Information Retrieval and Text Mining Fall 2016

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

**Roll No: Name**

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

(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 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?

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **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)** Create clusters using HAC (centroid clustering). Use Euclidean distance.

**Solution**

d1 (2,3)

d2(2,5)

d3(4,5)

d4(5,1)

d5(6.5,1)

d6(7,3)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | d1 (2,3) | d2(2,5) | d3(4,5) | d4(5,1) | d5(6.5,1) | d6(7,3) |
| d1 (2,3) | 0 |  |  |  |  |  |
| d2(2,5) | 2 | 0 |  |  |  |  |
| d3(4,5) | Sqrt(8) | 2 | 0 |  |  |  |
| d4(5,1) | Sqrt(13) | 5 | 17 | 0 |  |  |
| d5(6.5,1) | Sqrt(24.25) | Sqrt(36.25) | Sqrt(22.25) | 1.5 | 0 |  |
| d6(7,3) | 5 | Sqrt(29) | Sqrt(13) | Sqrt(8) | Sqrt(4.25) | 0 |

d4(5,1) and d5(6.5,1) have minimum distance so they will be merged in first iteration. Their centroid is d4-5(5.75,1)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | d1 (2,3) | d2(2,5) | d3(4,5) | d4-5(5.75,1) | d6(7,3) |
| d1 (2,3) | 0 |  |  |  |  |
| d2(2,5) | 2 | 0 |  |  |  |
| d3(4,5) | Sqrt(8) | 2 | 0 |  |  |
| d4-5(5.75,1) |  |  |  | 0 |  |
| d6(7,3) | 5 | Sqrt(29) | Sqrt(13) |  | 0 |