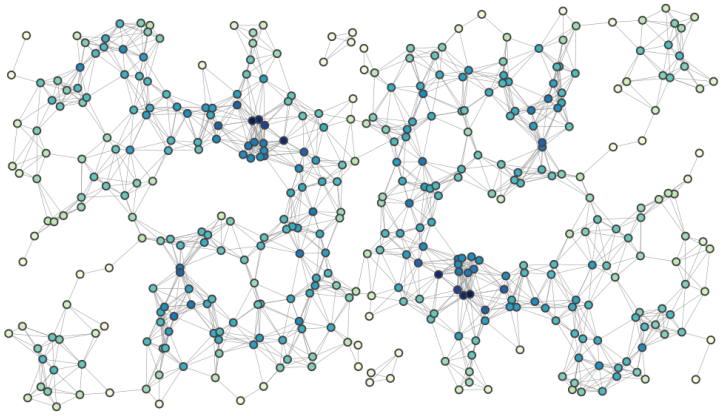


Old Dominion University

CS 432 Web Science

Assignment Seven

Derek Goddeau

Professor

Michael L. Nelson

April 15, 2017

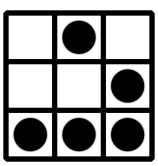
1 Create a blog-term matrix

To create the blog-term matrix I use the first 396 lines of the `glutton.py` program. To get the 98 random blog URIs use the `requests` package with a “next blog" request in a loop. Then I use `beautifulsoup` to parse out the RSS or Atom feed URIs out of the resulting page, and I save both to files so they do not need to be fetched again on new runs unless wanted.

To process each feed i use a while loop that iterates over all blog entries in the page and then attempts to get the next page, if an empty list is returned then the loop exits and returns the gathered data for the blog.

```
while entries:

    for entry in entries:
        .
        .
        .

    entries = []
    for link in feed_dict.feed.links:
        if link['rel'] == 'next':
        feed_dict = feedparser.parse(link['href'])
        entries = feed_dict.entries
```


Derek Goddeau (Datenstrom)
Old Dominion University



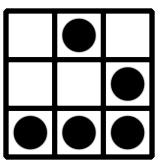
While processing each entry `beautifulsoup` is used to get the words from each entry, the code is very similar to code from a previous assignment.

```
soup = bs(html, 'lxml')


[s.extract() for s in soup(
['style', 'script', '[document]'])]


text = (''.join(string.findAll(text=True))
        for string in soup.findAll())


words = [(word.strip()).lower()
         for string in text
         for word in string.split()
         if word.isalpha()]


# Remove any empty strings, just in case
words = list(filter(len, words))
```

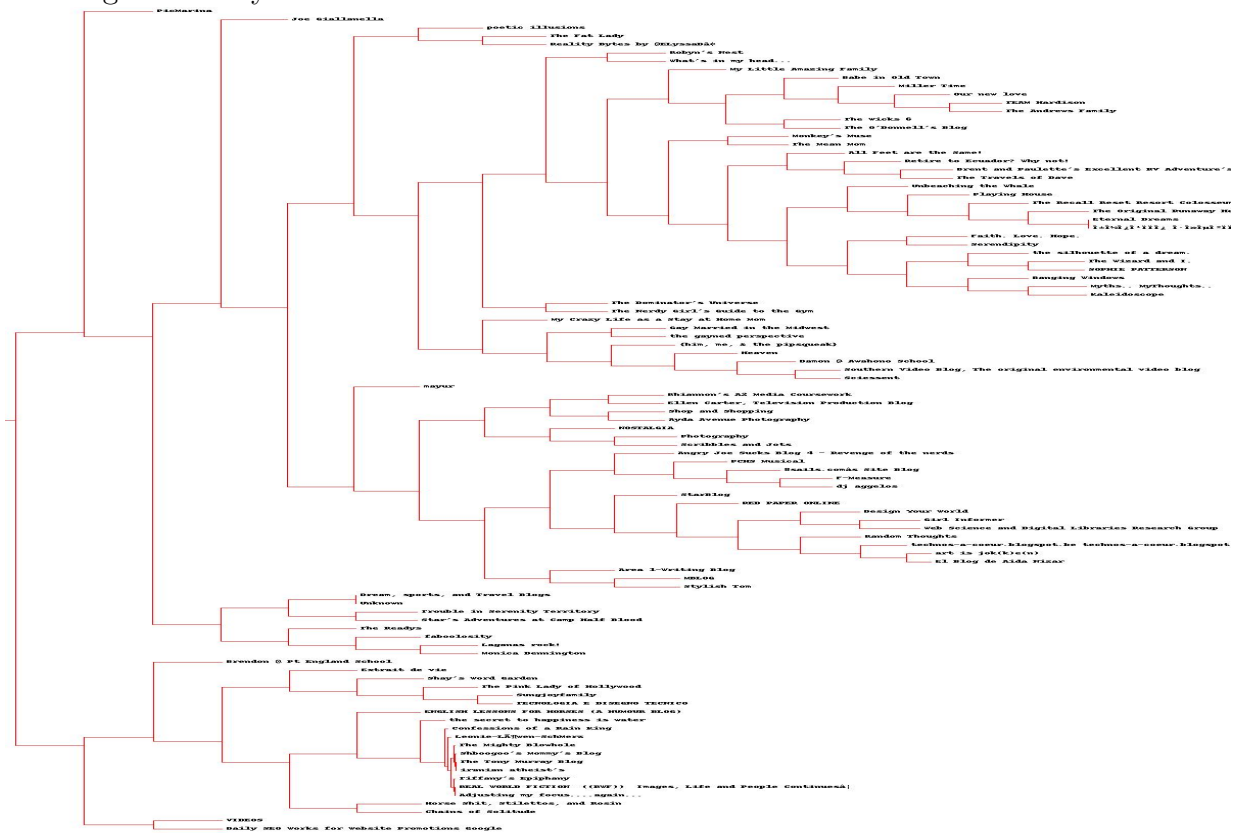
The data is then pickled and saved so that this step too can be skipped on future runs if desired. But I also dump the data to plain text using code very similar to the PCI code as requested.

