**INFORMATION RETRIEVAL \_ ASSIGNMENT 8**

**Content**: Text Clustering

**1. PREPARING:**

1.1. Go to page

<https://kdd.ics.uci.edu/databases/reuters21578/>

1.2. Download and extract reuters21578.tar.gz

**2. CLUSTERING**

2.1. Read README.txt to understand the format of the dataset.

2.2. Combine all documents of the dataset into one collection.

2.3. Choose randomly 5000 words for the vocubulary.

2.4. Build tf-idf vectors for documents.

2.5. Cluster the tf-idf vectors using HAC method given the expected number of clusters k