Assignment Eight

Jacob Berlin

CS432 – Spring 2016

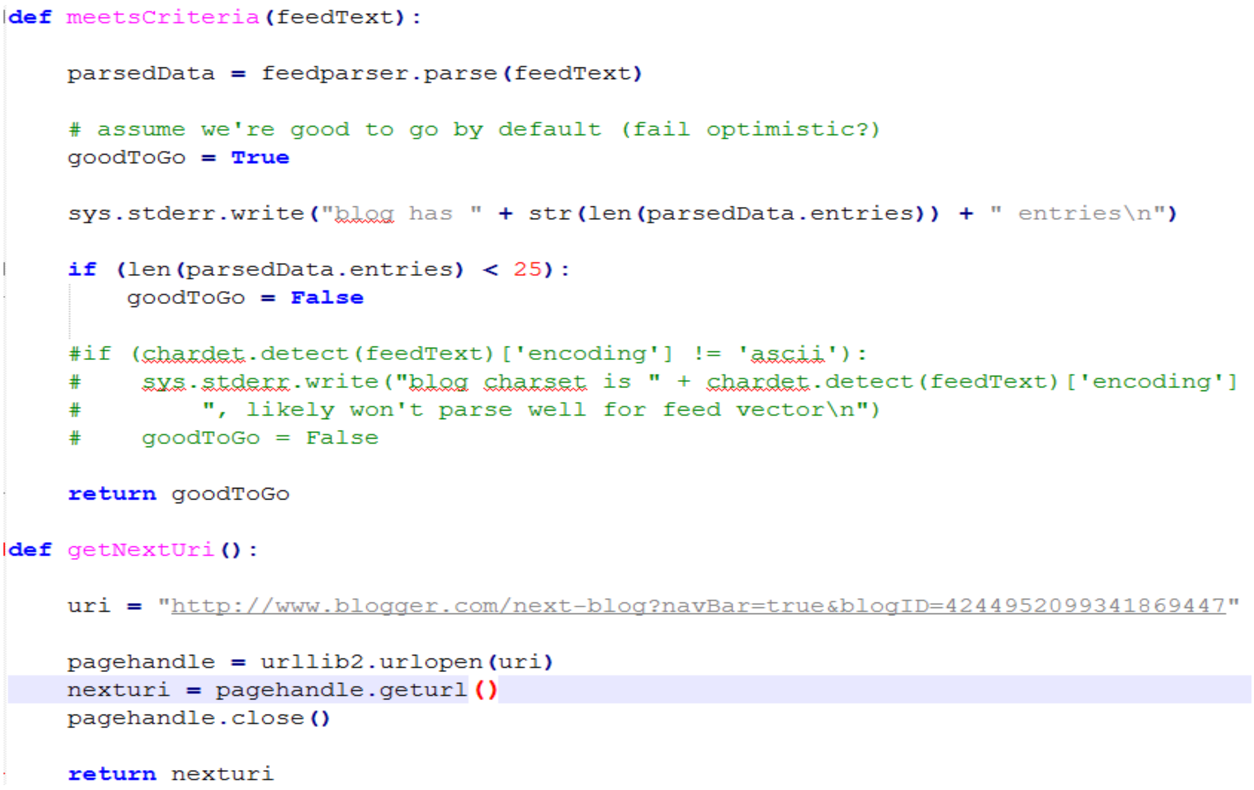
1. Create a blog-term matrix. Start by grabbing 100 blogs; include:

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

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

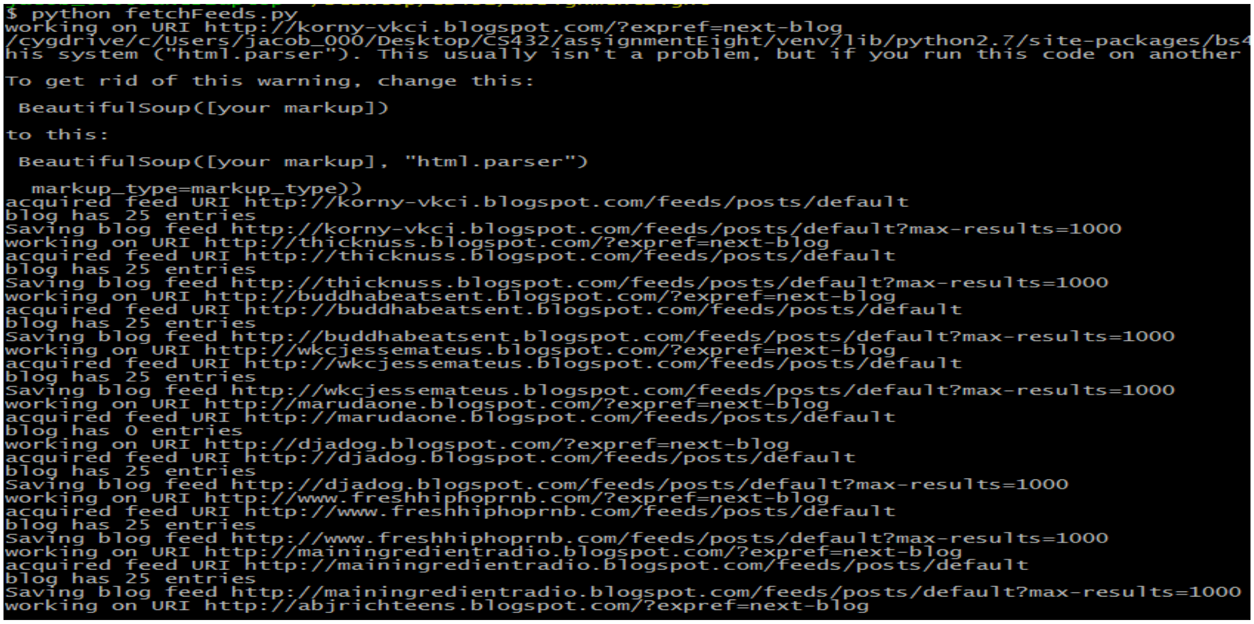
and grab 98 more as per the method shown in class. Note that this method randomly chooses blogs and each student will separately do this process, so it is unlikely that these 98 blogs will be shared among students. In other words, no sharing of blog data. Upload to github your code for grabbing the blogs and provide a list of blog URIs, both in the report and in github..