2 Create ASCII and JPEG dendograms

The last 200 lines of code in `glutton.py` were an attempt to use "real" packages instead of the PCI to complete the rest. It uses `pandas` to read the pickled data into a data frame and

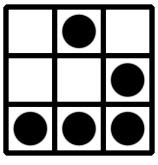


then attempts to process the data using `scipy.cluster.hierarchy`. First I transform the data frame and then get the correlation between the rows, and then condense the matrix. Unfortunately I was unable to get it to work, after eating $\frac{3}{4}$ of my 32GB of RAM it throws `ValueError: Distance matrix 'X' diagonal must be zero`. Out of time I resorted to following essentially the directions from the slides.



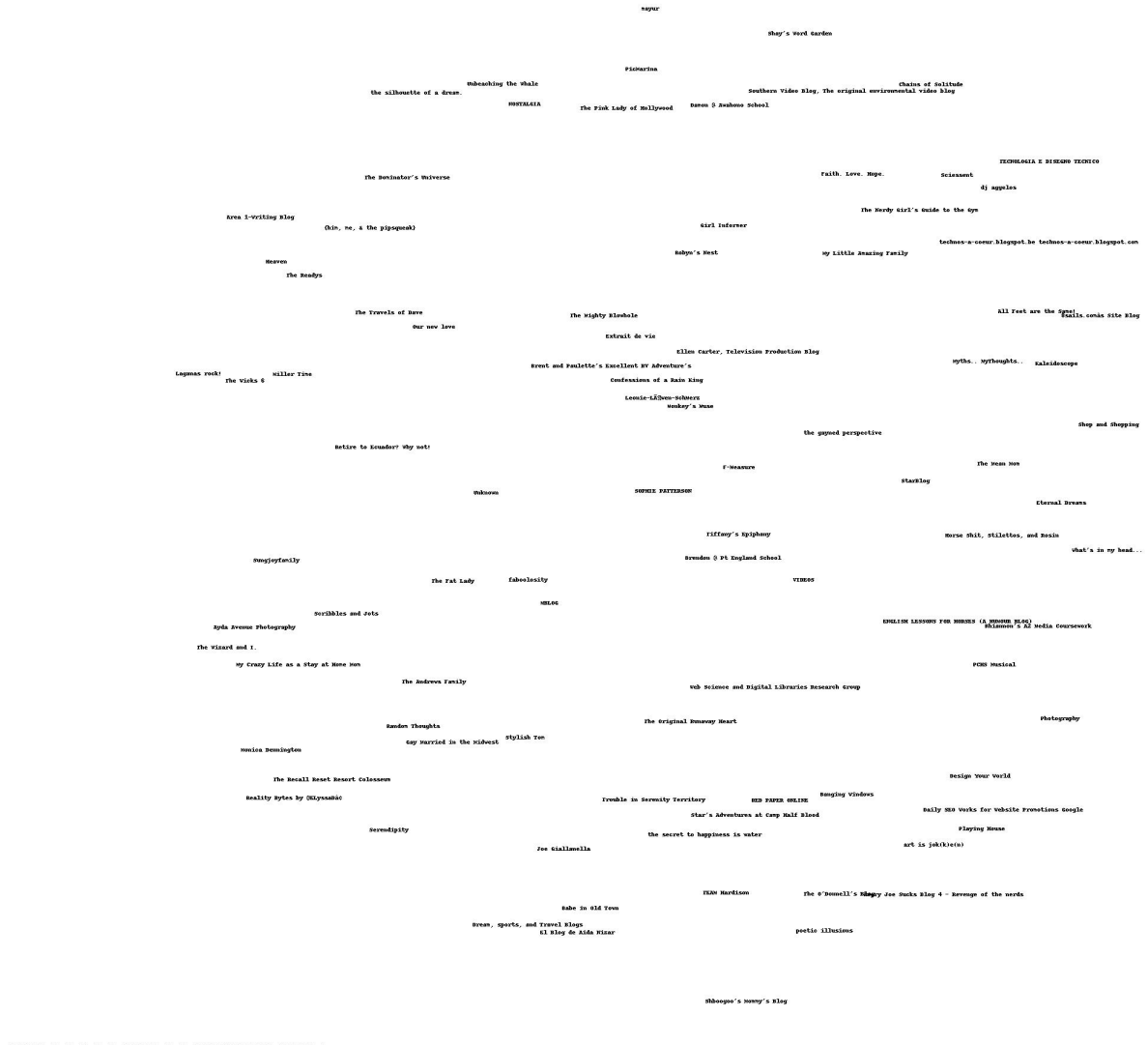
3 Cluster the blogs using K-Means

I wanted to perform this using SciPy also but it just isn't working for me. With the PCI code I got the following.



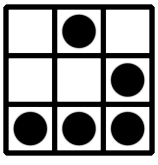
4 Use MDS to create a JPEG of the blogs

It only ran for three iterations.



5 Blog URIs

`http://zacky15.blogspot.com/?expref=next-blog`



http://falling-frombelow.blogspot.com/?expref=next-blog

http://ikeandbrandy.blogspot.com/?expref=next-blog

