Heidelberg University
Winter semester 2017/2018

Group Parallel and Distributed Systems (PVS)
Artur Andrzejak, Diego Costa

Problem Set 6 for lecture Mining Massive Datasets

Due December 4, 2017, 11:59 pm

Exercise 1 (2 points)

The Jaccard similarity can be applied to sets of elements. Sometimes, documents (or other objects) may be represented as multi-sets/bags rather than sets. In a multi-set, an element can be a member more than once, whereas a set can only hold each element at most once. Try to define a similarity metric for multi-sets. This metric should take exactly the same values as Jaccard similarity in the special case where both multi-sets are in fact sets.

Exercise 2 (6 points)

Recall the concept of shingling documents which makes it possible to represent text documents as sets. Write a (scalable!) implementation in Spark (either Python or Java) which computes for a collection of documents the corresponding sets of shingles.

Here, we mean shingles of characters (as opposed to e.g., shingles of words). The records in a RDD should be the lines of a text. Be careful to take into consideration both line breaks and the hyphenation at the of lines. I.e., in the following example, with $k = 9$, the shingles should include: "uished bu", "is possib", "hen he st", "mpossible", "robably w". Other special cases needed for proper shingling may be ignored.

> When a distinguished but elderly scientist states that something is possible, he is almost certainly right. When he states that something is impossible, he is very probably wrong.

Run your implementation on a set of 10 documents. One of them should be *The White Spark* by Orville Livingston Leach[1], in order for the solutions to be comparable. Output the number of distinct shingles for a run with $k = 5$ and $k = 9$, respectively.

Exercise 3 (3 points)

Fig. 1 shows a table (or matrix) representing four sets S_1, S_2, S_3 and S_4 (subsets of $\{0, 1, 2, 3, 4, 5\}$).

a) Compute the minhash signature for each set using the following three hash functions:
$h_1(x) = 2x + 1 \mod 6$
$h_2(x) = 3x + 2 \mod 6$
$h_3(x) = 5x + 2 \mod 6$

b) Which of these hash functions are true permutations? What collisions do occur in the other hash functions? Name the corresponding inputs and outputs.

[1] http://www.gutenberg.org/cache/epub/44016/pg44016.txt

c) Compare the similarity of the minhash signatures against the corresponding Jaccard similarities, for each of the $\binom{4}{2} = 6$ pairs of columns.

Element	S_1	S_2	S_3	S_4
0	0	1	0	1
1	0	1	0	0
2	1	0	0	1
3	0	0	1	0
4	0	0	1	1
5	1	0	0	0

Figure 1: Matrix for exercise 3

Exercise 4 (1 point)

Prove that if the Jaccard similarity of two columns is 0, then minhashing always gives a correct estimate of the Jaccard similarity.

Exercise 5 (2 points)

Recall the property which relates the Jaccard similarity to the probability of minhashing to equal values, for the particular case of Fig.2.

(a) Compute the Jaccard similarity of S_1 and S_2 in Fig. 2.

(b) Find out the fraction of the 120 permutations of the rows that make the two columns hash to the same value, without enumerating each permutation one by one. Explain how you derive your result.

Element	S_1	S_2
a	0	1
b	1	0
c	0	1
d	1	1
e	1	0

Figure 2: Matrix for exercise 5

Exercise 6 (2 points)

Give two examples for hash functions as alternatives to the arithmetic remainder function `mod`. See this video[2] which was created in the weeks after Adobe's password breach in October 2013 and explain the connection of hash functions to password security. What is a rainbow table? What is salt in this context?

[2] Computerphile: How NOT to store passwords! https://www.youtube.com/watch?v=8ZtInClXe1Q