Name : **Pawan Suresh Bathe** Roll No.: **CS18M519**

CS6745: Mining Massive DataSets

Tutorial 5

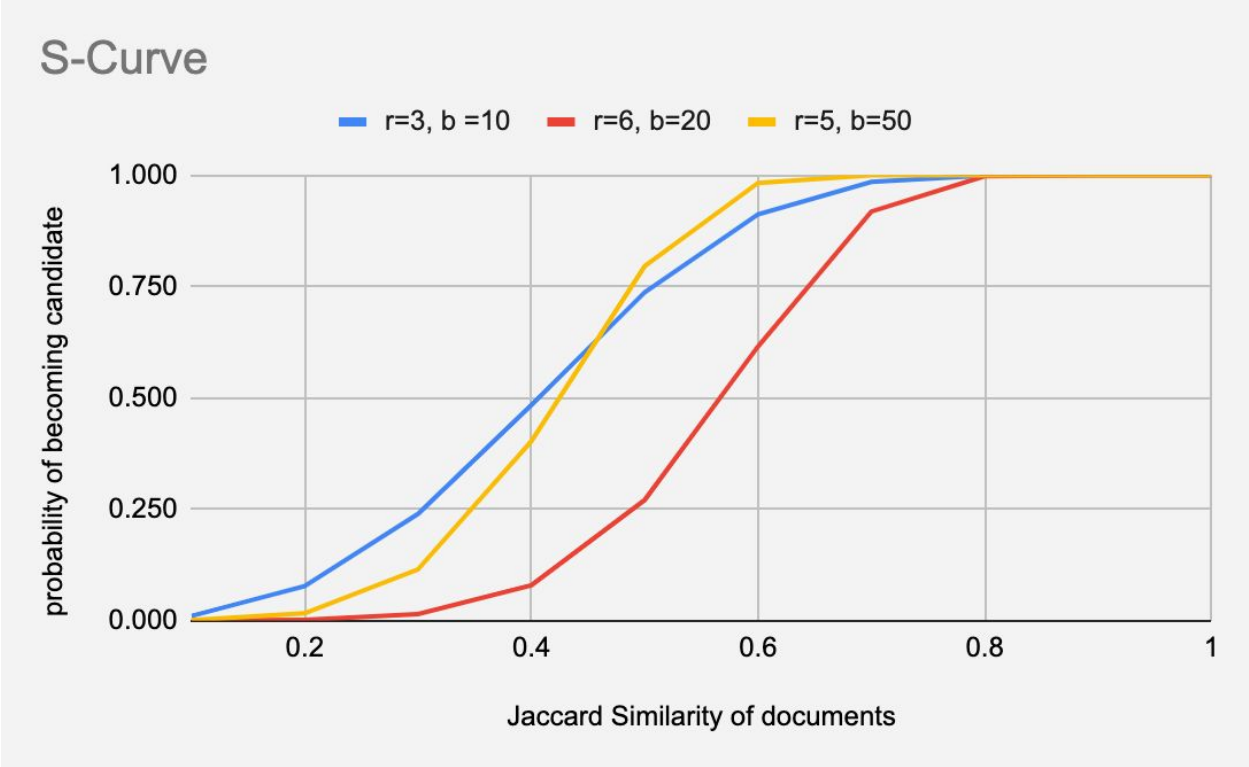
October 8, 2019

- Write your name and roll number in the space provided
- Be neat, and use the space judiciously.
- Rough sheets won't be evaluated.

1. (3 marks) Evaluate the S-curve $1 - (1 - s^r)^b$ for s = 0.1,0.2, ...,0.9, for the following values of r and b:

- **r = 3 and b = 10**
- **r = 6 and b = 20**
- **r = 5 and b = 50**

Answer:



s	1-(1-s^r)^b		
	r=3, b =10	r=6, b=20	r=5, b=50
0.1	0.010	0.000	0.000
0.2	0.077	0.001	0.016
0.3	0.239	0.014	0.115
0.4	0.484	0.079	0.402
0.5	0.737	0.270	0.796
0.6	0.912	0.615	0.983
0.7	0.985	0.918	1.000
0.8	0.999	0.998	1.000
0.9	1.000	1.000	1.000
1	1.000	1.000	1.000

2. (7 marks) Suppose we wish to implement LSH by MapReduce. Specifically, assume chunks of the signature matrix consist of columns, and elements are key-value pairs where the key is the column number and the value is the signature itself (i.e. a vector of values).

(a) Show how to produce the buckets for all the bands as output of a single MapReduce process. Hint: Remember that a Map function can produce several key-value pairs from a single element.

Answer: We can make use of grouping the elements by band id and combining them together in groups corresponding to each band.

Map Function:

Map each key element i.e column number and it's signature to its band id as key and column number as a value

Input : <column number, signature>

Output: <band id, column number >

Reduce Function:

Combine the output of map function, sort all elements by bucket id i.e key and aggregate the elements having the same key i.e same band id elements into one list or vector. Such that band id is key and aggregated list/vector of column number in same band id as it's value

Input : <band id, column number >

Output: <band id, column vector of columns having same band_id >

(b) Show how another MapReduce process can convert the output of (a) to a list of pairs that need to be compared. Specifically, for each column i, there should be a list of those columns j>i with which i needs to be compared.

Answer:

Map Function :

Take an output (a) which in the format <band id, column vector> and map it to key-value pair such that key is two element key **<band_id , id of column vector>** and its value is **column vector with corresponding column vector id**

Group all elements having same key i.e <band_id , id of column vector> pair, and it's value as a combination of all column vectors with same column vector id.

Input: <band id, column vector>

Output: <(band_id , id of column vector), list(all column vectors with same column vector id and band_id)>

Reduce Function :

For each key. < band id, and column vector id > pair , and each column i in the column vector, corresponding to column vector id output the candidate pairs as other elements i.e column number j from that vector where j>i against which i needs to be compared so that no duplication of column id happens as we increase column number only remaining column number would be compared against it.

Input: <band_id, list(all column vectors with corresponding column vector id)>

Output: <column number, list of column numbers to be compared>