Use the blog title as the identifier for each blog (and row of the matrix). Use the terms from every item/title (RSS) or entry/title (Atom) for the columns of the matrix. The values are the frequency of occurrence. Essentially you are replicating the format of the "blogdata.txt" file included with the PCI book code. Limit the number of terms to the most "popular" (i.e., frequent) 500 terms, this is *after* the criteria on p. 32 (slide 7) has been satisfied.

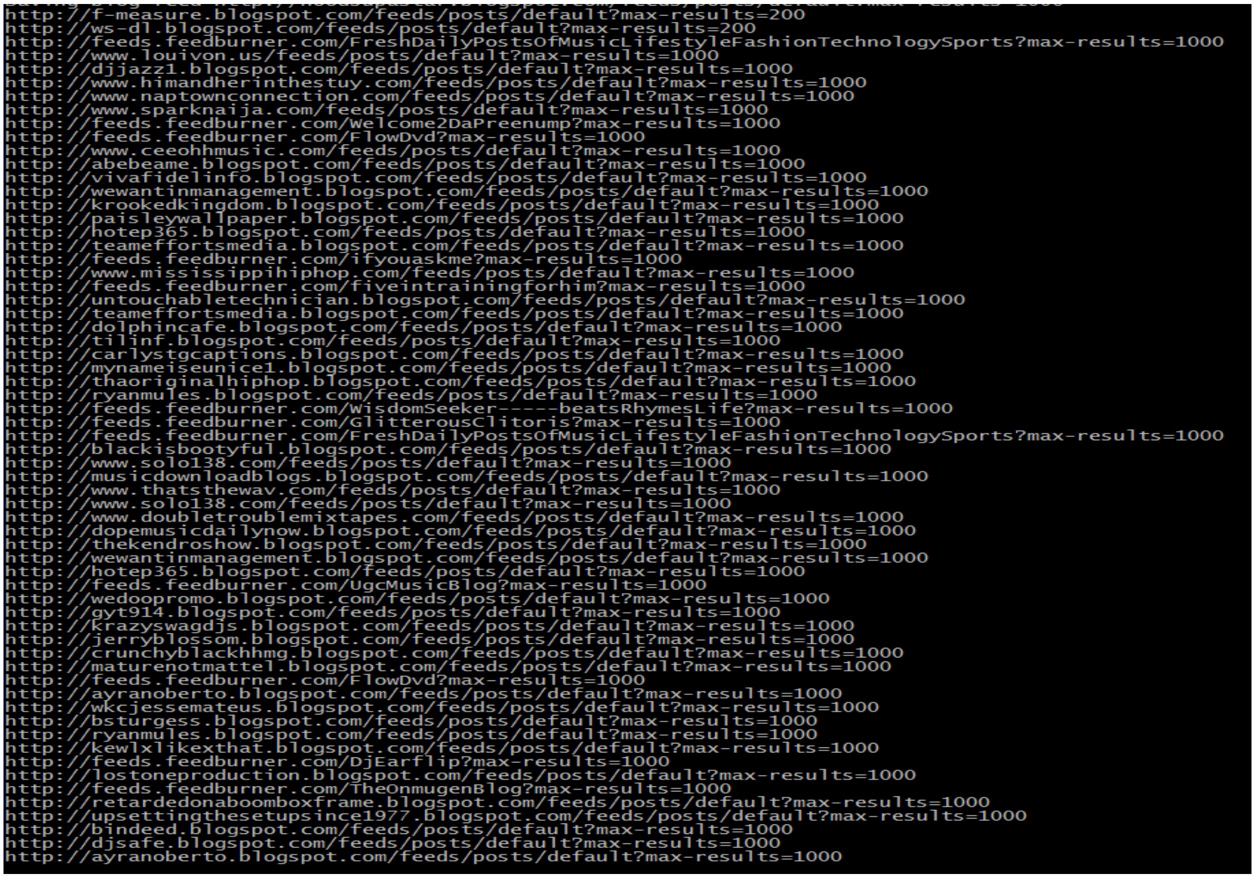


```
def meetsCriteria(feedText):

    parsedData = feedparser.parse(feedText)

    # assume we're good to go by default (fail optimistic?)
    goodToGo = True

    sys.stderr.write("blog has " + str(len(parsedData.entries)) + " entries\n")

    if (len(parsedData.entries) < 25):
        goodToGo = False

    #if (chardet.detect(feedText)['encoding'] != 'ascii'):
    #    sys.stderr.write("blog charset is " + chardet.detect(feedText)['encoding']
    #        ", likely won't parse well for feed vector\n")
    #    goodToGo = False

    return goodToGo

def getNextUri():

    uri = "http://www.blogger.com/next-blog?navBar=true&blogID=4244952099341869447"

    pagehandle = urllib2.urlopen(uri)
    nexturi = pagehandle.geturl()
    pagehandle.close()

    return nexturi
```

To start this problem, I used the file ‘fetchFeeds.py’ made by Shawn M. Jones which would take a particular blog ID (found in the html of the blog page) and would use the URI “http://www.blogger.com/next-blog?navBar=true&blogID=4244952099341869447” to list a random blog. This was achieved by using both the Urllib and BeautifulSoup libraries and with a small edit on my side posted all of the received blog’s URIs to ‘feedNames.txt’. I used two different starting blog IDs to get my list of 98 other blogs; ‘3395148574627097365’ and ‘4244952099341869447.’

