Machine Learning - Assignment 4 - k-Means clustering

Evgeny Manturov

November 15, 2023

The code presented is not complicated at all due to the use of **sklearn** machine learning library.
The example dataset is 6 text files, including 3 topics, 2 texts assigned to each topic:

- Topic 1 - Electrical and electronic engineering - acynch_gen.txt, transmission_lines.txt
- Topic 2 - The Forgotten Realms (fantasy universe) - astralplane.txt, githyanki.txt
- Topic 3 - Aircraft (inc. military aircraft-related weapon article) - interceptor.txt, sidewinder.txt

The procedure is as follows:

- Sort filenames, read and assign test labels (the only metric measured is "same cluster", so the label list can be any number combination of [x,y,y,z,z,x], e.g. [0,1,1,2,2,0])
- Perform TF-IDF vectorizing (settings are: exclude low-weight words (rare words), with Euclidean (L2) norm, default English language stopwords
- Perform dimensionality reduction via SVD with sklearn.decomposition.PCA
- Calculate k-Means for a range of cluster numbers, record labeled and non-label metrics and compare them

The results are:

```
Number of clusters: 2
[('acynch_gen.txt', 1), ('astralplane.txt', 0), ('githyanki.txt', 0),
('interceptor.txt', 1), ('sidewinder.txt', 1), ('transmission_lines.txt', 1)]
Homogenuity score: 0.5793801642856953
Completeness score: 1.0000000000000004
V Measure score: 0.7336804366512113
Additional metrics:
Silhouette score = 0.01746196806639581
Calinski Harabasz score = 1.0959115360278036
Davies Bouldin score = 1.7265560727297924
Number of clusters: 3
[('acynch_gen.txt', 1), ('astralplane.txt', 0), ('githyanki.txt', 0),
('interceptor.txt', 2), ('sidewinder.txt', 2), ('transmission_lines.txt', 1)]
Homogenuity score: 1.0
Completeness score: 1.0
```

```
V Measure score: 1.0
Additional metrics:
Silhouette score = 0.027575407835355933
Calinski Harabasz score = 1.121734722758693
Davies Bouldin score = 1.3398359123620738
Number of clusters: 4
[('acynch_gen.txt', 3), ('astralplane.txt', 1), ('githyanki.txt', 1),
('interceptor.txt', 2), ('sidewinder.txt', 2), ('transmission_lines.txt', 0)]
Homogenuity score: 1.0
Completeness score: 0.82623465712856
V Measure score: 0.9048504844691448
Additional metrics:
Silhouette score = 0.021152629875361328
Calinski Harabasz score = 1.099265040564918
Davies Bouldin score = 0.939606483378594
```

Clustering for 3 clusters makes sense, is precise and accurate. Clusters are clearly visible on a 2-D SVD-dimensioned plot:

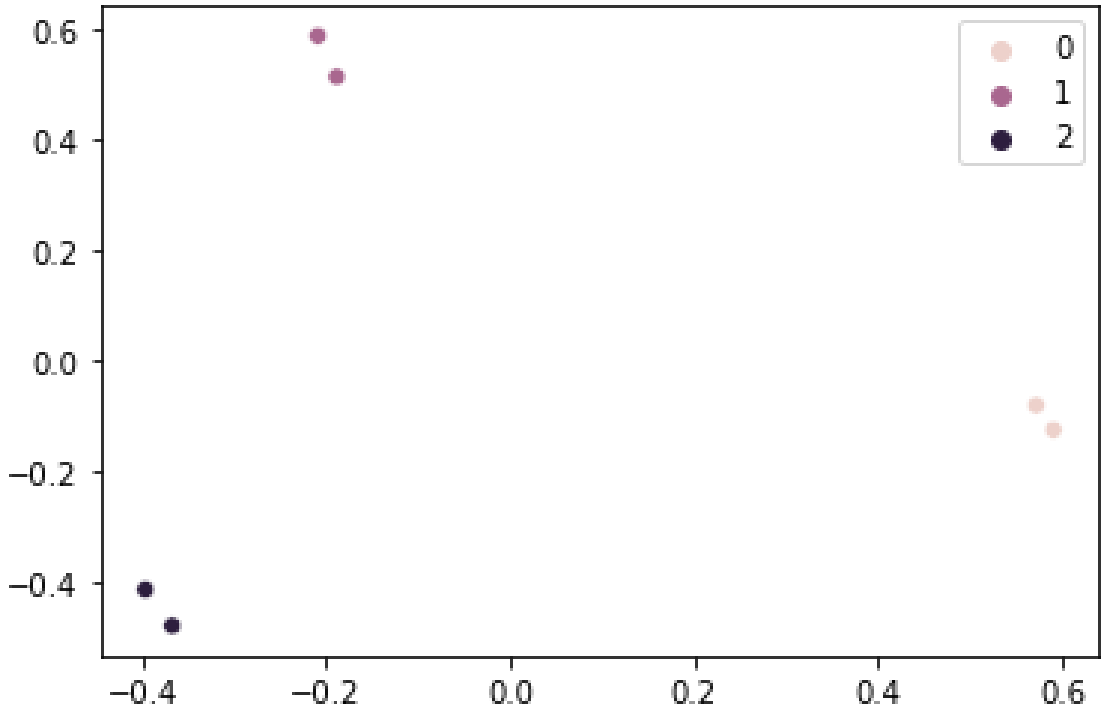

Figure 1: SVD TF-IDF plot

However, unsupervised metrics display this clustering as imprecise and clusters as overlapping. This may be due to a small amount of data to be clustered. Perhaps the scoring will improve if more texts are given.