Unsupervised Machine Learning - Coursera Project

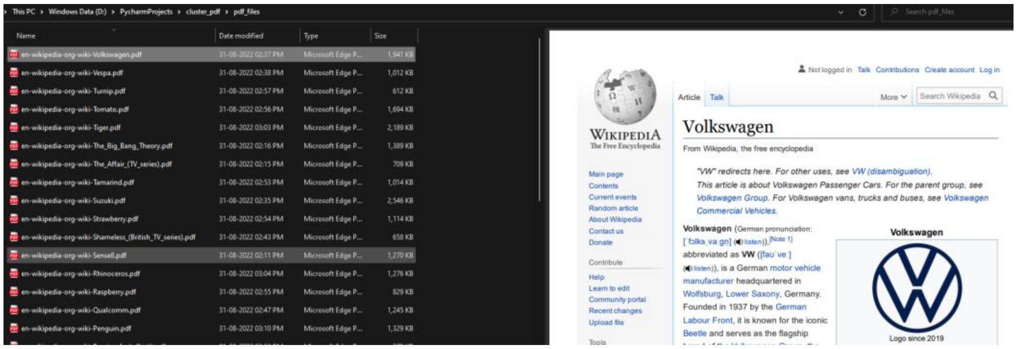
Automatic Folder Organization

Objective:

A clustering model was proposed to divide a folder containing multiple PDF documents into different folders, where each folder has PDF files corresponding to a topic. The benefit of this application is that a user can reorganize a folder automatically based on the text present inside them.

Data:

84 Wikipedia pages were scraped and saved as a PDF (using webtopdf.com) in a folder (as shown below in screenshot). It consisted of pages of TV series/movies, automobile companies, fruits and vegetables, birds and animals, software companies, semiconductor companies. Each webpage when converted to PDF resulted in multiple pages, but only the text from the first page from each PDF was chosen for clustering the document.



› This PC › Windows Data (D:) › PycharmProjects › cluster_pdf › pdf_files

Name	Date modified	Type	Size
en-wikipedia-org-wiki-Volkswagen.pdf	31-08-2022 02:37 PM	Microsoft Edge P...	1,941 KB
en-wikipedia-org-wiki-Vespa.pdf	31-08-2022 02:38 PM	Microsoft Edge P...	1,012 KB
en-wikipedia-org-wiki-Turnip.pdf	31-08-2022 02:57 PM	Microsoft Edge P...	612 KB
en-wikipedia-org-wiki-Tomato.pdf	31-08-2022 02:56 PM	Microsoft Edge P...	1,694 KB
en-wikipedia-org-wiki-Tiger.pdf	31-08-2022 03:03 PM	Microsoft Edge P...	2,189 KB
en-wikipedia-org-wiki-The_Big_Bang_Theory.pdf	31-08-2022 02:16 PM	Microsoft Edge P...	1,389 KB
en-wikipedia-org-wiki-The_Affair_(TV_series).pdf	31-08-2022 02:15 PM	Microsoft Edge P...	709 KB
en-wikipedia-org-wiki-Tamarind.pdf	31-08-2022 02:53 PM	Microsoft Edge P...	1,014 KB
en-wikipedia-org-wiki-Suzuki.pdf	31-08-2022 02:35 PM	Microsoft Edge P...	2,546 KB
en-wikipedia-org-wiki-Strawberry.pdf	31-08-2022 02:54 PM	Microsoft Edge P...	1,114 KB
en-wikipedia-org-wiki-Shameless_(British_TV_series).pdf	31-08-2022 02:43 PM	Microsoft Edge P...	658 KB
en-wikipedia-org-wiki-Sense8.pdf	31-08-2022 02:11 PM	Microsoft Edge P...	1,270 KB
en-wikipedia-org-wiki-Rhinoceros.pdf	31-08-2022 03:04 PM	Microsoft Edge P...	1,276 KB
en-wikipedia-org-wiki-Raspberry.pdf	31-08-2022 02:55 PM	Microsoft Edge P...	829 KB
en-wikipedia-org-wiki-Qualcomm.pdf	31-08-2022 02:47 PM	Microsoft Edge P...	1,245 KB
en-wikipedia-org-wiki-Penguin.pdf	31-08-2022 03:10 PM	Microsoft Edge P...	1,329 KB

Volkswagen

From Wikipedia, the free encyclopedia

"VW" redirects here. For other uses, see VW (disambiguation).