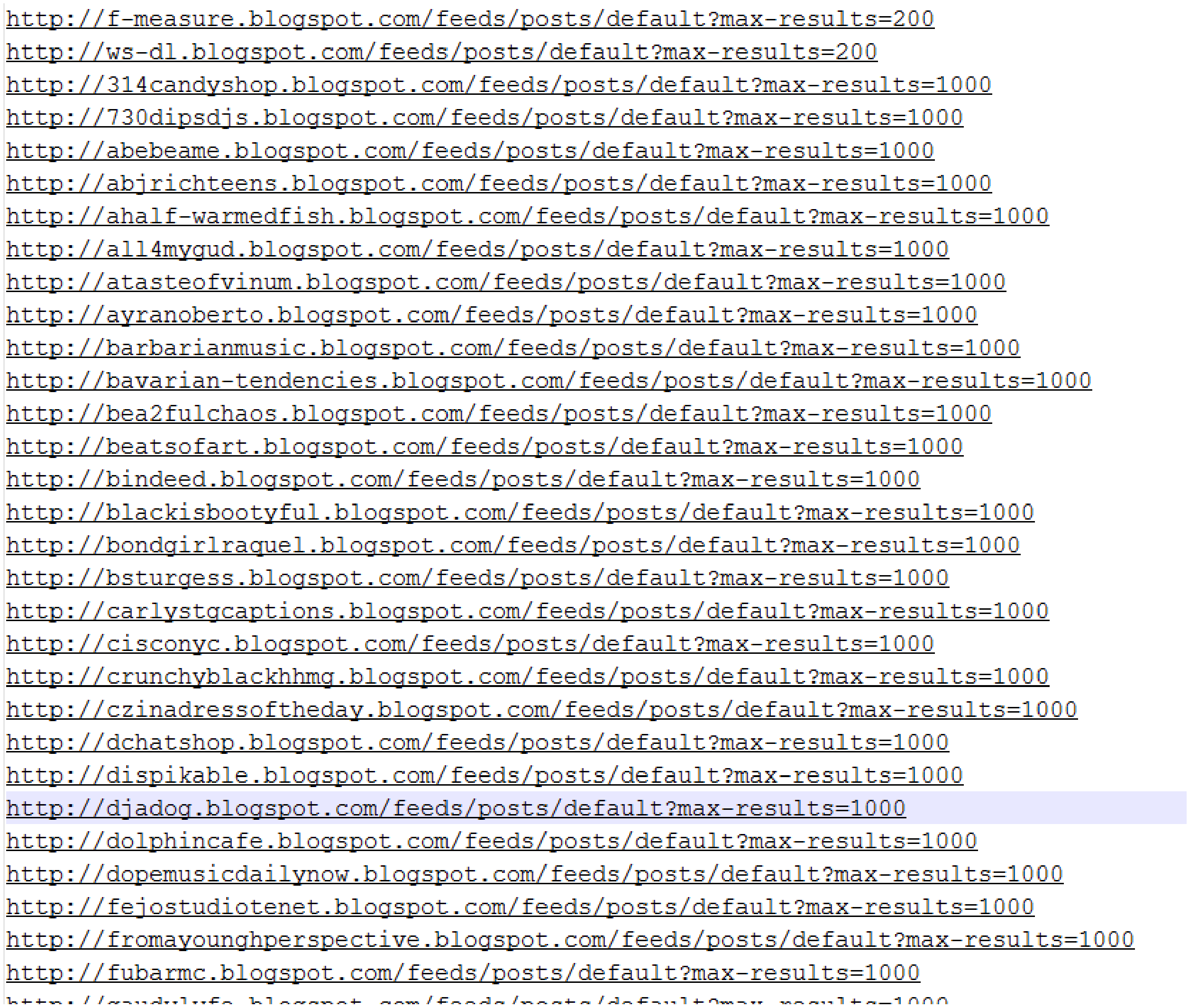


```
$ python fetchFeeds.py
working on URI http://korny-vkci.blogspot.com/?expref=next-blog
/cygdrive/c/Users/jacob_000/Desktop/CS432/assignmentEight/venv/lib/python2.7/site-packages/bs4
his system ("html.parser"). This usually isn't a problem, but if you run this code on another

To get rid of this warning, change this:

 BeautifulSoup([your markup])

to this:

 BeautifulSoup([your markup], "html.parser")

  markup_type=markup_type))
acquired feed URI http://korny-vkci.blogspot.com/feeds/posts/default
blog has 25 entries
Saving blog feed http://korny-vkci.blogspot.com/feeds/posts/default?max-results=1000
working on URI http://thicknuss.blogspot.com/?expref=next-blog
acquired feed URI http://thicknuss.blogspot.com/feeds/posts/default
blog has 25 entries
Saving blog feed http://thicknuss.blogspot.com/feeds/posts/default?max-results=1000
working on URI http://buddhabeatsent.blogspot.com/?expref=next-blog
acquired feed URI http://buddhabeatsent.blogspot.com/feeds/posts/default
blog has 25 entries
Saving blog feed http://buddhabeatsent.blogspot.com/feeds/posts/default?max-results=1000
working on URI http://wkcjessemateus.blogspot.com/?expref=next-blog
acquired feed URI http://wkcjessemateus.blogspot.com/feeds/posts/default
blog has 25 entries
Saving blog feed http://wkcjessemateus.blogspot.com/feeds/posts/default?max-results=1000
working on URI http://marudaone.blogspot.com/?expref=next-blog
acquired feed URI http://marudaone.blogspot.com/feeds/posts/default
blog has 0 entries
working on URI http://djadog.blogspot.com/?expref=next-blog
acquired feed URI http://djadog.blogspot.com/feeds/posts/default
blog has 25 entries
Saving blog feed http://djadog.blogspot.com/feeds/posts/default?max-results=1000
working on URI http://www.freshhiphoprnb.com/?expref=next-blog
acquired feed URI http://www.freshhiphoprnb.com/feeds/posts/default
blog has 25 entries
Saving blog feed http://www.freshhiphoprnb.com/feeds/posts/default?max-results=1000
working on URI http://mainingredientradio.blogspot.com/?expref=next-blog
acquired feed URI http://mainingredientradio.blogspot.com/feeds/posts/default
blog has 25 entries
Saving blog feed http://mainingredientradio.blogspot.com/feeds/posts/default?max-results=1000
working on URI http://abjrichteens.blogspot.com/?expref=next-blog
```



```
http://f-measure.blogspot.com/feeds/posts/default?max-results=200
http://ws-dl.blogspot.com/feeds/posts/default?max-results=200
http://feeds.feedburner.com/FreshDailyPostsOfMusicLifestyleFashionTechnologySports?max-results=1000
http://www.louivon.us/feeds/posts/default?max-results=1000
http://djjazz1.blogspot.com/feeds/posts/default?max-results=1000
http://www.himandherinthestuy.com/feeds/posts/default?max-results=1000
http://www.naptownconnection.com/feeds/posts/default?max-results=1000
http://www.sparknaija.com/feeds/posts/default?max-results=1000
http://feeds.feedburner.com/Welcome2DaPreenump?max-results=1000
http://feeds.feedburner.com/FlowDvd?max-results=1000
http://www.ceeohhmusic.com/feeds/posts/default?max-results=1000
http://abebeame.blogspot.com/feeds/posts/default?max-results=1000
http://vivafidelinfo.blogspot.com/feeds/posts/default?max-results=1000
http://wewantinmanagement.blogspot.com/feeds/posts/default?max-results=1000
http://krookedkingdom.blogspot.com/feeds/posts/default?max-results=1000
