# REPORT

## Assignment 2c

-Shreshth Saxena (38)
Thongminlien Kuki (46)

### Objective:

Perform LSA using reduced latent space with 4 dimensions.
For each topic identify the set of 5 top weighted terms.
Find the similarity matrix for the documents in the reduced space.
Apply hierarchical clustering. Cut the dendrogram at k and identify clusters of similar documents.

### Packages:

NLTK (The Natural Language Toolkit): for text processing like tokenization, stemming, tagging and parsing.
scipy: for clustering
sklearn: for LSA and cosine similarity matrices' calculation
numpy: for Scientific Computing.
matplotlib: for plotting

### Latent Semantic Analysis :

```
TruncatedSVD(algorithm='randomized', n_components=4, n_iter=100,

    random_state=None, tol=0.0)
```

### Top 5 terms in each topic:

```
Topic 0:
engin
wa
page
use
index

Topic 1:
gopher
netscap
menu
resourc
five
```

Topic 2:
engin
wa
looksmart
purchas
webcrawl

Topic 3:
engin
voic
advertis
answer
googl

## Reduced 4d Document Vectors:

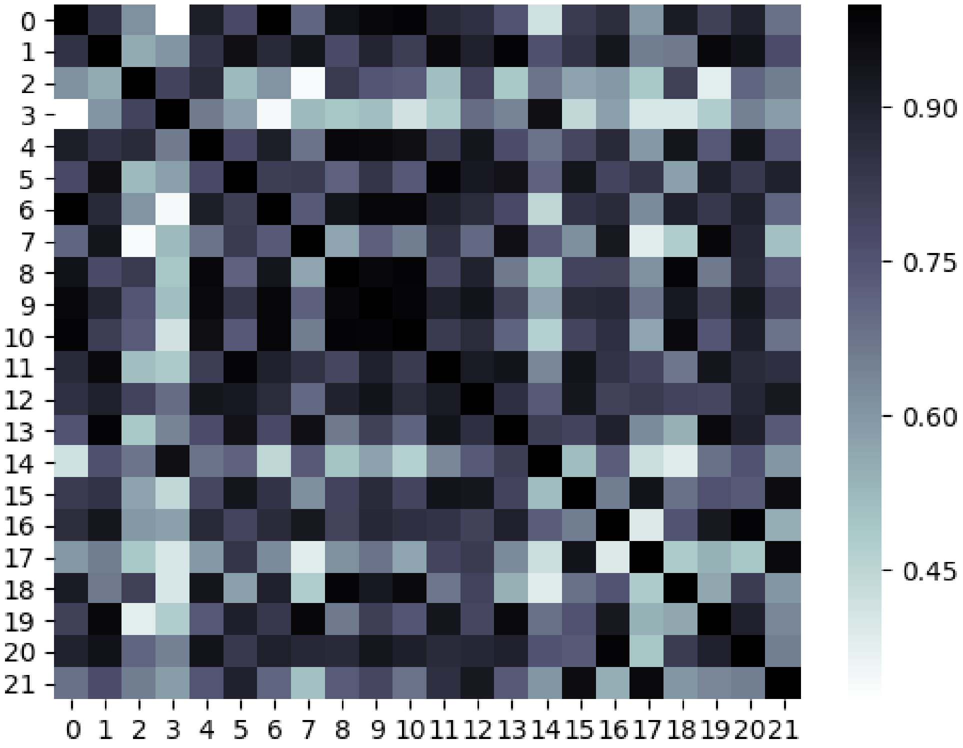
```
>>> doc_top.shape

(22, 4)

>>> doc_top
```