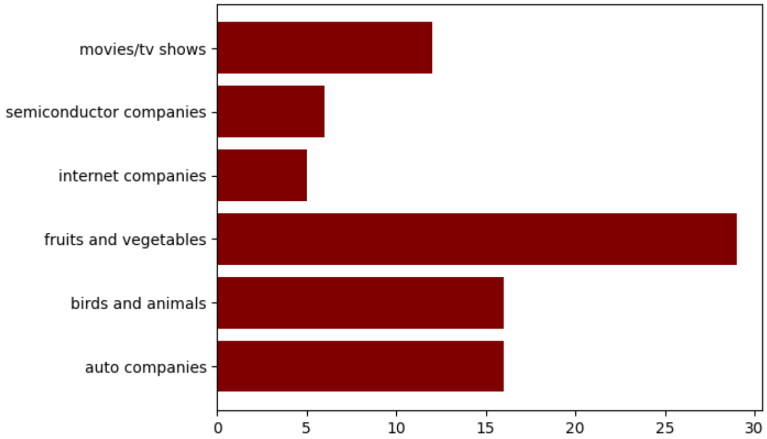
This article is about Volkswagen Passenger Cars. For the parent group, see Volkswagen Group. For Volkswagen vans, trucks and buses, see Volkswagen Commercial Vehicles.

Volkswagen (German pronunciation: [ˈfɔlksˌvaːɡn̩] (listen)),[Note 1] abbreviated as **VW** ([faʊˈveː] (listen)), is a German motor vehicle manufacturer headquartered in Wolfsburg, Lower Saxony, Germany. Founded in 1937 by the German Labour Front, it is known for the iconic Beetle and serves as the flagship

Volkswagen

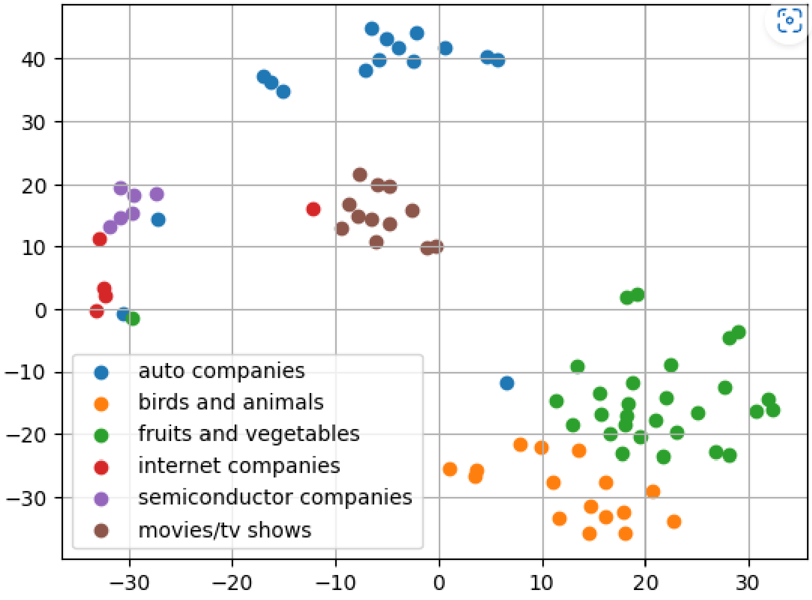
Logo since 2019

The count of PDF files for each cluster is also plotted below:



Data Exploration and Feature Engineering:

A possible snag in the preliminary hypothesis of the data, is that only the first page of each PDF was used to reduce the time taken for clustering the PDF files and move into a new folder for the purpose of reducing clustering latency. For each first page in a PDF, the stop words were removed, the remaining words were then lemmatized and TD-IDF (Term Frequency-Inverse Document frequency) vectors were computed. Only unigrams were chosen for creating the TF-IDF vector as presence of same bigram in documents of same topic was less common. The max number of features for the TD-IDF vectorizer was chosen as 1024. The t-SNE plot for our dataset is shown below. It shows the 6 clusters present in our dataset.

