WA 2 Panagiotis Paltsokas std163861

Topic 2: MapReduce

(15 total points) A bi-gram is called a **lexicographically ordered** pair of words occurring in successive places. The purpose of this topic is to count the number of occurrences of all-bi-grams using MapReduce. Assume the employment of 2 Mappers and 2 Reducers.

The first mapper is presented two documents:

"one four two three" and "two three one".

The second mapper is also presented two documents:

"one three one four" and "four two four one".

Suppose that after the mapping phase, the partitioner sends to the first reducer all keys whose first letter belongs in the range [a-m]; the rest are sent to the second reducer.

(a) (10 points) For your answer, fill in the following table. That is, for each input, you should present the output emitted in both mappers and reducers, if any. In your answer, sort the words of each individual bi-gram in lexicographical ascending order.

pairs lexicogr.

Mapper #1	**Mapper #2**
Input: one four two three (4,1),(4,2),(3,2)	**Input:** one three one four (1,3),(1,3),(4,1)
Emits: ((four, one), 1), ? ((four,two),1) ((three,two),1)	**Emits:** ? ((one,three),1),((one,three),1) ((four,one),1)
Input: two three one (3,2),(1,3)	**Input:** four two four one (4,2),(4,2),(4,1)
Emits: ((three,two),1), ((one,three),1)	**Emits:** ((four,two),1), ((four,two),1) ((four,one),1)
Shuffle and Sort	
Reducer #1 [a-h]	**Reducer #2** [i-z]
Input: ? ((four,one),(1,1,1)) ((four,two),(1,1,1))	**Input:** ?((two,three),(1,1)) ((one,three),(1,1,1))
Emits: ? ((four,one),3) ((four,two),3)	**Emits:** ?((three,two),2) ((one,three),3)

(b) (5 points) Solve the same problem by using In-Mapper Combiners.

Using In-Mapper Combiners, the mapper processes all its inputs and emits the aggregated results in the form (key, count)

Mapper #1	**Mapper #2**
Input: one four two three (4,1),(4,2),(3,2)	**Input:** one three one four (1,3),(1,3),(4,1)
Emits: ((four, one), 1), ? emit aggregated bi-grams ((four,two),1),((three,two),1)	**Emits:** ? ((one,three),2) ((four,one),1)
Input: two three one (3,2),(1,3)	**Input:** four two four one (4,2),(4,2),(4,1)
Emits: ? ((three,two),1),((one,three),1) ((four,one),1) ((four,two),1)	**Emits:** ?((four,two),2),((four,one),2) ((one,three),2)
Shuffle and Sort	
Reducer #1 [a-m]	**Reducer #2** [n-z]
Input: ? ((four,one),(1,2)) ((four,two),(1,2))	**Input:** ?((one,three),(2,1)) ((three,two),2)
Emits: ?((four,one),3) ((four,two),3)	**Emits:** ?((one,three),3) ((three,two),2)

the bi-grams are aggregated across both inputs

Topic 3: Locality Sensitive Hashing

(15 total points) Assume the following documents being represented as sets of shingles,

D1 = {a,b,c,d,e,g,h,j,k,m,n,o,p,s,t,v,w,y} $|D_1| = 18$

D2 = {a,b,c,f,g,h,j,k,s,t,u,w,x,z} $|D_2| = 14$

D3 = {e,p,q,u,x,z} $|D_3| = 6$

D4 = {a,b,d,e,f,g,i,m,q,r,t,u,y} $|D_4| = 13$

with each lowercase Latin letter representing a shingle that appears in the document. The universal set of shingles comprises all the lowercase Latin characters (26 shingles), U="a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,q,r,s,t,u,v,w,x,y,z".

(a) (4 points) Compute the Jaccard similarity between each pair of documents and fill in **the missing values** (indicated by **?**) in the following table. Report the results with 3 decimal places of accuracy (do not round up or down the numbers).

	D1	D2	D3	D4
D1	1	0.454	0.090	0,347
D2	0.454	1	0,176	0,285
D3	0.090	0,176	1	0,187
D4	0,347	0,285	0,187	1

You are given an illustration of how the two given values are computed. Do the same for the rest of them.

