REPORT

Assignment 2a

-Shreshth Saxena (38)
Thongminlien Kuki (46)

Objective : You are given a folder containing documents (text/pdf files).Normalize the text and create a similarity matrix using Jaccard index.Apply hierarchical clustering. Cut the dendrogram at k and identify clusters of similar documents.

Packages used :

NLTK (Natural Language Tool Kit)
numpy
scipy
SKlearn
matplotlib

Stop-words removed:

['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't", 'search', 'engine', 'web', 'internet']

Stemming:

['History', 'Engines', 'Google', 'become', 'synonyms', 'research', 'nowadays', 'anymore', 'google', 'become', 'verb', 'Oh', 'write', 'note', 'history', 'engines', 'google', 'Google', 'came', 'long', 'first', 'came', 'existence', 'get', 'history', 'engines', 'must'…..]

['histori', 'engin', 'googl', 'becom', 'synonym', 'research', 'nowaday', 'anymor', 'googl', 'becom', 'verb', 'Oh', 'write', 'note', 'histori', 'engin', 'googl', 'googl', 'came', 'long', 'first', 'came', 'exist', 'get', 'histori', 'engin', 'must'…..]

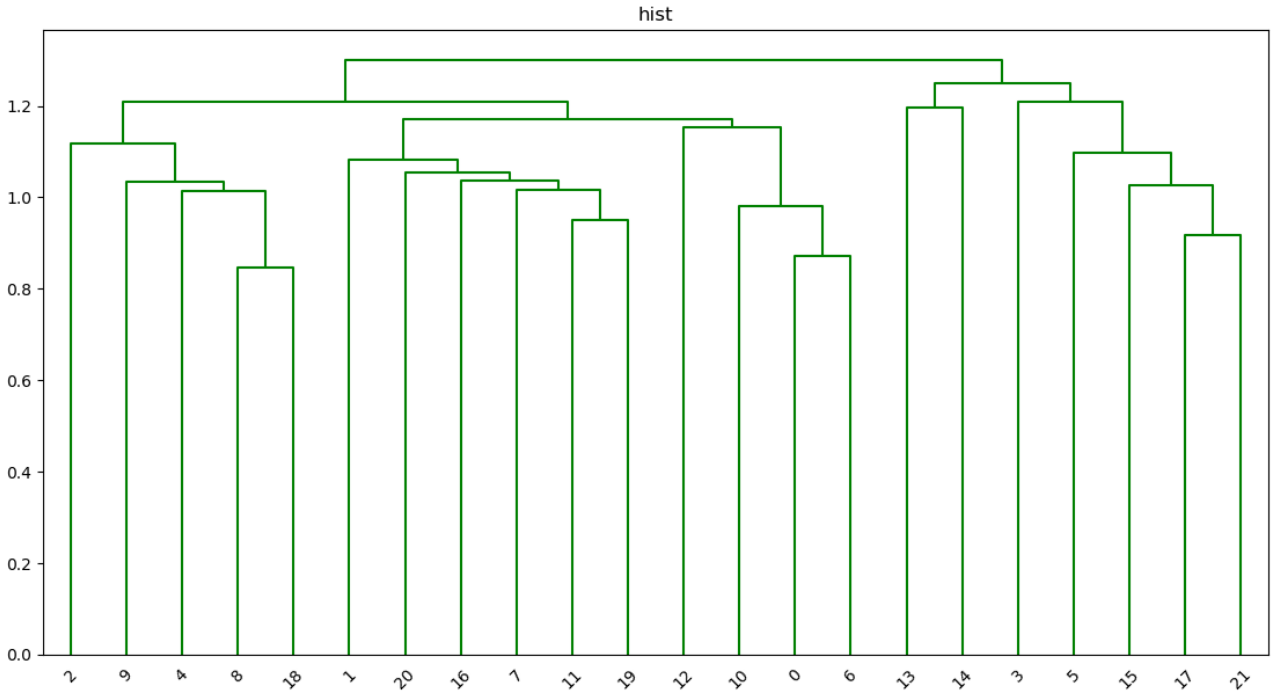
Similarity Matrix using Jaccard similarity

