Medius Health – Programming Challenge for Data Science

Task Description

Grouping documents into clusters. The documents in a cluster will have same semantic description. For example, the documents in cluster 1 talks about a "soccer" game etc.

The Data Description

1. For this task, this dataset is provided in "data" folder. Each file in the data folder is considered as a document.
2. There are 300 documents in the directory.
3. Each document has got some texts.

Task Description

1. The task is required to be completed in python.
2. Process the text data in each document/file. It might require having some knowledge in NLP, data processing, text mining and python.
3. Develop a model to partition the data into multiple clusters. It is required to develop the end-to-end model in python instead of using any data clustering libraries or pre-trained models.
4. The outcome of the model is number of clusters and the data points in each cluster.
5. Report the number of clusters found in the data.
6. Find out the topics of each cluster. (you can run any benchmark off-the-shelf topic modelling algorithm like Latent Dirichlet Allocation (LDA) or PLSA)
7. If possible, can you visualize the cluster. (bonus point)