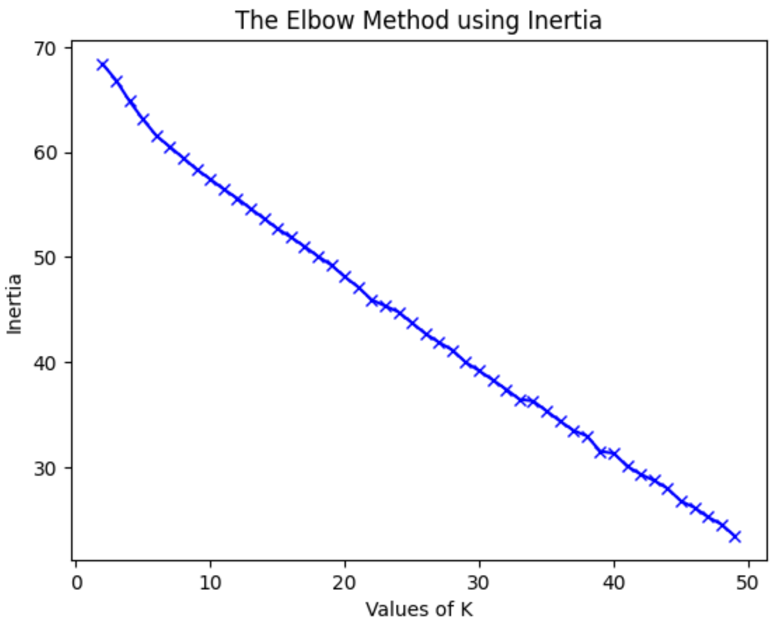


Clustering Approach

We tried 2 clustering approaches and showed that Agglomerative Clustering is the better one.

K-Means Clustering

We used elbow method using inertia to determine the number of clusters for K-means. It was observed that the inflection point is not observed in the plot. It is mainly because the shape of the cluster need not be spherical in our use case. Therefore, we try agglomerative clustering which has more control over the shape of the cluster.

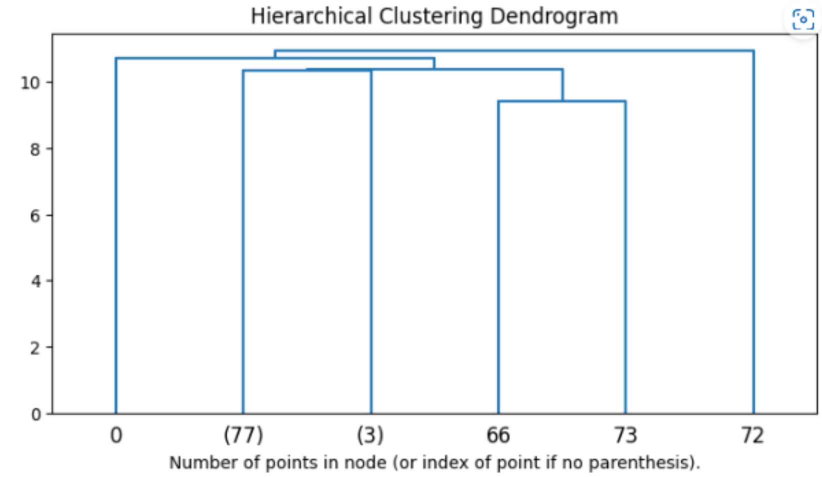


Hierarchical Agglomerative Clustering

Since the number of clusters for a given set of documents is not know and the number of PDF corresponding to a cluster may not be even, hierarchical clustering was chosen. It is also scalable in terms of large number of documents to be clustered and large number of clusters which might be present.