http://paisleywallpaper.blogspot.com/feeds/posts/default?max-results=1000
http://hotep365.blogspot.com/feeds/posts/default?max-results=1000
http://teameffortsmedia.blogspot.com/feeds/posts/default?max-results=1000
http://feeds.feedburner.com/ifyouaskme?max-results=1000
http://www.mississippihiphop.com/feeds/posts/default?max-results=1000
http://feeds.feedburner.com/fiveintrainingforhim?max-results=1000
http://untouchabletechnician.blogspot.com/feeds/posts/default?max-results=1000
http://teameffortsmedia.blogspot.com/feeds/posts/default?max-results=1000
http://dolphincafe.blogspot.com/feeds/posts/default?max-results=1000
http://tilinf.blogspot.com/feeds/posts/default?max-results=1000
http://carlystgcaptions.blogspot.com/feeds/posts/default?max-results=1000
http://mynameiseunice1.blogspot.com/feeds/posts/default?max-results=1000
http://thaoriginalhiphop.blogspot.com/feeds/posts/default?max-results=1000
http://ryanmules.blogspot.com/feeds/posts/default?max-results=1000
http://feeds.feedburner.com/WisdomSeeker-----beatsRhymesLife?max-results=1000
http://feeds.feedburner.com/GlitterousClitoris?max-results=1000
http://feeds.feedburner.com/FreshDailyPostsOfMusicLifestyleFashionTechnologySports?max-results=1000
http://blackisbootyful.blogspot.com/feeds/posts/default?max-results=1000
http://www.solo138.com/feeds/posts/default?max-results=1000
http://musicdownloadblogs.blogspot.com/feeds/posts/default?max-results=1000
http://www.thatsthewav.com/feeds/posts/default?max-results=1000
http://www.solo138.com/feeds/posts/default?max-results=1000
http://www.doubletroublemixtapes.com/feeds/posts/default?max-results=1000
http://dopemusicdailynow.blogspot.com/feeds/posts/default?max-results=1000
http://thekendroshow.blogspot.com/feeds/posts/default?max-results=1000
http://wewantinmanagement.blogspot.com/feeds/posts/default?max-results=1000
http://hotep365.blogspot.com/feeds/posts/default?max-results=1000
http://feeds.feedburner.com/UgcMusicBlog?max-results=1000
http://wedoopromo.blogspot.com/feeds/posts/default?max-results=1000
http://gyt914.blogspot.com/feeds/posts/default?max-results=1000
http://krazyswagdjs.blogspot.com/feeds/posts/default?max-results=1000
http://jerryblossom.blogspot.com/feeds/posts/default?max-results=1000
http://crunchyblackhhmg.blogspot.com/feeds/posts/default?max-results=1000
http://maturenotmattel.blogspot.com/feeds/posts/default?max-results=1000
http://feeds.feedburner.com/FlowDvd?max-results=1000
http://ayranoberto.blogspot.com/feeds/posts/default?max-results=1000
http://wkcjessemateus.blogspot.com/feeds/posts/default?max-results=1000
http://bsturgess.blogspot.com/feeds/posts/default?max-results=1000
http://ryanmules.blogspot.com/feeds/posts/default?max-results=1000
http://kewlxlikexthat.blogspot.com/feeds/posts/default?max-results=1000
http://feeds.feedburner.com/DjEarflip?max-results=1000
http://lostoneproduction.blogspot.com/feeds/posts/default?max-results=1000
http://feeds.feedburner.com/TheOnmugenBlog?max-results=1000
http://retardedonaboomboxframe.blogspot.com/feeds/posts/default?max-results=1000
http://upsettingthesetupsince1977.blogspot.com/feeds/posts/default?max-results=1000
http://bindeed.blogspot.com/feeds/posts/default?max-results=1000
http://djsafe.blogspot.com/feeds/posts/default?max-results=1000
http://ayranoberto.blogspot.com/feeds/posts/default?max-results=1000
```