D1 = "abcde ghjkmnopst vw y"
D2 = "abc fghjk stu wx z"

D1 = "abcdeghjkmnop st vw y"
D3 = " e pq u x z"

(a)

D1 = {a,b,c,d,e,g,h,j,k,m,n,o,p,s,t,v,w,y}
D4 = {a,b,d,e,f,g,i,m,q,r,t,u,y}

$$Jaccard(D_1, D_4) = \frac{|D_1 \cap D_4|}{|D_1 \cup D_4|} = \frac{8}{23} = 0,347$$

D2 = {a,b,c,f,g,h,j,k,s,t,u,w,x,z}
D4 = {a,b,d,e,f,g,i,m,q,r,t,u,y}

$$Jaccard(D_2, D_4) = \frac{6}{21} = 0,285$$

D3 = {e,p,q,u,x,z}
D4 = {a,b,d,e,f,g,i,m,q,r,t,u,y}

$$Jaccard(D_3, D_4) = \frac{3}{16} = 0,187$$

D2 = {a,b,c,f,g,h,j,k,s,t,u,w,x,z}
D3 = {e,p,q,u,x,z}

$$Jaccard(D_2, D_3) = \frac{3}{17} = 0,176$$

(b) (4 points) Compute signatures for the documents using the following six permutations of the universal set of shingles (where 0 stands for element "a" and 25 stands for element "z") and fill in the missing values (indicated by **?**) in the following signature matrix.

- Permutation 1: 19, 13, 8, 3, 20, 17, 25, 24, 21, 12, 9, 1, 22, 14, 10, 11, 5, 23, 0, 18, 2, 7, 4, 15, 16, 6
- Permutation 2: 16, 25, 10, 5, 9, 2, 11, 18, 17, 15, 7, 24, 6, 19, 0, 12, 20, 3, 14, 4, 21, 13, 1, 23, 22, 8
- Permutation 3: 7, 24, 14, 13, 8, 1, 9, 15, 2, 0, 16, 10, 20, 3, 12, 6, 18, 4, 17, 19, 11, 23, 5, 22, 25, 21
- Permutation 4: 10, 0, 16, 11, 5, 13, 7, 12, 4, 9, 17, 14, 24, 3, 1, 2, 8, 25, 21, 20, 19, 15, 6, 22, 23, 18
- Permutation 5: 12, 25, 2, 11, 13, 16, 18, 3, 22, 7, 8, 15, 4, 24, 19, 14, 1, 10, 6, 9, 20, 21, 0, 23, 5, 17
- Permutation 6: 15, 16, 20, 9, 19, 3, 14, 2, 6, 13, 12, 10, 0, 7, 11, 8, 23, 21, 25, 1, 18, 17, 4, 5, 24, 22

The permutations are interpreted as the new order of the shingles. So, for example, in permutation 1 the first shingle is 19 that corresponds to element "t", the second shingle is 13 that corresponds to "n", etc.

	D1	D2	D3	D4
Permutation 1	t	t	?	?
Permutation 2	k	z	?	?
Permutation 3	h	h	?	?
Permutation 4	k	k	?	?
Permutation 5	m	z	?	?
Permutation 6	p	u	?	?

Estimate the Jaccard similarity for the pairs of documents by computing the similarities of the pairs of signatures (fraction of common rows to the total number of rows) and fill in the missing values (indicated by **?**) in the next table. Report the results with 3 decimal places of accuracy (do not round up or down the numbers).

	D1	D2	D3	D4
D1	1	0,500	?	?
D2	0,500	1	?	?
D3	?	?	1	?
D4	?	?	?	1

(b)

Convert each documents shingles to indices.