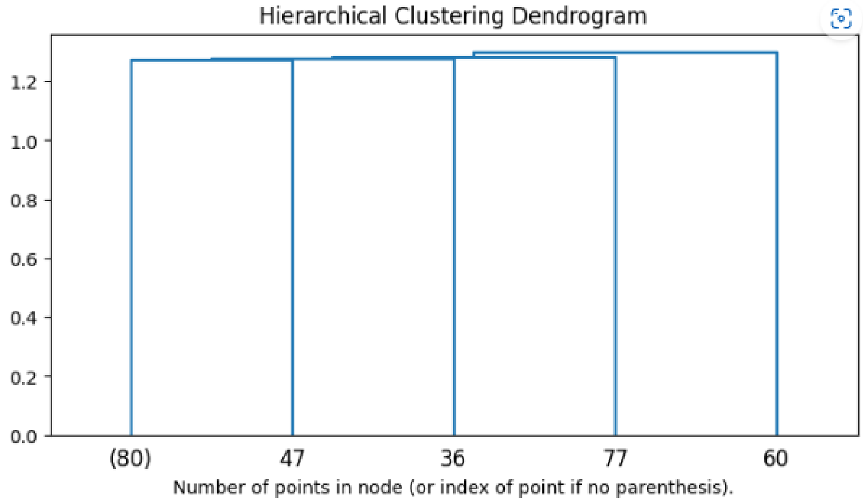
Single Linkage with L1 distance metric

From the dendrogram, we saw that **it failed to cluster the document properly** as most (77) of the documents fell into a single cluster.



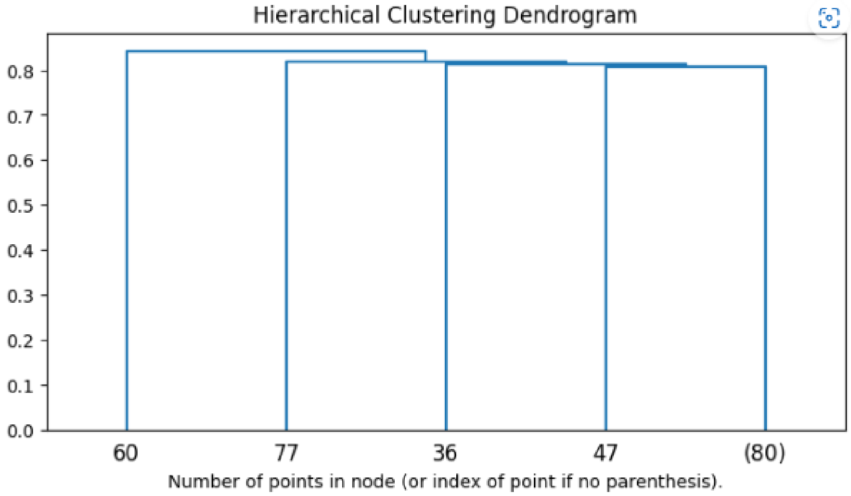
Single Linkage with L2 distance metric

From the dendrogram, we saw that it **failed to cluster the document properly** as most (80) of the documents fell into a single cluster.



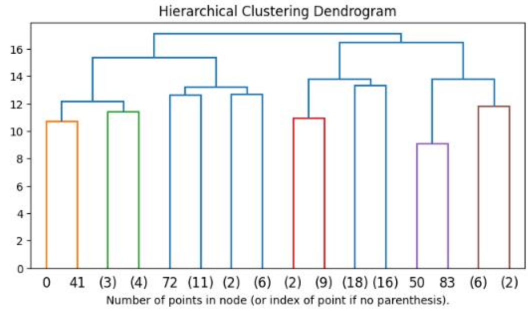
Single Linkage with cosine distance metric

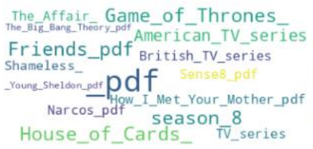
From the dendrogram, we saw that it **failed to cluster the document properly** as most (80) of the documents fell into a single cluster.

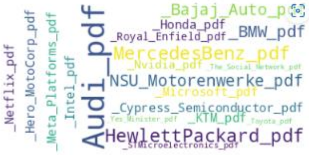


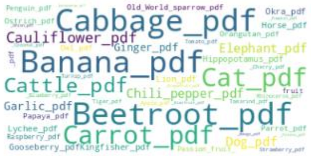
Complete Linkage with L1 distance metric

