CMSC 476 Information Retrieval: Phase 5 Report

Hafsa Chaudry

Department of Computer Science

and Electrical Engineering

University of Maryland

Baltimore County

***Objective*—This assignment is to analyze the 504-document HTML test corpus,** **to execute agglomerative clustering, and introduce some concepts related to document clustering.**

# I. METHODS

1. *Testing Environment*

Throughout the experiment, one computer was used to run and test the data in order to ensure the accuracy and consistency of the results. The UMBC GL server was used to execute the algorithms for reasons of convenience and usability. The computer utilized for testing has the following specifications shown in Table 1 below.

*Table 1: Testing Computer Specifications*

|  |  |
| --- | --- |
| Processor | Intel Core i7 CPU 2.80 GHz |
| RAM | 16 GB |
| Operating System | Windows 64 Bit |

1. *Algorithms*

All algorithms were developed in Python and run using Python version 3.7.1. I utilized the python libraries from the previous phases and built on the code from previous submissions. In order to complete this assignment, I used chapter 17’s algorithm of Simple Hierarchical Agglomerative Clustering as a reference. I used cosine similarity to find the similarity scores between two documents and then used that score to build a similarity matrix, which holds all the cosine similarities for each document pairing. Then, the clustering algorithm is implemented and continues until no two clusters (or documents) have similarity greater than 0.4.

The command format was python phase5.py <input directory> <output directory>

1. *Issues*

My algorithm takes a long time to compile. Thus, when trying to update or change any part of the code, it takes a long time before being able to see the end result. Or being able to see that there is a single error.

# II. RESULTS

1. *Which pair of HTML documents is the most similar?*

The closest document is document 420.html and 417.html with a similarity of 1.0000000000000042. They are this way because they are the first two documents to merge together to become a cluster.

1. *Which pair of documents is the most dissimilar?*

The farthest document is document 069.html and 439.html with a similarity of 0.0. To find the most dissimilar pairing, you need to compare the similarity queue for each iteration with a global minimum, and update that minimum if the similarity queue is in fact less.

1. *Which document is the closest to the corpus centroid?*

From my understanding, the document that is the closest to the corpus centroid is cluster pair 165.html and 176.html.

# III. “Shell Code” Proof

A screenshot of a cell phone

Description automatically generated