* Retrieve the wiki pages related to Diseases and Disorders and their subcategories. Tag the pages with their respective sub category. Build a document classification model. Split the same data for training and testing purposes.
* The data files "readme.txt" and "posts.rar" have about 11,000 blog posts from a blog. Your job is to cluster the text of these blog posts into at most 5 clusters, and output the top-10 representative words for each cluster. You can use TF-IDF, Vector-space model, and Kmeans clustering.