[[1. 0.207 0.186 0.154 0.219 0.192 0.389 0.202 0.28 0.275 0.311 0.232 0.208 0.193 0.139 0.215 0.249 0.205 0.288 0.231 0.245 0.183]
 [0.207 1. 0.185 0.191 0.245 0.236 0.253 0.278 0.194 0.232 0.204 0.262 0.185 0.225 0.147 0.229 0.238 0.182 0.205 0.326 0.261 0.218]
 [0.186 0.185 1. 0.215 0.283 0.149 0.193 0.191 0.227 0.231 0.201 0.193 0.195 0.162 0.144 0.206 0.182 0.186 0.271 0.189 0.247 0.159]
 [0.154 0.191 0.215 1. 0.174 0.163 0.174 0.193 0.156 0.171 0.139 0.196 0.161 0.157 0.144 0.182 0.191 0.161 0.144 0.191 0.184 0.183]
 [0.219 0.245 0.283 0.174 1. 0.19 0.24 0.224 0.291 0.284 0.271 0.212 0.199 0.183 0.142 0.206 0.248 0.179 0.323 0.211 0.238 0.175]
 [0.192 0.236 0.149 0.163 0.19 1. 0.212 0.211 0.187 0.181 0.202 0.224 0.229 0.208 0.206 0.249 0.198 0.238 0.178 0.226 0.261 0.238]
 [0.389 0.253 0.193 0.174 0.24 0.212 1. 0.247 0.281 0.289 0.323 0.255 0.212 0.181 0.148 0.276 0.239 0.201 0.291 0.251 0.268 0.218]
 [0.202 0.278 0.191 0.193 0.224 0.211 0.247 1. 0.205 0.214 0.222 0.31 0.215 0.172 0.176 0.254 0.271 0.187 0.174 0.287 0.262 0.184]
 [0.28 0.194 0.227 0.156 0.291 0.187 0.281 0.205 1. 0.277 0.258 0.258 0.205 0.181 0.15 0.237 0.219 0.183 0.405 0.224 0.23 0.203]
 [0.275 0.232 0.231 0.171 0.284 0.181 0.289 0.214 0.277 1. 0.258 0.255 0.212 0.17 0.145 0.261 0.261 0.172 0.28 0.26 0.244 0.192]
 [0.311 0.204 0.201 0.139 0.271 0.202 0.323 0.222 0.258 0.258 1. 0.202 0.231 0.151 0.162 0.227 0.223 0.171 0.271 0.205 0.258 0.188]
 [0.232 0.262 0.193 0.196 0.212 0.224 0.255 0.31 0.258 0.255 0.202 1. 0.218 0.192 0.176 0.273 0.272 0.182 0.238 0.333 0.263 0.212]
 [0.208 0.185 0.195 0.161 0.199 0.229 0.212 0.215 0.205 0.212 0.231 0.218 1. 0.15 0.178 0.221 0.197 0.182 0.192 0.191 0.231 0.173]
 [0.193 0.225 0.162 0.157 0.183 0.208 0.181 0.172 0.181 0.17 0.151 0.192 0.15 1. 0.161 0.169 0.188 0.156 0.178 0.236 0.202 0.164]
 [0.139 0.147 0.144 0.144 0.142 0.206 0.148 0.176 0.15 0.145 0.162 0.176 0.178 0.161 1. 0.176 0.174 0.137 0.138 0.159 0.203 0.147]
 [0.215 0.229 0.206 0.182 0.206 0.249 0.276 0.254 0.237 0.261 0.227 0.273 0.221 0.169 0.176 1. 0.243 0.293 0.205 0.228 0.261 0.302]
 [0.249 0.238 0.182 0.191 0.248 0.198 0.239 0.271 0.219 0.261 0.223 0.272 0.197 0.188 0.174 0.243 1. 0.177 0.223 0.275 0.266 0.203]
 [0.205 0.182 0.186 0.161 0.179 0.238 0.201 0.187 0.183 0.172 0.171 0.182 0.182 0.156 0.137 0.293 0.177 1. 0.166 0.189 0.187 0.354]
 [0.288 0.205 0.271 0.144 0.323 0.178 0.291 0.174 0.405 0.28 0.271 0.238 0.192 0.178 0.138 0.205 0.223 0.166 1. 0.21 0.27 0.162]
 [0.231 0.326 0.189 0.191 0.211 0.226 0.251 0.287 0.224 0.26 0.205 0.333 0.191 0.236 0.159 0.228 0.275 0.189 0.21 1. 0.283 0.234]
 [0.245 0.261 0.247 0.184 0.238 0.261 0.268 0.262 0.23 0.244 0.258 0.263 0.231 0.202 0.203 0.261 0.266 0.187 0.27 0.283 1. 0.211]
 [0.183 0.218 0.159 0.183 0.175 0.238 0.218 0.184 0.203 0.192 0.188 0.212 0.173 0.164 0.147 0.302 0.203 0.354 0.162 0.234 0.211 1.]]

Distance matrix (1-similarity):