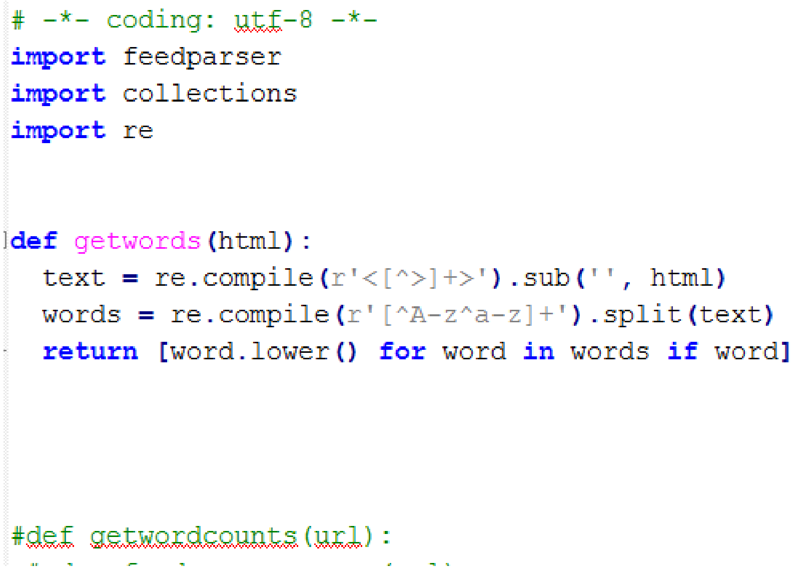
When running the program through the command prompt, I received my initial list of blogs that contained both duplicated and graphic blogs. I removed these by using the UNIX command ‘sort –u feedNames.txt’ and by manually going through to delete the provocative blogs. Each URI in this case has the max limit of pages set to 1,000 due to me not finding any file that might have more than that.



