CSC 391/691 – Spring 2014
Test 1 – Take Home
Due – 4pm, Friday, 3/7/14

Name _____Shuowen Wei___

THIS WORK MUST BE DONE BY YOU AND YOU ALONE!

No Exceptions.

If you use printed or Internet resources, other than the text, you must indicate the source appropriately.

I will try to clarify any question that you don't understand, but otherwise do your own work. I will be out of the office beginning 3/5/14 but you can send your questions via e-mail. Submit your responses on paper or electronically through the Sakai Assignment link. Paper submissions should be given to Dr. Turkett.

I. (10 pts) Let's think about using a MapReduce (MR) environment to compute TF-IDF for terms in a very large collection of documents (corpus). We'll do this through a cascaded sequence of MR steps. Your job is to fill in the blanks. NOTE: In practice, some of these steps would be combined.

[Reminder: Here are the values we need to compute the TF-IDF for each word w_i in document d_j which is in a corpus of *D* documents.

D – the number of documents in the corpus. For the sake of simplicity, we'll just assume that this value is known globally.

$f_{i,j}$ – the number of times that word w_i appears in document d_j

$maxk_j$ – the maximum number of times any individual word appears in document d_j

n_i – the number of documents in which word w_i appears]

Step 1: Determine word count in individual documents.

Input: collection of (docID, contents of document) pairs distributed across mappers

What should a mapper emit? ((word, docID), 1)

What does the Reducer do? for every key-value pair ((word, docID), 1), sum up all 1s of the mapper output, set it to be f, and output ((word, docID), f).

Reducer Output: list of key-value pairs ((word, docID), f) where f is the number of times *word* appears in *docID.*

Step 2: Determine the count of the most frequent word in each document

Input: list of key-value pairs ((word, docID), f)

What should a mapper emit? (docID, (word, f))

What does the Reducer do? for every docID, find the maximum value of f, set it to *maxk* and associate it with each input key-value pairs , output ((word, docID), (f, *maxk*)).

Reducer Output: list of key-value pairs ((word, docID), (f, maxk)) where *maxk* is the largest f in *docID*

Step 3: Determine the number of documents in which *word* appears

Input: list of key-value pairs ((word, docID), (f, maxk))

What should a mapper emit? (word, (docID, f, *max*k, 1))

What does the Reducer do? for every key-value pair (word, (docID, f, *max*k,1)), sum up all 1s of the mapper output, set it to be n, and associate it every input key-value pairs, output ((word, docID), (f, *max*k, n).

Reducer Output: list of key-value pairs ((word, docID), (f, maxk, n) where *n* is the number of documents in which word appears

Step 4: Calculate TF-IDF for each word relative to each document. Assume that D, the number of documents in the corpus, is known.

Input: list of key-value pairs ((word, docID), (f, *max*k, n))

What should a mapper emit? ((word, docID), (f, *max*k, n)), the identical function

What does the Reducer do? for each key-value pairs input, since D is known, compute $\frac{f}{maxk}\log_2\frac{D}{n}$ and set it TF-IDF, then associate it with each (word, docID), output ((word, docID), TF-IDF)

Reducer Output: list of key-value pairs ((word, docID), TF-IDF)

II. (10 pts) Suppose the input to a MR operation consists of integer keys (the values are not important, we indicate them with an underscore). The Map function takes an integer *i* and produces the list of pairs (*p*,*i*) such that *p* is a prime divisor of *i*. For example, map((12,_)) emits [(2,12), (3,12)]. Note that (2, 12) is not repeated even though the factorization of 12 is 2 * 2 * 3.

The Reduce function is addition. That is, Reduce (p, [i_1, i_2, ...,i_k]) is (p,i_1+i_2+...+i_k).

What is the output, if the input is the set {(15, _), (21, _), (24, _), (30, _), (49, _)}.

