CS 473-Fall 2019
Name: Shafay Haq
Purdue ID: haqs
Date: November 13, 2019

Project 2

Due Date: November 13th, 2019
Instructor: Chris Clifton
Taking one late day

Part 1: Parse the Corpus

Used Beutiful soup to parse sgml. extracted title, id and body for each document.

Stop word removal is done by the nltk libirary these are downloaded online. I also removed a regex since it eliminated punctuation

words are stemmed when d=processing each doc

Processing steps are as follows
for each doc if body not empty
extract title, body, topics, id
stem body words
calculate term Freq, TF
for each term:
create a term map, of t-¿ Term Object
add docs this term appears in
calculate IDF for this doc term pair
add column and row to similarity matrix for this doc
calculate cosine similarity between this doc and previously processed docs
end of parsing

Packages used are Beautiful Soup, nltk. for parsing and stop words.

Part 2: Run Complete and Single-Linkage Clustering

The complete linkage contains a smaller number of levels, vs the single linkage
The documents don't really have a sequence, a lot of them are grouped sparsely so they have very little similarity in general.
single linkage has results that went to deeper and more number of clusters
This might be due to the fact that the documents don't have much in common hence why the dedogram is wide

Part 3: Evaluation

Question 1

The measure is a score of the number of documents in a paticular topic with common clusters divided by the total number of documents in that cluster.
This will give us a measure of how well the docs in the same topic are grouped together!
we would calculate this as a per topic level and a average can be computed for all the topics for a certain linkage

Question 2

$\frac{d}{n}$, where d is the number of documents in a paticular topic with common clusters, and n is the total number of documents in that cluster.

Question 3

This metric is relevant since it shows us the ratio of the number of docs it clustered together with the same topic to the total number of docs.
Since if we assume that Topics is a good measure for relevance and hence clustering criteria, we can use this to measure how much the clustering of TFIDF and cosine similarity matches up with topics. this way if they are very similar we can consider the linkage and clusters to be good.

Question 4

Single Linkage: 435.74
Complete Linkage: 4.02

Question 5

The score for single linkage is much higher than complete. This is evident from the files generated which have terms and scores. The complete linkage has very small scores of this measure.
The single linkage has larger scores, and the average reflects this as well.
This score metric also takes into account the heirarchial structure since it counts all the parents of the heirarchy as well. The complete linkage also pairs documents with highest distance which is why these scores are evident. Given the data set and matches Single linkage is probably the best case here since some topics have a lot of docs in common.