http://nerdygirlsguidetothegym.blogspot.com/?expref=next-blog

http://starsadventuresatcamphalfblood.blogspot.com/?expref=next-blog

http://gaymarriedinthemidwest.blogspot.com/?expref=next-blog

http://pesbrendonl.blogspot.com/?expref=next-blog

http://tecnologiaedisegnotecnico.blogspot.com/?expref=next-blog

http://thegirlinthemoonlight.blogspot.com/?expref=next-blog

http://pchsmusical.blogspot.com/?expref=next-blog

http://area1randwick.blogspot.com/?expref=next-blog

http://thereadysblog.blogspot.com/?expref=next-blog

http://sdjohns1.blogspot.com/?expref=next-blog

http://trthompson11.blogspot.com/?expref=next-blog

http://djaggelosr.blogspot.com/?expref=next-blog

http://tonymurrayblog.blogspot.com/?expref=next-blog

http://heavenmspie.blogspot.com/?expref=next-blog

http://aydaavenuephotography.blogspot.com/?expref=next-blog

http://thedominatorsuniverse.blogspot.com/?expref=next-blog

http://ecuador-bound.blogspot.com/?expref=next-blog

http://thetravelsofdave.blogspot.com/?expref=next-blog

http://redpaperonline.blogspot.com/?expref=next-blog

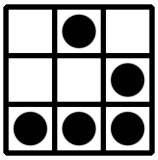
http://rewritingthewest.blogspot.com/?expref=next-blog