$D_1 = \{0, 1, 2, 3, 4, 6, 7, 9, 10, 12, 13, 14, 15, 18, 19, 21, 22, 24\}$

$D_2 = \{0, 1, 2, 5, 6, 7, 9, 10, 18, 19, 20, 22, 23, 25\}$

$D_3 = \{4, 15, 16, 20, 23, 25\}$

$D_4 = \{0, 1, 3, 4, 5, 6, 8, 12, 16, 17, 19, 20, 24\}$

- Permutation 1: 19, 13, 8, 3, 20, 17, 25, 24, 21, 12, 9, 1, 22, 14, 10, 11, 5, 23, 0, 18, 2, 7, 4, 15, 16, 6

	D1	D2	D3	D4
Permutation 1	t	t	u	t

- Permutation 2: 16, 25, 10, 5, 9, 2, 11, 18, 17, 15, 7, 24, 6, 19, 0, 12, 20, 3, 14, 4, 21, 13, 1, 23, 22, 8

Permutation 2	k	z	q	q

Likewise

- Permutation 3: 7, 24, 14, 13, 8, 1, 9, 15, 2, 0, 16, 10, 20, 3, 12, 6, 18, 4, 17, 19, 11, 23, 5, 22, 25, 21

Permutation 3	h	h	p	y

- Permutation 4: 10, 0, 16, 11, 5, 13, 7, 12, 4, 9, 17, 14, 24, 3, 1, 2, 8, 25, 21, 20, 19, 15, 6, 22, 23, 18

Permutation 4	k	k	q	x

- Permutation 5: 12, 25, 2, 11, 13, 16, 18, 3, 22, 7, 8, 15, 4, 24, 19, 14, 1, 10, 6, 9, 20, 21, 0, 23, 5, 17

Permutation 5	m	z	z	w

- Permutation 6: 15, 16, 20, 9, 19, 3, 14, 2, 6, 13, 12, 10, 0, 7, 11, 8, 23, 21, 25, 1, 18, 17, 4, 5, 24, 22

Permutation 6	p	u	p	q

Ultimately

	D1	D2	D3	D4
Permutation 1	t	t	u	t
Permutation 2	k	z	q	q
Permutation 3	h	h	p	y
Permutation 4	k	k	q	x
Permutation 5	m	z	z	w
Permutation 6	p	u	p	q

with similarities

$D_1 - D_2$: 3/6 = 0.5
$D_1 - D_3$: 1/6 = 0,166
$D_1 - D_4$: 2/6 = 0,333
$D_2 - D_3$: 1/6 = 0,166
$D_2 - D_4$: 1/6 = 0,166
$D_3 - D_4$: 1/6 = 0,166

	D1	D2	D3	D4
D1	1	0,500	0,166	0,333
D2	0,500	1	0,166	0,166
D3	0,166	0,166	1	0,166
D4	0,333	0,166	0,166	1

(c) (4 points) Using the following six (6) hash functions, $h_1, h_2, h_3, h_4, h_5, h_6$:

$$h_1 = (11 * r + 1)\ mod\ 26$$
$$h_2 = (7 * r + 3)\ mod\ 26$$
$$h_3 = (21 * r + 5)\ mod\ 26$$
$$h_4 = (19 * r + 7)\ mod\ 26$$
$$h_5 = (23 * r + 11)\ mod\ 26$$
$$h_6 = (29 * r + 13)\ mod\ 26$$

where r stands for the row of each shingle in the characteristic matrix, ranging from 0 for "a" to 25 for "z". Compute the signatures of the documents and fill in the **missing values** (indicated by ?) in the following signature matrix.

	D1	D2	D3	D4
h_1	0	0	10	1
h_2	0	0	4	3
h_3	0	0	3	0
h_4	0	0	2	0
h_5	0	3	3	1
h_6	0	1	4	2

Estimate the Jaccard similarity of the documents by computing the similarities of the new signatures (fraction of common rows to the total number of rows) and fill in the missing values (**indicated by ?**) in the following table. Report the results with 3 decimal places of accuracy (do not round up or down the numbers).

	D1	D2	D3	D4
D1	1	0,666	0	0,333
D2	0,666	1	0,166	0,333
D3	0	0,166	1	0
D4	0,333	0,333	0	1

(d) (3 points) Assuming three bands of two rows each (b=3, r=2) for the signatures of question (b), find the candidate pairs of similar documents using locality-sensitive hashing.
Do the same with two bands of three rows each (b=2, r=3).

Answer:

Candidate pairs for case (b=3, r=2): ?

Candidate pairs for case (b=2, r=3): ?