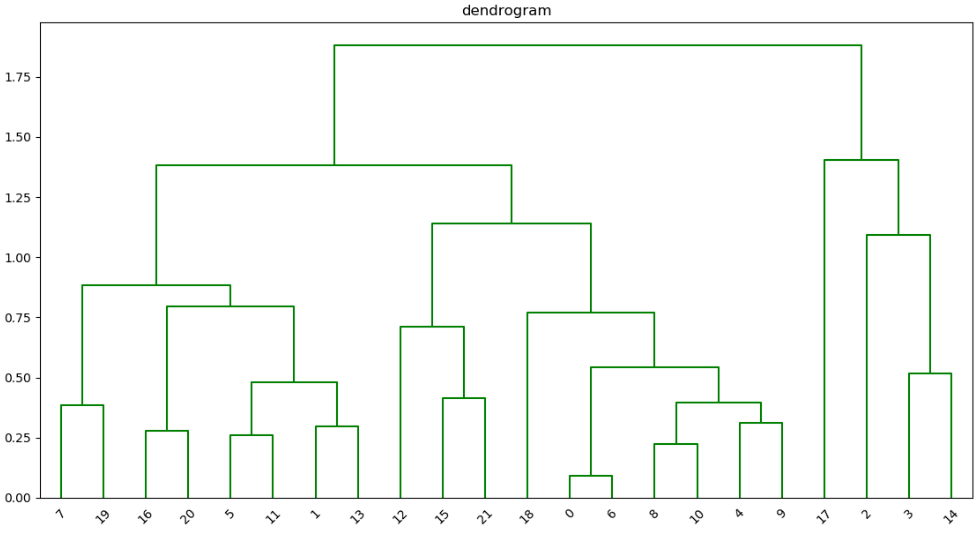
```
array([[ 0.92599498,  0.29947713, -0.11968999, -0.1962678 ],
       [ 0.95390416, -0.24337764, -0.13219241, -0.11558268],
       [ 0.73747105,  0.23170522,  0.12758101,  0.62142758],
       [ 0.62305539, -0.38727326,  0.02855758,  0.67897412],
       [ 0.95461739,  0.19771659, -0.06702441,  0.21241827],
       [ 0.92507245, -0.30430198,  0.14920939, -0.17139959],
       [ 0.93989245,  0.24970688, -0.10439102, -0.20820944],
       [ 0.8009208 , -0.38950333, -0.42251776, -0.16820156],
       [ 0.91996648,  0.37921176,  0.00299073,  0.09925302],
       [ 0.97841251,  0.20276467, -0.0293492 , -0.02709393],
       [ 0.92758729,  0.35525921, -0.1107149 , -0.03339034],
       [ 0.95137796, -0.16798124,  0.05373569, -0.25253665],
       [ 0.9766933 , -0.03348851,  0.18680989,  0.10025363],
       [ 0.90050727, -0.39864396, -0.14385122, -0.0973472 ],
       [ 0.70702141, -0.51953735, -0.11013652,  0.46698139],
       [ 0.90419773, -0.03572272,  0.36946352, -0.2112985 ],
       [ 0.9018036 , -0.07316626, -0.4258498 ,  0.00699392],
       [ 0.73198216, -0.11388281,  0.65037147, -0.1680767 ],
       [ 0.83805472,  0.52546596, -0.07669811,  0.12516876],
       [ 0.87641978, -0.29891738, -0.27362718, -0.26012484],
       [ 0.95207169, -0.01983316, -0.29451225,  0.08017899],
       [ 0.84579746, -0.12500055,  0.51862801, -0.0051492 ]])
```

## COSINE Similarity Heatmap:



## HEIRARICHAL CLUSTERING

## Dendrogram



## Cluster Labels for files:

```
>>> cluster.labels_
array([0, 4, 0, 0, 2, 2, 9, 0, 5, 7, 3, 8, 1, 3, 4, 0, 1, 1, 4, 5, 6, 4],
      dtype=int64)
```

| File | Cluster |
|---|---|
| ass1-1019.txt | 0 |
| ass1-1037.txt | 4 |
| ass1-1046.txt | 0 |
| ass1-1138.txt | 0 |
| ass1-1147.txt | 2 |
| ass1-202.txt | 2 |
| ass1-211.txt | 9 |
| ass1-321.txt | 0 |
| ass1-440.txt | 5 |
| ass1-505.txt | 7 |
| ass1-532.txt | 3 |
| ass1-541.txt | 8 |
| ass1-606.txt | 1 |
| ass1-743.txt | 3 |
| ass1-817.txt | 4 |
| ass1-826.txt | 0 |
| ass1-909.txt | 1 |
| ass1_1349.txt | 1 |
| ass1_422.txt | 4 |
| ass1_734.txt | 5 |
| ass1_808.txt | 6 |
| ass1_936.txt | 4 |