ 4858 11040  7411 ...   519 14642  1710]
 [ 3268 16920 11002 ...  3268  7075  6706]]
```

3. Local Sensitivity Hashing Table

```
[[2.42815561e+09  2.61687555e+09 ... 2.94125897e+09  2.43340210e+09]
 [3.15663816e+09  3.04152007e+09 ... 3.61470857e+09  3.20712201e+09]
 [3.38655209e+09  3.83028871e+09 ... 1.96893080e+09  2.40676553e+09]
 ...
 [3.12568361e+09  2.94625217e+09 ... 2.18965851e+09  4.08602568e+09]
 [2.72879242e+09  3.32631730e+09 ... 1.97086841e+09  3.63960968e+09]
 [2.78486404e+09  2.85245843e+09 ... 4.22640403e+09  3.16024730e+09]]
```

Problem 5

Statement: Effective way to detect all pairs of reviews that are close to one another

Approach:

- Each band of the local sensitivity hashing table is checked with other documents and a document is similar if any one of the band values matches at that position.
- Similar pairs are matched to each other and they are mapped to the same bucket if they have Jaccard distance of less than 0.2
- Calculate the Jaccard distance of similar pairs are calculated using the sparse matrix

Parameter Justification:

The following diagram shows the relationship between similarity and the probability of hitting. The value of Band(b1) = 20 and rows per band r = 6 give better results as we get less false negatives. The values of r1 and b give better results with probability 0.9977. The probability of finding similar pairs is given by P = 1 – (1 – (S ^ r)) ^ b.

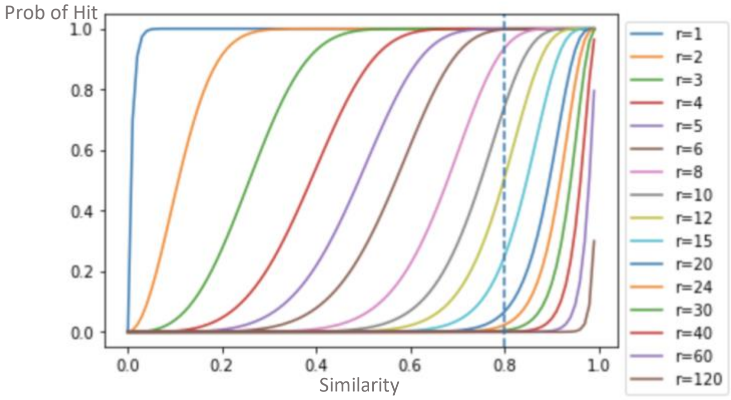
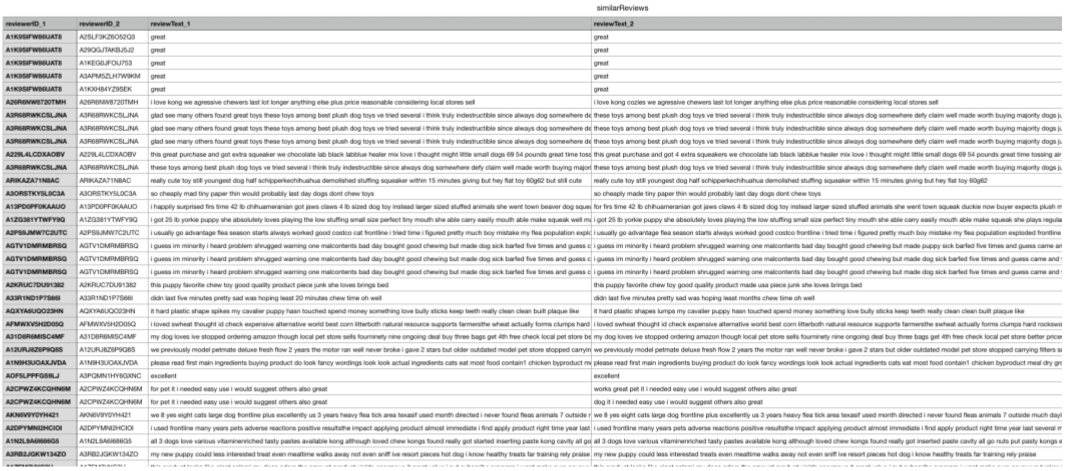


Figure 2 LSH Parameter Tuning

Result:

We found 1113 pairs when Jaccard distance < 0.2.