Output of Mapper:

Map({(15, _), (21, _), (24, _), (30, _), (49, _)}) = { [(3,15), (5,15)],
[(3,21), (7,21)],
[(2,24), (3,24)],
[(2,30), (3,30), (5,30)],
[(7,49)], }

Output of Reducer:

Reduce({(2, [24, 30]), (3, [15, 21, 24, 30]), (5, [15, 30]), (7, [21, 49])}) = {(2, 54), (3,90), (5, 45), (7, 70)}

III. (10 pts) Apply the matrix-vector multiplication approach described in Section 2.3.1 to the matrix and vector:

1	2	3	4		1
5	6	7	8		2
9	10	11	12		3
13	14	15	16		4

Show the output of the Map function? You don't need to complete the multiplication.

(1, 1) (1, 4) (1, 9) (1, 16)
(2, 5) (2, 12) (2, 21) (2, 32)
(3, 9) (3, 20) (3, 33) (3, 48)
(4, 13) (4, 28) (4, 45) (4, 64)

IV. (12 pts) Suppose we use the two-stage algorithm of Section 2.3.9 to compute the product of matrices M and N. Let M have x rows and y columns, while N has y rows and z columns. As a function of x, y, and z, express the answers to the following questions:

1. The output of the first Map function has how many different keys? How many key-value pairs are there with each key? How many key-value pairs are there in all?
2. The output of the first Reduce function has how many keys? What is the length of the value (a list) associated with each key?
3. The output of the second Map function has how many different keys? How many key-value pairs are there with each key? How many key-value pairs are there in all?

1. The first Map function has y different keys. There are (x+z) key-value pairs with each key. And there are (xy+yz) key-value pairs in total.
2. The output of the first Reduce function has xz keys, and there are y values associated with each key.
3. The output of the second Map function has xz different keys, and there are y key-value pairs with each key, and there are xyz key-value pair in all.

V. (15 pts) Here are five bit vectors in a 10-dimensional space:

V1: 1111000000 V2: 0100100101 V3: 0000011110 V4: 0111111111 V5: 1011111111

a. Suppose cos(x,y) denotes the similarity of vectors x and y under the cosine similarity measure. Compute all the pairwise similarities among t, u,v, and w. Which two vectors are the most similar under this measure?
b. Compute the Jaccard distance (not Jaccard "measure") between each pair of vectors. Which two vectors are the most similar under this measure?
c. Compute the Manhattan distance (L_1 norm) between each pair of vectors. Which two vectors are the most similar under this measure?

a. $\cos(V1, V2) = \frac{1}{4}$ $\cos(V1, V3) = \frac{0}{4}$ $\cos(V1, V4) = \frac{3}{6}$ $\cos(V1, V5) = \frac{3}{6}$

$\cos(V2, V3) = \frac{1}{4}$ $\cos(V2, V4) = \frac{4}{6}$ $\cos(V2, V5) = \frac{3}{6}$

$\cos(V3, V4) = \frac{4}{6}$ $\cos(V3, V5) = \frac{4}{6}$

$\cos(V4, V5) = \frac{8}{9}$

Thus, $V4$ and $V5$ are most similar under this measure.

b. $d(V1, V2) = 1 - \frac{1}{7}$ $d(V1, V3) = 1 - 0$ $d(V1, V4) = 1 - \frac{3}{10}$ $d(V1, V5) = 1 - \frac{3}{10}$

$d(V2, V3) = 1 - \frac{1}{7}$ $d(V2, V4) = 1 - \frac{4}{9}$ $d(V2, V5) = 1 - \frac{3}{10}$

$d(V3, V4) = 1 - \frac{4}{9}$ $d(V3, V5) = 1 - \frac{4}{9}$

$d(V4, V5) = 1 - \frac{8}{10}$

Thus, $V4$ and $V5$ are most similar under Jaccard distance since their distance is the smallest.

c. $d_1(V1, V2) = 6$ $d_1(V1, V3) = 8$ $d_1(V1, V4) = 7$ $d_1(V1, V5) = 7$
$d_1(V2, V3) = 6$ $d_1(V2, V4) = 5$ $d_1(V2, V5) = 7$
$d_1(V3, V4) = 5$ $d_1(V3, V5) = 5$
$d_1(V4, V5) = 2$

Thus, $V4$ and $V5$ are most similar under Manhattan distance since their distance is the smallest.