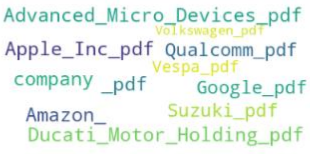
The dendrogram for complete linkage is shown below. The distance threshold was then chosen as 14 and the file names for each of the clusters thus obtained was plotted in the form of word cloud. Most of the movies/ TV shows, fruits and vegetables, and companies (in 2 clusters) were clustered separately. **The type of company like semiconductor, IT or automobile company got mixed up in this clustering.**





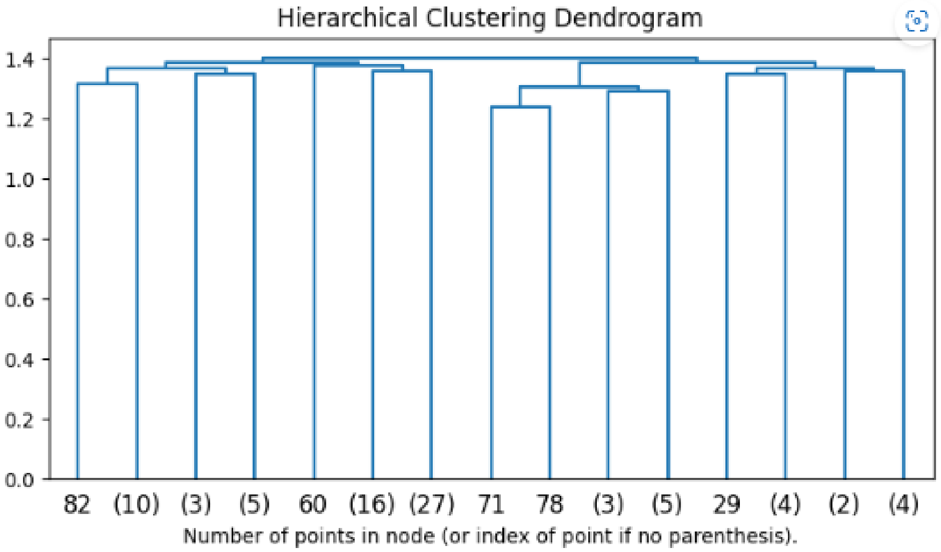






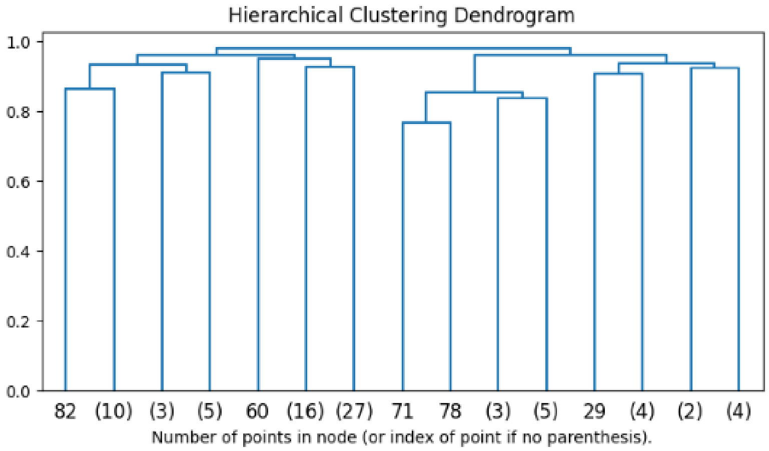
Complete Linkage with L2 distance metric

From the dendrogram, we saw that the range of threshold for which meaningful clusters were getting formed was very small. So choosing a **threshold would be very sensitive**.



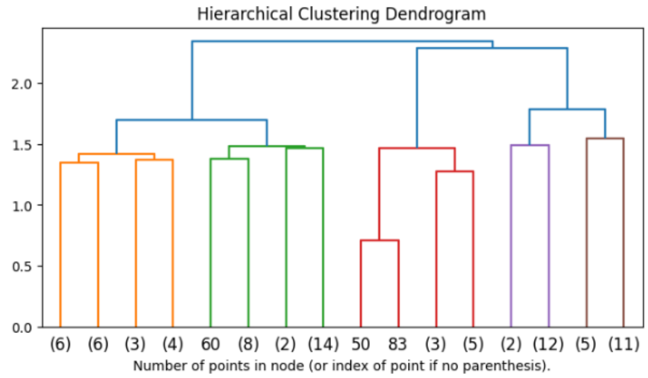
Complete Linkage with cosine distance metric