similarReviews

reviewerID_1	reviewerID_2	reviewText_1	reviewText_2
A1K9SIFW86UAT8	A2SLF3KZ6O52Q3	great	great
A1K9SIFW86UAT8	A29QGJTAKBJ5J2	great	great
A1K9SIFW86UAT8	A1KEG0JFOIJ753	great	great
A1K9SIFW86UAT8	A3APM5ZLH7W9KM	great	great
A1K9SIFW86UAT8	A1KXH84YZ9SEK	great	great
A26R6NW8720TMH	A26R6NW8720TMH	i love kong we agressive chewers last lot longer anything else plus price reasonable considering local stores sell	i love kong cozies we agressive chewers last lot longer anything else plus price reasonable considering local stores sell
A3R68RWKCSLJNA	A3R68RWKCSLJNA	glad see many others found great toys these toys among best plush dog toys ve tried several i think truly indestructible since always dog somewhere de	these toys among best plush dog toys ve tried several i think truly indestructible since always dog somewhere defy claim well made worth buying majority dogs ju
A3R68RWKCSLJNA	A3R68RWKCSLJNA	glad see many others found great toys these toys among best plush dog toys ve tried several i think truly indestructible since always dog somewhere de	these toys among best plush dog toys ve tried several i think truly indestructible since always dog somewhere defy claim well made worth buying majority dogs ju
A3R68RWKCSLJNA	A3R68RWKCSLJNA	glad see many others found great toys these toys among best plush dog toys ve tried several i think truly indestructible since always dog somewhere de	these toys among best plush dog toys ve tried several i think truly indestructible since always dog somewhere defy claim well made worth buying majority dogs ju
A229L4LCDXAOBV	A229L4LCDXAOBV	this great purchase and got extra squeaker we chocolate lab black labblue healer mix love i thought might little small dogs 69 54 pounds great time toss	this great purchase and got 4 extra squeakers we chocolate lab black labblue healer mix love i thought might little small dogs 69 54 pounds great time tossing air
A3R68RWKCSLJNA	A3R68RWKCSLJNA	these toys among best plush dog toys ve tried several i think truly indestructible since always dog somewhere defy claim well made worth buying majori	these toys among best plush dog toys ve tried several i think truly indestructible since always dog somewhere defy claim well made worth buying majority dogs ju
ARIKAZA71N8AC	ARIKAZA71N8AC	really cute toy still youngest dog half schipperkechihuahua demolished stuffing squeaker within 15 minutes giving but hey flat toy 60g62 but still cute	really cute toy still youngest dog half schipperkechihuahua demolished stuffing squeaker within 15 minutes giving but hey flat toy 60g62
A3ORSTKY5L0C3A	A3ORSTKY5L0C3A	so cheaply mad tiny paper thin would probably last day dogs dont chew toys	so cheaply made tiny paper thin would probably last day dogs dont chew toys
A13PD0PF0KAAUO	A13PD0PF0KAAUO	i happily surprised firs time 42 lb chihuameranian got jaws claws 4 lb sized dog toy instead larger sized stuffed animals she went town beaver dog sque	for firs time 42 lb chihuameranian got jaws claws 4 lb sized dog toy instead larger sized stuffed animals she went town squeak duckie now buyer expects plush m
A1ZG381YTWFY9Q	A1ZG381YTWFY9Q	i got 25 lb yorkie puppy she absolutely loves playing the low stuffing small size perfect tiny mouth she able carry easily mouth able make squeak well m	i got 25 lb yorkie puppy she absolutely loves playing the low stuffing small size perfect tiny mouth she able carry easily mouth able make squeak she plays regula
A2PS9JMW7C2UTC	A2PS9JMW7C2UTC	i usually go advantage flea season starts always worked good costco cat frontline i tried time i figured pretty much boy mistake my flea population explo	i usually go advantage flea season starts always worked good costco frontline i tried time i figured pretty much boy mistake my flea population exploded frontline
AGTV1DMRMBRSQ	AGTV1DMRMBRSQ	i guess im minority i heard problem shrugged warning one malcontents bad day bought good chewing but made dog sick barfed five times and guess c	i guess im minority i heard problem shrugged warning one malcontents bad day bought good chewing but made puppy sick barfed five times and guess came an
AGTV1DMRMBRSQ	AGTV1DMRMBRSQ	i guess im minority i heard problem shrugged warning one malcontents bad day bought good chewing but made dog sick barfed five times and guess c	i guess im minority i heard problem shrugged warning one malcontents bad day bought good chewing but made dog sick barfed five times and guess came and
AGTV1DMRMBRSQ	AGTV1DMRMBRSQ	i guess im minority i heard problem shrugged warning one malcontents bad day bought good chewing but made dog sick barfed five times and guess c	i guess im minority i heard problem shrugged warning one malcontents bad day bought good chewing but made dog sick barfed five times and guess came and
A2KRUC7DU91382	A2KRUC7DU91382	this puppy favorite chew toy good quality product piece junk she loves brings bed	this puppy favorite chew toy good quality product made usa piece junk she loves brings bed
A33R1ND1P7S66I	A33R1ND1P7S66I	didn last five minutes pretty sad was hoping least 20 minutes chew time oh well	didn last five minutes pretty sad was hoping least months chew time oh well
AQXYA6UQO23HN	AQXYA6UQO23HN	it hard plastic shape spikes my cavalier puppy hasn touched spend money something love bully sticks keep teeth really clean clean built plaque like	it hard plastic shapes lumps my cavalier puppy hasn touched spend money something love bully sticks keep teeth really clean clean built plaque like