	D1	D2	D3	D4
Permutation 1	t	t	**u**	**t**
Permutation 2	k	z	**q**	**q**
Permutation 3	h	h	**p**	**y**
Permutation 4	k	k	**q**	**a**
Permutation 5	m	z	**z**	**m**
Permutation 6	p	u	**p**	**q**

D3 = {e,p,q,u,x,z} → 4 15 16 20 23 25 — $|D_3| = 6$

D4 = {a,b,d,e,f,g,i,m,q,r,t,u,y} → 0 1 3 4 5 6 8 12 16 17 19 20 24 — $|D_4| = 13$

for D3

- $h_1 = (11 \cdot r + 1) \bmod 26$

$h_1(4) = 45 \bmod 26 = 19$

$h_1(15) = 166 \bmod 26 = 10$ → min value

$h_1(16) = 177 \bmod 26 = 21$

$h_2(20) = 221 \bmod 26 = 13$

$h_1(23) = 254 \bmod 26 = 20$

$h_1(25) = 276 \bmod 26 = 16$

↓ we repeat the same process for $h_2, h_3, \dots, h_6$ selecting the minimum value each time

then, we do the same for D4

for the Jaccard similarity, we work just like before

$D_1 - D_3$: 0/6 = 0 $D_3 - D_4$: 0/6 = 0

$D_1 - D_4$: 2/6 = 0,333

$D_2 - D_3$: 1/6 = 0,166

$D_2 - D_4$: 2/6 = 0,333

3 bands of 2 rows — b=3, r=2

$(h_1, h_2) \rightarrow \begin{bmatrix} t & t & u & t \\ k & z & q & q \end{bmatrix}$ $D_1 = (t,k)$ $D_2 = (t,z)$ $D_3 = (u,q)$ $D_4 = (t,q)$ — no match

$(h_3, h_4) \rightarrow \begin{bmatrix} h & h & p & y \\ k & k & q & a \end{bmatrix}$ $D_1 = (h,k)$ $D_2 = (h,k)$ $D_3 = (p,y)$ $D_4 = (q,a)$ — $D_1 = D_2$

$(h_5, h_6) \rightarrow \begin{bmatrix} m & z & z & m \\ p & u & p & q \end{bmatrix}$ $D_1 = (m,p)$ $D_2 = (z,u)$ $D_3 = (z,p)$ $D_4 = (u,q)$ — no match

2 bands of 3 rows — b=2, r=3

$(h_1, h_2, h_3) \rightarrow \begin{bmatrix} t & t & u & t \\ k & z & q & q \\ h & h & p & y \end{bmatrix}$ $D_1 = (t,k,h)$ $D_2 = (t,z,h)$ $D_3 = (u,q,p)$ $D_4 = (t,q,y)$ — no match

$(h_4, h_5, h_6) \rightarrow \begin{bmatrix} k & k & q & a \\ m & z & z & m \\ p & u & p & q \end{bmatrix}$ $D_1 = (k,m,p)$ $D_2 = (k,z,u)$ $D_3 = (q,z,p)$ $D_4 = (a,m,q)$ — no match.

Topic 4: Data Streams

(15 total points) Consider the elements in the sequence <8, 12, 5, 7, 3, 15> that come as a stream. First you will construct a Bloom filter to lookup unseen values from this steam and then you will use the Flajolet-Martin algorithm to estimate the number of distinct elements in the sequence.

(a) (**4 points**) Construct a Bloom filter with N=10 bits (starting from position 0 to position 9), using the hash functions :

$h_1(x) = (3 * x + 7) \bmod N$

$h_2(x) = (5 * x + 11) \bmod N$

Fill in the missing information (**?**) and define the Bloom Filter bit array.

Answer:

Stream Element (x)	Binary Representation of $h_1(x)$	Binary Representation of $h_2(x)$
12	0011	0001
7	1000	0110
15	0010	0110

(a) $\langle 8, 12, 5, 7, 3, 15 \rangle$ $N=10$ x mod 10 returns the last digit of an integer

$h_2(12) = 71 \bmod 10 = 1 \hookrightarrow 0001$

$h_1(7) = 28 \bmod 10 = 8 \hookrightarrow 1000$

$h_1(15) = 52 \bmod 10 = 2 \hookrightarrow 0010$

$h_2(15) = 86 \bmod 10 = 6 \hookrightarrow 0110$

now lets define the bloom filter array.