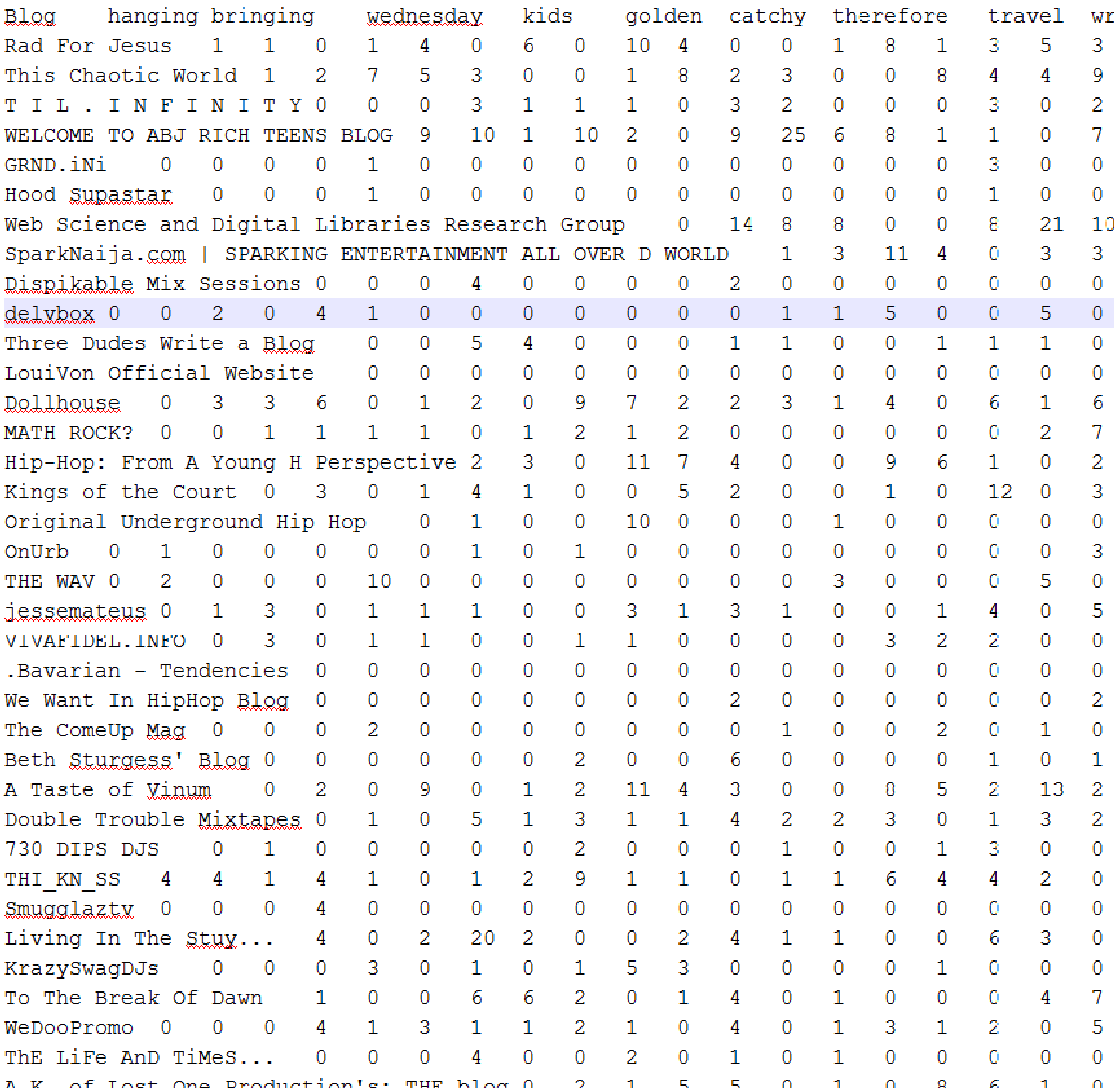
```
http://f-measure.blogspot.com/feeds/posts/default?max-results=200
http://ws-dl.blogspot.com/feeds/posts/default?max-results=200
http://314candyshop.blogspot.com/feeds/posts/default?max-results=1000
http://730dipsdjs.blogspot.com/feeds/posts/default?max-results=1000
http://abebeame.blogspot.com/feeds/posts/default?max-results=1000
http://abjrichteens.blogspot.com/feeds/posts/default?max-results=1000
http://ahalf-warmedfish.blogspot.com/feeds/posts/default?max-results=1000
http://all4mygud.blogspot.com/feeds/posts/default?max-results=1000
http://atasteofvinum.blogspot.com/feeds/posts/default?max-results=1000
http://ayranoberto.blogspot.com/feeds/posts/default?max-results=1000
http://barbarianmusic.blogspot.com/feeds/posts/default?max-results=1000
http://bavarian-tendencies.blogspot.com/feeds/posts/default?max-results=1000
http://bea2fulchaos.blogspot.com/feeds/posts/default?max-results=1000
http://beatsofart.blogspot.com/feeds/posts/default?max-results=1000
http://bindeed.blogspot.com/feeds/posts/default?max-results=1000
http://blackisbootyful.blogspot.com/feeds/posts/default?max-results=1000
http://bondgirlraquel.blogspot.com/feeds/posts/default?max-results=1000
http://bsturgess.blogspot.com/feeds/posts/default?max-results=1000
http://carlystgcaptions.blogspot.com/feeds/posts/default?max-results=1000
http://cisconyc.blogspot.com/feeds/posts/default?max-results=1000
http://crunchyblackhhmg.blogspot.com/feeds/posts/default?max-results=1000
http://czinadressoftheday.blogspot.com/feeds/posts/default?max-results=1000
http://dchatshop.blogspot.com/feeds/posts/default?max-results=1000
http://dispikable.blogspot.com/feeds/posts/default?max-results=1000
http://djadog.blogspot.com/feeds/posts/default?max-results=1000
http://dolphincafe.blogspot.com/feeds/posts/default?max-results=1000
http://dopemusicdailynow.blogspot.com/feeds/posts/default?max-results=1000
http://fejostudiotenet.blogspot.com/feeds/posts/default?max-results=1000
http://fromayounghperspective.blogspot.com/feeds/posts/default?max-results=1000
http://fubarmc.blogspot.com/feeds/posts/default?max-results=1000
http://gaudylufa.blogspot.com/feeds/posts/default?max-results=1000
```