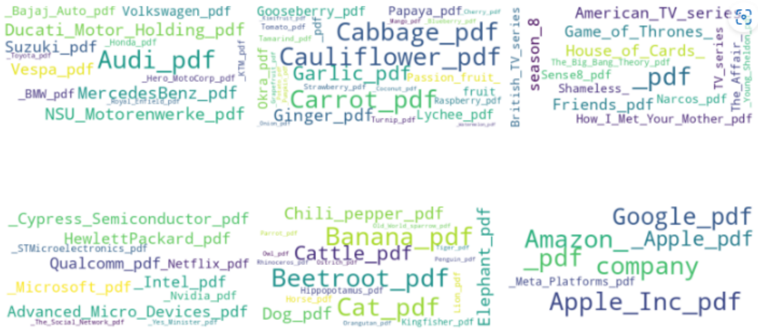
From the dendrogram, we saw that the range of threshold for which meaningful clusters were getting formed was very small. So choosing a **threshold would be very sensitive**.



WARD Linkage with L2 distance metric

The dendrogram for ward linkage is shown below. The distance threshold was then chosen as 1.5 and the file names for each of the clusters thus obtained was plotted in the form of word cloud. It can be seen that most of the automobile companies, fruits and vegetables, movies/ TV shows, semiconductor companies, fruits and vegetables, and internet companies were clustered separately. **This has the best formed clusters.**

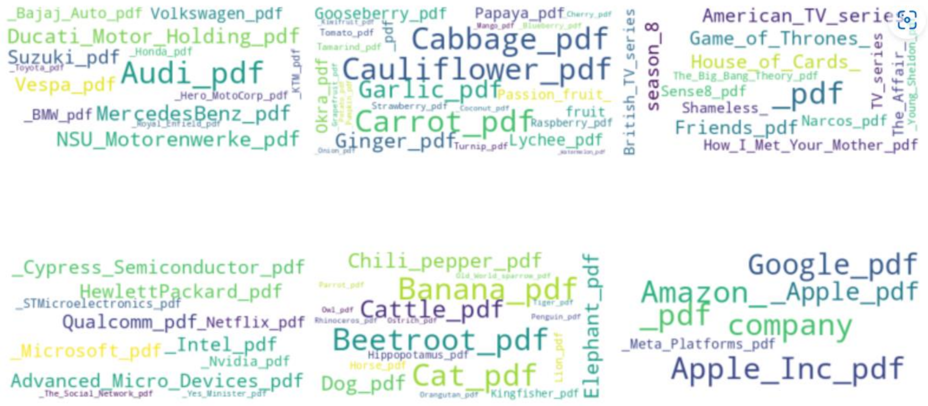




Note: L1 and cosine distance metric are not supported for WARD linkage.

Best Model

The best model was the one with WARD Linkage with Euclidean distance metric. It was able to even cluster the companies which were fine grained like semiconductor, internet and automobile companies. The corresponding word cloud is shown below.



Summary

A clustering approach was used to automatically restructure a folder of PDF documents into folder of PDF files belonging to a topic based on the text present in them. **Hierarchical Agglomerative Clustering was successfully used for this task with WARD Linkage with Euclidean distance metric**. It was able to cluster various topics present in the dataset including clusters which were fine grained like semiconductor, internet and automobile companies. **K-means clustering was not suitable** for our use case as the clusters are not spherical. Our final model was better as it works for non-spherical clusters as well.

Code

Github Link : CV-and-Neural-Nets-Basic/Automatic Folder Organization at master · kritiksoman/CV-and-Neural-Nets-Basic (github.com)

Future Work/Current Drawbacks

All the following are missing in the current approach:

- Support for **more file formats**: A folder will usually contain multiple file formats. We should add support to common ones like docx, pptx, xlsx, html etc.
- Adding **more features**: like file metadata, support for different file formats, using image embeddings as features are few improvements which can be made to this work.
- Improving the clustering approach: It should also have **support for handling outliers**.
- Extract features from **all the pages** of the document.
- In practice, the **user should have some control over the clusters which are formed**. So, it might be useful to take some keywords as input from the user and use that for clustering.