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]) (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

**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 |

**Q3)** Create clusters using HAC (centroid clustering). Use Euclidean distance. (4 Marks)