After the sorting and compilation I came out with the revised ‘feedNames.txt’ shown partially above.

For the second part of this question I used the file ‘generateFeed.py’ made by the creators of PCI, which took all of the blogs that I received earlier and created an index based off of the most popular words for each. This was limited to 500 words at most and ordered everything as required. The output was placed into ‘blogdata.txt’



```
# -*- coding: utf-8 -*-
import feedparser
import collections
import re


def getwords(html):
  text = re.compile(r'<[^>]+>').sub('', html)
  words = re.compile(r'[^A-z^a-z]+').split(text)
  return [word.lower() for word in words if word]




#def getwordcounts(url):
```



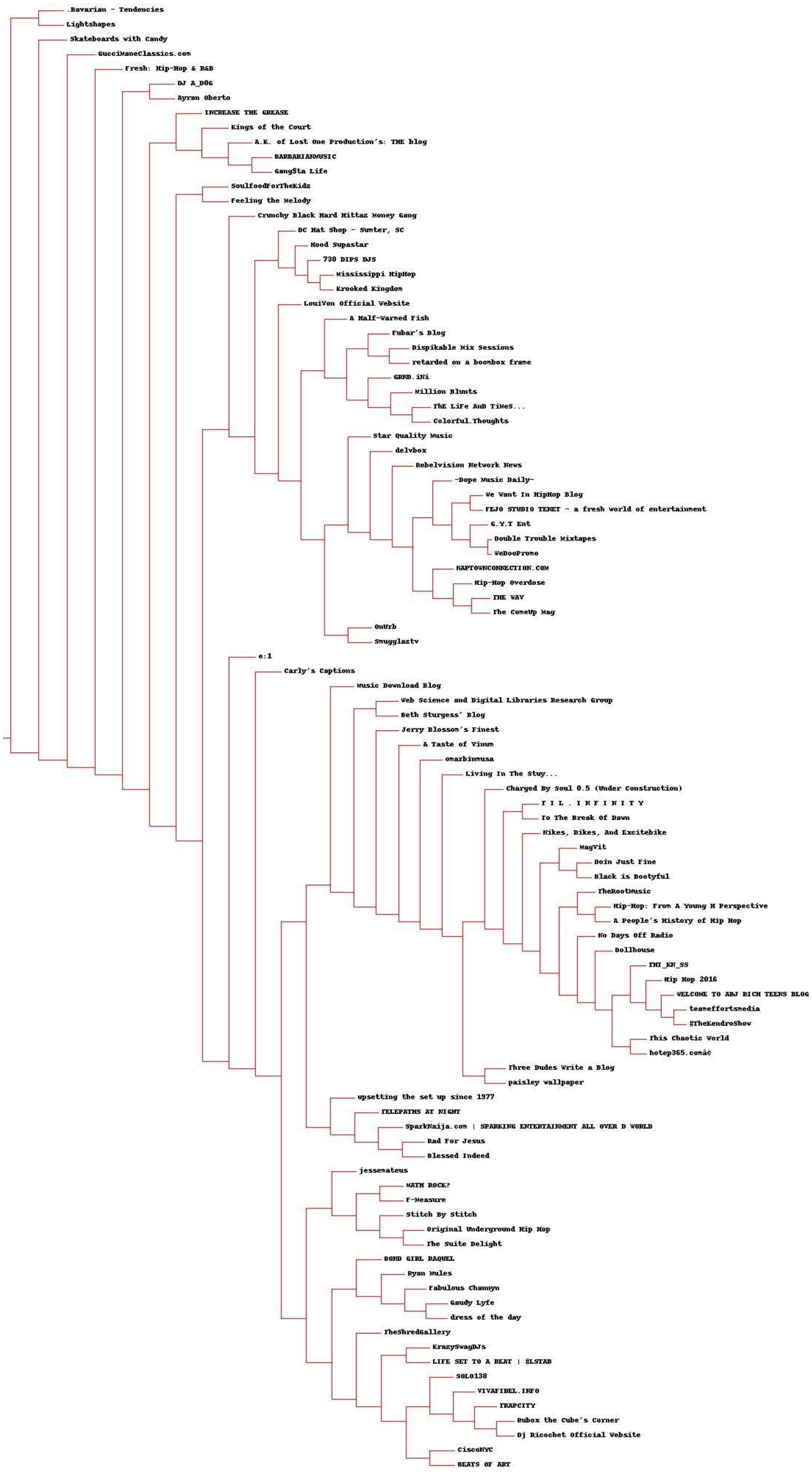
```
Blog    hanging bringing    wednesday   kids    golden  catchy  therefore   travel  wr
Rad For Jesus   1   1   0   1   4   0   6   0   10  4   0   0   1   8   1   3   5   3
This Chaotic World  1   2   7   5   3   0   0   1   8   2   3   0   0   8   4   4   9
T I L . I N F I N I T Y 0   0   0   3   1   1   1   0   3   2   0   0   0   3   0   2
WELCOME TO ABJ RICH TEENS BLOG  9   10  1   10  2   0   9   25  6   8   1   1   0   7
GRND.iNi    0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   3   0   0
Hood Supastar   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   1   0   0
Web Science and Digital Libraries Research Group    0   14  8   8   0   0   8   21  10
SparkNaija.com | SPARKING ENTERTAINMENT ALL OVER D WORLD    1   3   11  4   0   3   3
Dispikable Mix Sessions 0   0   0   4   0   0   0   0   2   0   0   0   0   0   0   0
delvbox 0   0   2   0   4   1   0   0   0   0   0   0   0   1   1   5   0   0   5   0
Three Dudes Write a Blog    0   0   5   4   0   0   0   1   1   0   0   1   1   1   0
LouiVon Official Website    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
Dollhouse   0   3   3   6   0   1   2   0   9   7   2   2   3   1   4   0   6   1   6
MATH ROCK?  0   0   1   1   1   1   0   1   2   1   2   0   0   0   0   0   0   2   7
Hip-Hop: From A Young H Perspective 2   3   0   11  7   4   0   0   9   6   1   0   2
Kings of the Court  0   3   0   1   4   1   0   0   5   2   0   0   1   0   12  0   3
Original Underground Hip Hop    0   1   0   0   10  0   0   0   1   0   0   0   0   0
OnUrb   0   1   0   0   0   0   0   1   0   1   0   0   0   0   0   0   0   0   0   3
THE WAV 0   2   0   0   0   10  0   0   0   0   0   0   0   0   3   0   0   0   5   0
jessemateus 0   1   3   0   1   1   1   0   0   3   1   3   1   0   0   1   4   0   5
VIVAFIDEL.INFO  0   3   0   1   1   0   0   1   1   0   0   0   0   3   2   2   0   0
.Bavarian - Tendencies  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
We Want In HipHop Blog  0   0   0   0   0   0   0   0   2   0   0   0   0   0   0   2
The ComeUp Mag  0   0   0   2   0   0   0   0   0   0   0   1   0   0   2   0   1   0
Beth Sturgess' Blog 0   0   0   0   0   0   2   0   0   6   0   0   0   0   1   0   1
A Taste of Vinum    0   2   0   9   0   1   2   11  4   3   0   0   8   5   2   13  2
Double Trouble Mixtapes 0   1   0   5   1   3   1   1   4   2   2   3   0   1   3   2
730 DIPS DJS    0   1   0   0   0   0   0   2   0   0   0   1   0   0   1   3   0   0
THI_KN_SS   4   4   1   4   1   0   1   2   9   1   1   0   1   1   6   4   4   2   0
Smugglaztv  0   0   0   4   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
Living In The Stuy...   4   0   2   20  2   0   0   2   4   1   1   0   0   6   3   0
KrazySwagDJs    0   0   0   3   0   1   0   1   5   3   0   0   0   0   1   0   0   0
To The Break Of Dawn    1   0   0   6   6   2   0   1   4   0   1   0   0   0   4   7
WeDooPromo  0   0   0   4   1   3   1   1   2   1   0   4   0   1   3   1   2   0   5
ThE LiFe AnD TiMeS...   0   0   0   4   0   0   2   0   1   0   1   0   0   0   0   0
A K  of Lost One Production's: THE blog 0   2   1   5   5   0   1   0   8   6   1   0
```

