Information Retrieval Fall 2016

**Quiz 4 ( Total Marks = 10)**

**Roll No: Name**

**Q1)** Consider following posting list of a term. (document Id, count, [positions]) (4 Marks)

(3,3,[4,7,12]) (5,1,[84]) (12,4,[13,15,20,24])

1. Delta encode document Ids and delta encode term positions
2. Encode resulting list from part a using Elias Gamma Encoding
3. How many bits are required for encoding entire list in part b? How many bits will be required for encoding list from part a using fixed length encoding of 8 bits per number

**Solution:**

**a) (**3,3,[4,3,5]) (2,1,[84]) (7,4,[13,2,5,4])

b) 101 101 11000 101 11001 100 0 1111110010100 11011 11000 1110101 100 11001 11000

c) 3 + 3+ 5+ 3+5+3 +1+ 13+ 5+5+7+3+5+5 = 5\*6 + 3\*5 + 1+13+7 = 30+15+21 = 66

encoding list from part a using fixed length encoding = 14\*8 = 112

**Q2)** Following table gives RSS (Residual Sum of Squares) for different value of K using K Means clustering algorithm for some n documents. Which value of K will you choose and why? (2 Marks)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **K** | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **RSS** | 2000 | 1800 | 1610 | 1565 | 1300 | 1120 | 900 | 700 | 500 |

**Ans:** K = 4 since K = 5 doses not give much reduction in RSS . This is Knee in plot.

Q3) Show the different steps of HAC algorithm using the distance matrix below.

Give partial results after each step. [4 Marks]

a) Calculate Similarity of Clusters using Complete Link

b) Calculate Similarity of Clusters using Single Link

****

**Solution**

**a) Complete Link**

Document 4 and 5 have minimum distance so they will be merged first

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4-5 |
| 1 | 0 |  |  |  |
| 2 | 2 | 0 |  |  |
| 3 | 4 | 3 | 0 |  |
| 4-5 | 10 | 7 | 9 | 0 |

Documents 1 and 2 have minimum distance so they will be merged

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1-2 | 3 | 4-5 |
| 1-2 | 0 |  |  |
| 3 | 4 | 0 |  |
| 4-5 | 10 | 9 | 0 |

Cluster 3 and Cluster 1-2 will be merged

So we will have last 2 clusters 1-2-3 and 4-5. They will be merged to get 1-2-3-4-5

**a) Single Link**

Document 4 and 5 have minimum distance so they will be merged first

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4-5 |
| 1 | 0 |  |  |  |
| 2 | 2 | 0 |  |  |
| 3 | 4 | 3 | 0 |  |
| 4-5 | 8 | 5 | 6 | 0 |

Documents 1 and 2 have minimum distance so they will be merged

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1-2 | 3 | 4-5 |
| 1-2 | 0 |  |  |
| 3 | 3 | 0 |  |
| 4-5 | 5 | 6 | 0 |

Cluster 3 and Cluster 1-2 will be merged

So we will have last 2 clusters 1-2-3 and 4-5. They will be merged to get 1-2-3-4-5