**COMP9313 2018s2**

**Assignment**

**z5089358**

**Fengting YANG**

**Question 1:**

Pseudo code:

for each **line** in inputdata:

flatMap(**line** =>

for each **item** in **line**.split(“:”)(1).split(“,”):

yield((**line**.split(“:”)(0),**item**))

filter(**tuple** => **tuple**.\_1 < **tuple**.\_2)

for each **tuple**:

flatMap(**tuple** =>

for each **othertuple** in other tuples:

if **othertuple**.\_1 == **tuple**.\_1 or

**othertuple**.\_1 == **tuple**.\_2 or

**othertuple**.\_2 == **tuple**.\_1 or

**othertuple**.\_2 == **tuple**.\_2:

yield ((**tuple**, **othertuple**))

reduceByKey(\_:::\_)

sortValue

sortByKey

print or save to file

**Question 2:**

(i)

For sentence A: A (“the sky is blue the sun is bright”) and sentence B: B (“the sun in the sky is bright”)

Their 2-shingles group:

|  |  |
| --- | --- |
| 2-shingles word set from A | 2-shingles word set from B |
| the sky | the sun |
| sky is | sun in |
| is blue | in the |
| blue the | the sky |
| the sun | sky is |
| sun is | is bright |
| is bright |  |

Jaccard similarity:

A∩B = 4

A∪B = 9

Jaccard Similarity = 4/9 = 0.44

(ii)

For two hash functions

h1(n) = 5n – 1 mod M

h2(n) = 2n + 1 mod M

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| #Row | Token | C(A) | C(B) | h1(n) | h2(n) |
| 0 | the sky | 1 | 1 | 8 | 1 |
| 1 | sky is | 1 | 1 | 4 | 3 |
| 2 | is blue | 1 | 0 | 0 | 5 |
| 3 | blue the | 1 | 0 | 5 | 7 |
| 4 | the sun | 1 | 1 | 1 | 0 |
| 5 | sun is | 1 | 0 | 6 | 2 |
| 6 | is bright | 1 | 1 | 2 | 4 |
| 7 | sun in | 0 | 1 | 7 | 6 |
| 8 | in the | 0 | 1 | 3 | 8 |

Initilize:

|  |  |  |
| --- | --- | --- |
|  | SigA | SigB |
| h1 | ∞ | ∞ |
| h2 | ∞ | ∞ |

Row 0:

|  |  |  |
| --- | --- | --- |
|  | SigA | SigB |
| h1 | 8 | 8 |
| h2 | 1 | 1 |

Row 1:

|  |  |  |
| --- | --- | --- |
|  | SigA | SigB |
| h1 | 4 | 4 |
| h2 | 1 | 1 |

Row 2:

|  |  |  |
| --- | --- | --- |
|  | SigA | SigB |
| h1 | 0 | 4 |
| h2 | 1 | 1 |

Row 3:

|  |  |  |
| --- | --- | --- |
|  | SigA | SigB |
| h1 | 0 | 4 |
| h2 | 1 | 1 |

Row 4: (Here, SigA get an early start)

|  |  |  |
| --- | --- | --- |
|  | SigA | SigB |
| h1 | 0 | 1 |
| h2 | 0 | 0 |

Row 5:

|  |  |  |
| --- | --- | --- |
|  | SigA | SigB |
| h1 | 0 | 1 |
| h2 | 0 | 0 |

Row 6:

|  |  |  |
| --- | --- | --- |
|  | SigA | SigB |
| h1 | 0 | 1 |
| h2 | 0 | 0 |

Row 7:

|  |  |  |
| --- | --- | --- |
|  | SigA | SigB |
| h1 | 0 | 1 |
| h2 | 0 | 0 |

Row 8:

|  |  |  |
| --- | --- | --- |
|  | SigA | SigB |
| h1 | 0 | 1 |
| h2 | 0 | 0 |

So the final signature for A is 0 0 and for B is 1 0