Clustering Algorithms

Intro to Web Science

10 points

In this assignment you will be modifying two Python scripts from chapter 3 of Toby Segaran's book *Programming Collective Intelligence*. You can download the scripts from here:

http://kiwitobes.com/PCI_Code.zip

You are to first download 50 URLs of your choosing. They can be popular blogs, popular websites, Wikipedia articles, or whatever you think would be an interesting dataset. Modify the genatefeedvector.py script to download and produce the data file that contains a matrix of terms used by the web pages.

Then use hierarchical clustering or k-means in clusters.py to produce an image that visualizes the clusters in the dataset. You can earn 3 bonus points by using NodeXL to produce an image.

Submit your modified genatefeedvector.py script and the cluster image you produced on Easel before the due date.