VI. (8 pts) Consider the following matrix:

	C1	**C2**	**C3**	**C4**
R1	0	1	1	0
R2	1	0	1	1
R3	0	1	0	1
R4	0	0	1	0
R5	1	0	1	0
R6	0	1	0	0

Compute the Jaccard similarity between each pair of columns.

$\text{sim}(C1, C2) = \frac{0}{5}$ $\text{sim}(C1, C3) = \frac{2}{4}$ $\text{sim}(C1, C4) = \frac{1}{3}$
$\text{sim}(C2, C3) = \frac{1}{6}$ $\text{sim}(C2, C4) = \frac{1}{4}$
$\text{sim}(C3, C4) = \frac{1}{5}$

VII. (10 pts) Using the same matrix as in the previous question, perform a minhashing of the data, with the order of rows: R4, R6, R1, R3, R5, R2.

	C1	**C2**	**C3**	**C4**
R4	0	0	1	0
R6	0	1	0	0
R1	0	1	1	0
R3	0	1	0	1
R5	1	0	1	0
R2	1	0	1	1

Thus, we have

C1	**C2**	**C3**	**C4**
2	1	1	2
5	2	1	4

Then the Jaccard similarity between each pair of columns

$\text{sim}(C1, C2) = \frac{0}{2}$ $\text{sim}(C1, C3) = \frac{0}{2}$ $\text{sim}(C1, C4) = \frac{1}{2}$
$\text{sim}(C2, C3) = \frac{1}{2}$ $\text{sim}(C2, C4) = \frac{0}{2}$
$\text{sim}(C3, C4) = \frac{0}{2}$

VIII. (15 pts) Find the set of 2-shingles for the "document": ABRACADABRA
and also for the "document": BRICABRAC

Answer the following questions:

a. How many 2-shingles does ABRACADABRA have?
b. How many 2-shingles does BRICABRAC have?
c. How many 2-shingles do they have in common?
d. What is the Jaccard similarity between the two documents"?

a. {AB, BR, RA, AC, CA, AD, DA}, there are 7 2-shingles.
b. {BR, RI, IC, CA, AB, RA, AC}, there are 7 2-shingles.
c. There are 5 in common
d The Jaccard similarity is $\frac{5}{9}$.

IX. (**UNDERGRADS ONLY**, 10 pts) Here are four "documents," each consisting of a single sentence (ignore caps and punctuation):

Pussycat, pussycat, where have you been?
I've been to London to visit the Queen.
Pussycat, pussycat, what did you there?
I frightened a little mouse under her chair.

Compute the term frequency (TF) for each of the following words: "pussycat", "chair", "queen", and "mouse" in each of the four “documents”. Also, compute the inverse document frequency (IDF) for each of these words. Generate a table to display the TF-IDF measure of each word within each “document”.

(**GRAD STUDENTS ONLY**, 10 pts)

We wish to take the join R(A,B) |><| S(B,C) |><| T(A,C) as a single map-reduce process, in a way that minimizes the communication cost. We shall use 512 Reduce tasks, and the sizes of relations R, S, and T are 2^20 = 1,048,576, 2^17 = 131,072, and 2^14 = 16,384, respectively. Using the technique of Section 2.5.3, compute the number of buckets into which each of the attributes A, B, and C are to be hashed. Then, determine the number of times each tuple of R, S, and T is replicated by the Map function.

By the conclusion of section 2.5.3, to minimum communication cost is $\left(r+s+t+2\sqrt{ktr}\right)$, here we have $k=2^9, s, r, t \in \{2^{20}, 2^{17}, 2^{14}\}$. Thus, it’s clear that we should let $s=2^{20}$ and $r, t \in \{2^{17}, 2^{14}\}$. Thus, (1)

the number of buckets into which attribute A is to be hashed is $\sqrt{2^9 \frac{2^{14}}{2^{17}}} = 2^3$

the number of buckets into which attribute B is to be hashed is $\sqrt{2^9 \frac{2^{17}}{2^{14}}} = 2^6$

the number of buckets into which attribute C is to be hashed is 2^9

So, (2)
each tuple of R is replicated 1 times by the Map function

each tuple of S is replicated 2^3 times by the Map function
each tuple of T is replicated 2^6 times by the Map function