**INFORMATION RETRIEVAL USING DOCUMENT SUMMARIES**

Text tiling

Initially the document is partitioned using text tiling. The text tiling function gives a vector as output which contains 0’s and 1’s where 1 represents the sentence at which the partition is to be made. Using this vector, the document is partitioned.

Topic modeling

The partitioned documents above are given to LDA (Latent Dirichlet Allocation) algorithm which does topic modeling and outputs a matrix which represents document topic distribution.

The number of topics is decided based on topic coherence.

Coherence =Σ scores.

Score = p(wi,wj)/p(wi).p(wj). [extrinsic uci measure]

p(wi,wj)-probability of co-occurrence of wi,wj in a document.

p(wi)-probability of occurrence of wi in a document.

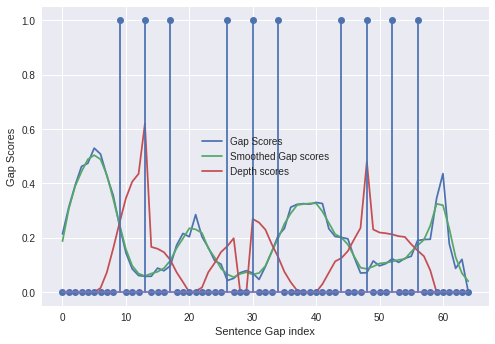
We plotted a graph between the no of topics and topic coherence. The graph is plotted till the change in coherence is less than the threshold (0.01) for at least three values of x-axis(number of topics).The number of topics where the change in coherence becomes less than threshold is considered as the optimal number of topics.

The output(document-topic distribution) serves as the input for K-means which groups the document partitions. We used Jensen Shannon distance as the distance measure in K-means.

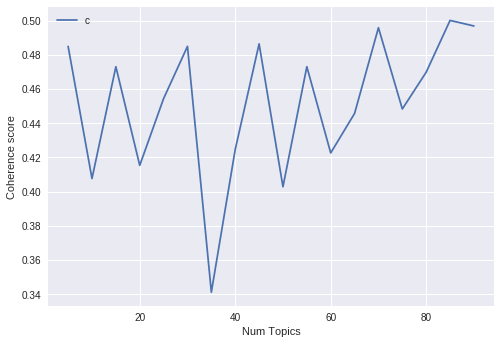
**Outputs:**

**Input 1**: test2.txt

**Image showing the output of text tiling on test2.txt:**



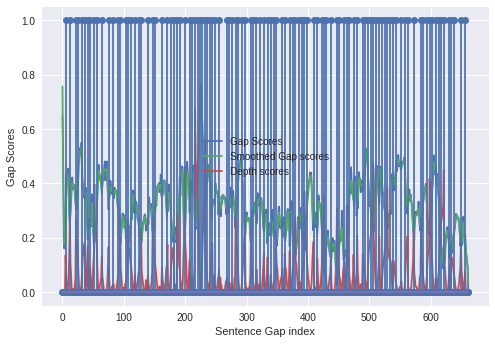
**Graph of number of topics versus Topic coherence:**



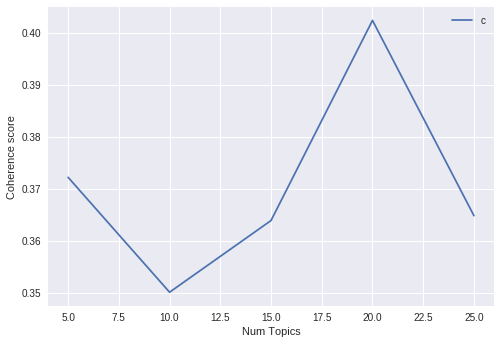
Using the above results, the number of topics is set to 75.

**Input2:**

**Image showing the output of text tiling on test4.txt:**

****

**Graph of number of topics versus Topic coherence:**

****

Using the above results, the number of topics is set to 10.