$N=10$. Start with $[0,0,0,0,0,0,0,0,0,0]$

- $x=8$ $h_1(8)=1$, $h_2(8)=1$ $\rightarrow [0,1,0,0,0,0,0,0,0,0]$
- $x=12$ $h_1(12)=3$, $h_2(12)=1$ $\rightarrow [0,1,0,1,0,0,0,0,0,0]$
- $x=5$ $h_1(5)=2$, $h_2(5)=6$ $\rightarrow [0,1,1,1,0,0,1,0,0,0]$
- $x=7$ $h_1(7)=8$, $h_2(7)=6$ $\rightarrow [0,1,1,1,0,0,1,0,1,0]$
- $x=3$ $h_1(3)=6$, $h_2(3)=6$ $\rightarrow [0,1,1,1,0,0,1,0,1,0]$
- $x=15$ $h_1(15)=2$, $h_2(15)=6$ $\rightarrow$ final bit array $[0,1,1,1,0,0,1,0,1,0]$

(b) (**4 points**) Using the generated Bloom filter, how likely is to see the following values: 3, 9, 15, 20. Please select the "Might Be" if the value could be seen, otherwise "Is Definitely Not", following the calculation of the values of $h_1(3)$, $h_2(3)$ etc.

Answer:

Number 3 has $h_1(3)$ and $h_2(3)$ so it MIGHT BE/ IS DEFINITELY NOT in the filter

Number 9 has $h_1(9)$ and $h_2(9)$ so it MIGHT BE/ IS DEFINITELY NOT in the filter

Number 15 has $h_1(15)$ and $h_2(15)$ so it MIGHT BE/ IS DEFINITELY NOT in the filter

Number 20 has $h_1(20)$ and $h_2(20)$ so it MIGHT BE/ IS DEFINITELY NOT in the filter

(b) $h_1(3)=h_2(3)=6$. The bit array has a 1 in position 6 might be

$h_1(9)=4, h_2(9)=6$ The bit array has a 0 in position 4 definitely not

$h_1(15)=2, h_2(15)=6$ Both positions are a 1 might be

$h_1(20)=7, h_2(20)=1$ The bit array has a 0 in position 7 definitely not

(c) (**4 points**) Now, for the same sequence as defined at the beginning of this topic, using the Flajolet-Martin algorithm, and the following hash functions

$f_1(x) = (7 * x + 13) \bmod 32$

$f_2(x) = (5 * x + 17) \bmod 32$

estimate the number of distinct elements (as defined in the slides) for each function, after having seen all stream elements.

(c) $\langle 8, 12, 5, 7, 3, 15 \rangle$ Since $\log_2(32)=5$, we should only consider the 5-bit binary representation

- $x=8$ $f_1(8)=5 \rightarrow 00101 \rightarrow 0$ (trailing zeroes) $f_2(8)=25 \rightarrow 11001 \rightarrow 0$
- $x=12$ $f_1(12)=1 \rightarrow 00001 \rightarrow 0$ $f_2(12)=13 \rightarrow 01101 \rightarrow 0$
- $x=5$ $f_1(5)=16 \rightarrow 10000 \rightarrow 4$ max $f_2(5)=10 \rightarrow 01010 \rightarrow 1$
- $x=7$ $f_1(7)=30 \rightarrow 11110 \rightarrow 1$ $f_2(7)=20 \rightarrow 10100 \rightarrow 2$
- $x=3$ $f_1(3)=2 \rightarrow 00010 \rightarrow 1$ $f_2(3)=0 \rightarrow 00000 \rightarrow 5$ max
- $x=15$ $f_1(15)=22 \rightarrow 10110 \rightarrow 1$ $f_2(15)=28 \rightarrow 11100 \rightarrow 2$

$2^{\text{max trailing zeroes}}$ $N_1 = 2^4 = 16$ $N_2 = 2^5 = 32$

(d) **(3 points)** How can we use both N_1 and N_2 to find a more accurate estimate of the number of distinct elements in the stream?

Answer:

A more accurate representation is provided by ? the two values: N =

To calculate a more accurate estimate of the number of distinct elements using N_1 and N_2 we calculate their average $N = \frac{N_1+N_2}{2} = \frac{48}{2} = 24$