[[0. 0.793 0.814 0.846 0.781 0.808 0.611 0.798 0.72 0.725 0.689 0.768 0.792 0.807 0.861 0.785 0.751 0.795 0.712 0.769 0.755 0.817]
 [0.793 0. 0.815 0.809 0.755 0.764 0.747 0.722 0.806 0.768 0.796 0.738 0.815 0.775 0.853 0.771 0.762 0.818 0.795 0.674 0.739 0.782]
 [0.814 0.815 0. 0.785 0.717 0.851 0.807 0.809 0.773 0.769 0.799 0.807 0.805 0.838 0.856 0.794 0.818 0.814 0.729 0.811 0.753 0.841]
 [0.846 0.809 0.785 0. 0.826 0.837 0.826 0.807 0.844 0.829 0.861 0.804 0.839 0.843 0.856 0.818 0.809 0.839 0.856 0.809 0.816 0.817]
 [0.781 0.755 0.717 0.826 0. 0.81 0.76 0.776 0.709 0.716 0.729 0.788 0.801 0.817 0.858 0.794 0.752 0.821 0.677 0.789 0.762 0.825]
 [0.808 0.764 0.851 0.837 0.81 0. 0.788 0.789 0.813 0.819 0.798 0.776 0.771 0.792 0.794 0.751 0.802 0.762 0.822 0.774 0.739 0.762]
 [0.611 0.747 0.807 0.826 0.76 0.788 0. 0.753 0.719 0.711 0.677 0.745 0.788 0.819 0.852 0.724 0.761 0.799 0.709 0.749 0.732 0.782]
 [0.798 0.722 0.809 0.807 0.776 0.789 0.753 0. 0.795 0.786 0.778 0.69 0.785 0.828 0.824 0.746 0.729 0.813 0.826 0.713 0.738 0.816]
 [0.72 0.806 0.773 0.844 0.709 0.813 0.719 0.795 0. 0.723 0.742 0.742 0.795 0.819 0.85 0.763 0.781 0.817 0.595 0.776 0.77 0.797]
 [0.725 0.768 0.769 0.829 0.716 0.819 0.711 0.786 0.723 0. 0.742 0.745 0.788 0.83 0.855 0.739 0.739 0.828 0.72 0.74 0.756 0.808]
 [0.689 0.796 0.799 0.861 0.729 0.798 0.677 0.778 0.742 0.742 0. 0.798 0.769 0.849 0.838 0.773 0.777 0.829 0.729 0.795 0.742 0.812]
 [0.768 0.738 0.807 0.804 0.788 0.776 0.745 0.69 0.742 0.745 0.798 0. 0.782 0.808 0.824 0.727 0.728 0.818 0.762 0.667 0.737 0.788]
 [0.792 0.815 0.805 0.839 0.801 0.771 0.788 0.785 0.795 0.788 0.769 0.782 0. 0.85 0.822 0.779 0.803 0.818 0.808 0.809 0.769 0.827]
 [0.807 0.775 0.838 0.843 0.817 0.792 0.819 0.828 0.819 0.83 0.849 0.808 0.85 0. 0.839 0.831 0.812 0.844 0.822 0.764 0.798 0.836]
 [0.861 0.853 0.856 0.856 0.858 0.794 0.852 0.824 0.85 0.855 0.838 0.824 0.822 0.839 0. 0.824 0.826 0.863 0.862 0.841 0.797 0.853]
 [0.785 0.771 0.794 0.818 0.794 0.751 0.724 0.746 0.763 0.739 0.773 0.727 0.779 0.831 0.824 0. 0.757 0.707 0.795 0.772 0.739 0.698]
 [0.751 0.762 0.818 0.809 0.752 0.802 0.761 0.729 0.781 0.739 0.777 0.728 0.803 0.812 0.826 0.757 0. 0.823 0.777 0.725 0.734 0.797]
 [0.795 0.818 0.814 0.839 0.821 0.762 0.799 0.813 0.817 0.828 0.829 0.818 0.818 0.844 0.863 0.707 0.823 0. 0.834 0.811 0.813 0.646]
 [0.712 0.795 0.729 0.856 0.677 0.822 0.709 0.826 0.595 0.72 0.729 0.762 0.808 0.822 0.862 0.795 0.777 0.834 0. 0.79 0.73 0.838]
 [0.769 0.674 0.811 0.809 0.789 0.774 0.749 0.713 0.776 0.74 0.795 0.667 0.809 0.764 0.841 0.772 0.725 0.811 0.79 0. 0.717 0.766]
 [0.755 0.739 0.753 0.816 0.762 0.739 0.732 0.738 0.77 0.756 0.742 0.737 0.769 0.798 0.797 0.739 0.734 0.813 0.73 0.717 0. 0.789]
 [0.817 0.782 0.841 0.817 0.825 0.762 0.782 0.816 0.797 0.808 0.812 0.788 0.827 0.836 0.853 0.698 0.797 0.646 0.838 0.766 0.789 0.]]

Dendrogram



Cluster Labels for files:

```
>>> cluster.labels_
array([2, 0, 2, 0, 2, 1, 2, 0, 2, 2, 2, 0, 2, 3, 4, 1, 0, 1, 2, 0, 0, 1],
      dtype=int64)
```

File	Cluster
ass1-1019.txt	2
ass1-1037.txt	0
ass1-1046.txt	2
ass1-1138.txt	0
ass1-1147.txt	2
ass1-202.txt	1
ass1-211.txt	2
ass1-321.txt	0
ass1-440.txt	2
ass1-505.txt	2
ass1-532.txt	2
ass1-541.txt	0
ass1-606.txt	2
ass1-743.txt	3
ass1-817.txt	4
ass1-826.txt	1
ass1-909.txt	0
ass1_1349.txt	1
ass1_422.txt	2
ass1_734.txt	0
ass1_808.txt	0
ass1_936.txt	1

Observations:

a) We get decent values for jaccard index even though documents are on the same topic.
b) Smallest clusters are 3 and 4 which contain only 1 element each, “ass1-743.txt”, “ass1-817.txt” respectively. Both are most dissimilar in entire collection with respect to each document.
c) Largest cluster is 2. It contain 9 documents.