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