AFMWXV5H2D05Q	AFMWXV5H2D05Q	i loved swheat thought id check expensive alternative world best corn litterboth natural resource supports farmersthe wheat actually forms clumps hard	i loved swheat thought id check expensive alternative world best corn litterboth natural resource supports farmersthe swheat actually forms clumps hard rockswo
A31D8R6MISC4MF	A31D8R6MISC4MF	my dog loves ive stopped ordering amazon though local pet store sells fourninety nine ongoing deal buy three bags get 4th free check local pet store be	my dog loves ive stopped ordering amazon though local pet store sells fourninety nine ongoing deal buy three bags get 4th free check local pet store better pricer
A12URJ8Z6P9Q8S	A12URJ8Z6P9Q8S	we previously model petmate deluxe fresh flow 2 years the motor ran well never broke i gave 2 stars but older outdated model pet store stopped carryin	we previously model petmate deluxe fresh flow 2 years the motor ran well never broke i gave 2 stars but older outdated model pet store stopped carrying filters s
A1N9H3UOAXJVDA	A1N9H3UOAXJVDA	please read first main ingredients buying product do look fancy wordings look look actual ingredients cats eat most food contain1 chicken byproduct m	please read first main ingredients buying product do look fancy wordings look look actual ingredients cats eat most food contain1 chicken byproduct meal dry gro
AOF5LPPFGS9LJ	A3PQMN1HY6GXNC	excellent	excellent
A2CPWZ4KCQHN6M	A2CPWZ4KCQHN6M	for pet it i needed easy use i would suggest others also great	works great pet it i needed easy use i would suggest others also great
A2CPWZ4KCQHN6M	A2CPWZ4KCQHN6M	for pet it i needed easy use i would suggest others also great	dog it i needed easy use i would suggest others also great
AKN6V9Y0YH421	AKN6V9Y0YH421	we 8 yes eight cats large dog frontline plus excellently us 3 years heavy flea tick area texasif used month directed i never found fleas animals 7 outside r	we 8 yes eight cats large dog frontline plus excellently us 3 years heavy flea tick area texasif used month directed i never found fleas animals 7 outside much dayl
A2DPYMNI2HCIOI	A2DPYMNI2HCIOI	i used frontline many years pets adverse reactions positive resultsthe impact applying product almost immediate i find apply product right time year last	i used frontline many years pets adverse reactions positive resultsthe impact applying product almost immediate i find apply product right time year last several m
A1N2L9A6I686G5	A1N2L9A6I686G5	all 3 dogs love various vitaminenriched tasty pastes available kong although loved chew kongs found really got started inserting paste kong cavity all go	all 3 dogs love various vitaminenriched tasty pastes available kong although loved chew kongs found really got inserted paste cavity all go nuts put pasty kongs e
A3RB2JGKW134ZO	A3RB2JGKW134ZO	my new puppy could less interested treat even mealtime walks away not even sniff ive resort pieces hot dog i know healthy treats far training rely praise	my new puppy could less interested treats even mealtime walks away not even sniff ive resort pieces hot dog i know healthy treats far training rely praise

Figure 3 Result of Nearest Pairs

Problem 6

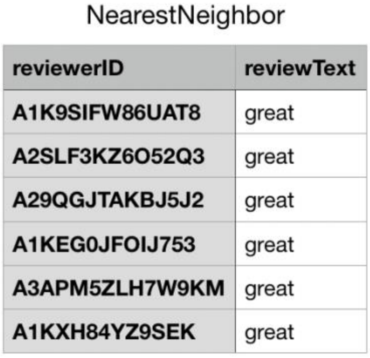
Statement: Find the nearest neighbor of a given input review

Approach:

- Pre-processing the review according to the required format
- Create shingle representation for this review as a binary matrix consisting 0 or 1
- Min hashing and local sensitivity hashing are applied to shingle representation of this review and the band matrix is obtained. The values of a1, b1, r, a2, b2, p used are the same as the signature matrix and the band matrix of other reviews
- The band values are checked against the document in the data set and if at least one bad match, the new user and the user in that column of the data set are similar
- If there are not band value match, return None
- If there are band value match, we determine the Jaccard distance to determine which is the nearest neighbour with minimum Jaccard distance

Result:

If we input a new review as ' i . me , , great! so: |', it will return the index of the nearest neighbour as [133 57364 85048 112897 119030 151045], and store the review text in a csv file.



NearestNeighbor

reviewerID	reviewText
A1K9SIFW86UAT8	great
A2SLF3KZ6O52Q3	great
A29QGJTAKBJ5J2	great
A1KEG0JFOIJ753	great
A3APM5ZLH7W9KM	great
A1KXH84YZ9SEK	great

Figure 4 Result of the Function to Nearest Neighbor

Problem 7

Statement: Briefly discuss the complexity of your implementation and how it is better than the naive implementations.

Answer:

The naïve implementation requires $O(N_2)$ comparison, so the time complexity is $O(n_2)$.
In the min hashing method and local sensitivity hashing, the hash function can be computed in time O(M) of the given sets, where M is the number of hash function. Specifically, for set size N documents, the min hashing method takes O(M*N) time. Assuming N>>M such the number of documents is always large, we could get O(*n*) time to maintain the queue of minimum hash values.
Thus, min hashing and local sensitivity hashing requires O(*n*) time, while the naïve way requires $O(n_2)$ time.