http://leonieloewenschmerz.blogspot.com/?expref=next-blog

http://tylerandlaurenmiller.blogspot.com/?expref=next-blog

http://angryjoesuckspart4.blogspot.com/?expref=next-blog

http://blazingdice.blogspot.com/?expref=next-blog



http://thefinalwordtv.blogspot.com/?expref=next-blog

http://www.chainsofsolitude.com/?expref=next-blog

http://art-is-jokken.blogspot.com/?expref=next-blog

http://thegaynedperspective.blogspot.com/?expref=next-blog

http://sayyestoadventure.blogspot.com/?expref=next-blog

https://duhparenting101.blogspot.com/?expref=next-blog

http://realworldfiction.blogspot.com/?expref=next-blog

http://designingyourworlds.blogspot.com/?expref=next-blog

http://vhui.blogspot.com/?expref=next-blog

http://mayur210.blogspot.com/?expref=next-blog

http://poeticallydead.blogspot.com/?expref=next-blog

http://jodythemeanmom.blogspot.com/?expref=next-blog

http://hansandpeter.blogspot.com/?expref=next-blog

https://shopandshoping.blogspot.com/?expref=next-blog

http://playinghouse-jenn.blogspot.com/?expref=next-blog

http://odonnellsblog.blogspot.com/?expref=next-blog

http://awahonodamons.blogspot.com/?expref=next-blog

http://markeldersonmarkelderson.blogspot.com/?expref=next-blog

http://adjustingmyfocusagain.blogspot.com/?expref=next-blog

http://fireblossom-wordgarden.blogspot.com/?expref=next-blog

http://rhiannontolea2media.blogspot.com/?expref=next-blog

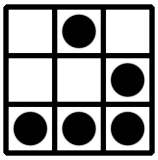
http://girlinformer.blogspot.com/?expref=next-blog

http://shesaysgoodmorningtome.blogspot.com/?expref=next-blog

http://picmarina.blogspot.com/?expref=next-blog

http://scribblingsandjots.blogspot.com/?expref=next-blog

http://sungjoyfamily.blogspot.com/?expref=next-blog



http://marissabenni.blogspot.com/?expref=next-blog

http://pearlwhite-chuchu.blogspot.com/?expref=next-blog

http://aminmoz.blogspot.com/?expref=next-blog

http://www.thepinkladyofhollywood.com/?expref=next-blog

http://1lovelylifetolive.blogspot.com/?expref=next-blog

http://findingeunice.blogspot.com/?expref=next-blog

http://sciessent.blogspot.com/?expref=next-blog

http://stylishtom.blogspot.com/?expref=next-blog

http://wicks6.blogspot.com/?expref=next-blog

http://unbeachingthewhale.blogspot.com/?expref=next-blog

http://ws-dl.blogspot.com/

http://diamflorakis4.blogspot.com/?expref=next-blog

http://johnrandomthoughts.blogspot.com/?expref=next-blog

http://monkeysmuse.blogspot.com/?expref=next-blog

http://coochi-coo.blogspot.com/?expref=next-blog

http://meganmccartney.blogspot.com/?expref=next-blog

http://bangingwindows.blogspot.com/?expref=next-blog

http://maggiesayers.blogspot.com/?expref=next-blog

http://southvidblog.blogspot.com/?expref=next-blog

http://sydneebreit.blogspot.com/?expref=next-blog

http://shboogoo.blogspot.com/?expref=next-blog

http://allfeetarethesame.blogspot.com/?expref=next-blog

http://faboolosity.blogspot.com/?expref=next-blog

http://babeinoldetown.blogspot.com/?expref=next-blog

http://darknightdurant.blogspot.com/?expref=next-blog

http://wordsinorange.blogspot.com/?expref=next-blog

http://mylittleamazingfamily.blogspot.com/?expref=next-blog

http://elblogdeaidanizar.blogspot.com/?expref=next-blog

http://dharmaphotography.blogspot.com/?expref=next-blog

http://giallanella.blogspot.com/?expref=next-blog

http://f-measure.blogspot.com/

http://runperryrun-runperryrun.blogspot.com/?expref=next-blog

https://tkozzz.blogspot.com/?expref=next-blog

http://confessionsofarainking.blogspot.com/?expref=next-blog

http://ellencartertvproduction.blogspot.com/?expref=next-blog

http://starboy91.blogspot.com/?expref=next-blog

http://laganasrock.blogspot.com/?expref=next-blog

https://myrunawayheart.blogspot.com/?expref=next-blog

http://harshasrao.blogspot.com/?expref=next-blog

http://matthewandann.blogspot.com/?expref=next-blog

http://kalyanisthoughts.blogspot.com/?expref=next-blog

http://8sails.blogspot.com/?expref=next-blog

http://gemini6870.blogspot.com/?expref=next-blog

http://mightyblowhole.blogspot.com/?expref=next-blog

https://technos-a-coeur.blogspot.com/?expref=next-blog

http://yakshi17.blogspot.com/?expref=next-blog

http://seoforwebpromotion.blogspot.com/?expref=next-blog