Shown above is a part of the data received from compiling ‘generateFeed.py’

2. Create an ASCII and JPEG dendrogram that clusters (i.e., HAC) the most similar blogs (see slides 12 & 13). Include the JPEG in your report and upload the ascii file to github (it will be too unwieldy for inclusion in the report).

Using the file ‘clusters.py’ provided by PCI, I was able to use both the drawdendrogram and printclust functions to generate both a .jpg and ascii file based off of the data that I generated in the first question.

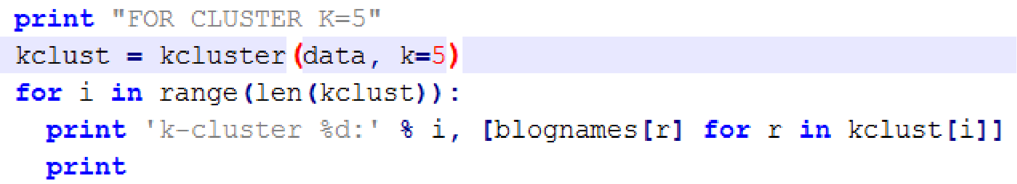
The ascii file is called ‘asciiDiagram.txt’ in github and the .jpg is called ‘blogclust.jpg’ (shown on the next page).


.Bavarian - Tendencies
Lightshapes
Skateboards with Candy
GucciManeClassics.com
Fresh: Hip-Hop & R&B
DJ A_DOG
Ayran Oberto
INCREASE THE GREASE
Kings of the Court
A.K. of Lost One Production's: THE blog
BARBARIANMUSIC
Gang$ta Life
SoulfoodForTheKidz
Feeling the Melody
Crunchy Black Hard Hittaz Money Gang
DC Hat Shop - Sumter, SC
Hood Supastar
730 DIPS DJS
Mississippi HipHop
Krooked Kingdom
LouiVon Official Website
A Half-Warmed Fish
Fubar's Blog
Dispikable Mix Sessions
retarded on a boombox frame
GRND.iNi
Million Blunts
ThE LiFe AnD TiMeS...
Colorful.Thoughts
Star Quality Music
delvbox
Rebelvision Network News
-Dope Music Daily-
We Want In HipHop Blog
FEJO STUDIO TENET - a fresh world of entertainment
G.Y.T Ent
Double Trouble Mixtapes
WeDooPromo
NAPTOWNCONNECTION.COM
Hip-Hop Overdose
THE WAV
The ComeUp Mag
OnUrb
Smugglaztv
e:1
Carly's Captions
Music Download Blog
Web Science and Digital Libraries Research Group
Beth Sturgess' Blog
Jerry Blossom's Finest
A Taste of Vinum
omarbinmusa
Living In The Stuy...
Charged By Soul 0.5 (Under Construction)
T I L . I N F I N I T Y
To The Break Of Dawn
Nikes, Dikes, And Excitebike
MagVit
Doin Just Fine
Black is Bootyful
TheRootMusic
Hip-Hop: From A Young H Perspective
A People's History of Hip Hop
No Days Off Radio
Dollhouse
THI_KN_SS
Hip Hop 2016
WELCOME TO ABJ RICH TEENS BLOG
teameffortsmedia
#TheKendroShow
This Chaotic World
hotep365.comâ¢
Three Dudes Write a Blog
paisley wallpaper
upsetting the set up since 1977
TELEPATHS AT NIGHT
SparkNaija.com | SPARKING ENTERTAINMENT ALL OVER D WORLD
Rad For Jesus
Blessed Indeed
jessemateus
MATH ROCK?
F-Measure
Stitch By Stitch
Original Underground Hip Hop
The Suite Delight
BOND GIRL RAQUEL
Ryan Mules
Fabulous Chamyn
Gaudy Lyfe
dress of the day
TheShredGallery
KrazySwagDJs
LIFE SET TO A BEAT | #LSTAB
SOLO138
VIVAFIDEL.INFO
TRAPCITY
Rubox the Cube's Corner
Dj Ricochet Official Website
CiscoNYC
BEATS OF ART

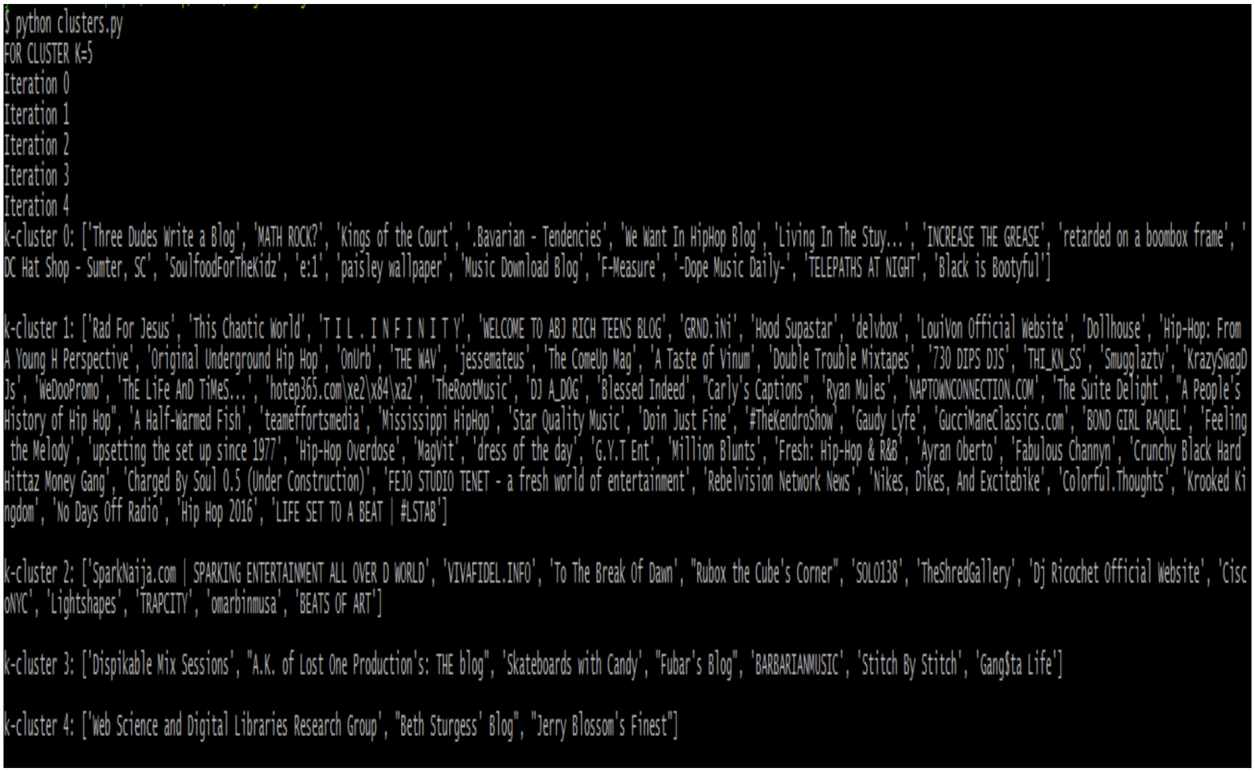
3. Cluster the blogs using K-Means, using k=5,10,20. (see slide 18). Print the values in each centroid, for each value of k. How many interations were required for each value of k?

For this question I used the data received in question 1 and the 'clusters.py' file (from PCI).

For k = 5: Computed in 5 total iterations.



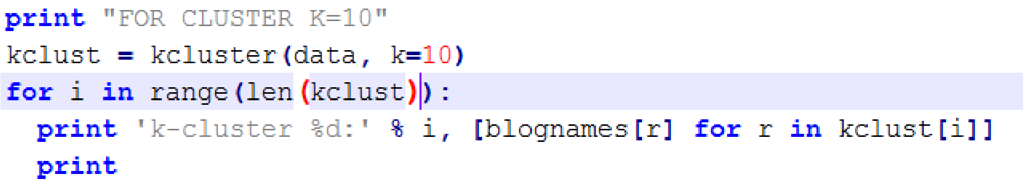
```
print "FOR CLUSTER K=5"
kclust = kcluster(data, k=5)
for i in range(len(kclust)):
  print 'k-cluster %d:' % i, [blognames[r] for r in kclust[i]]
  print
```



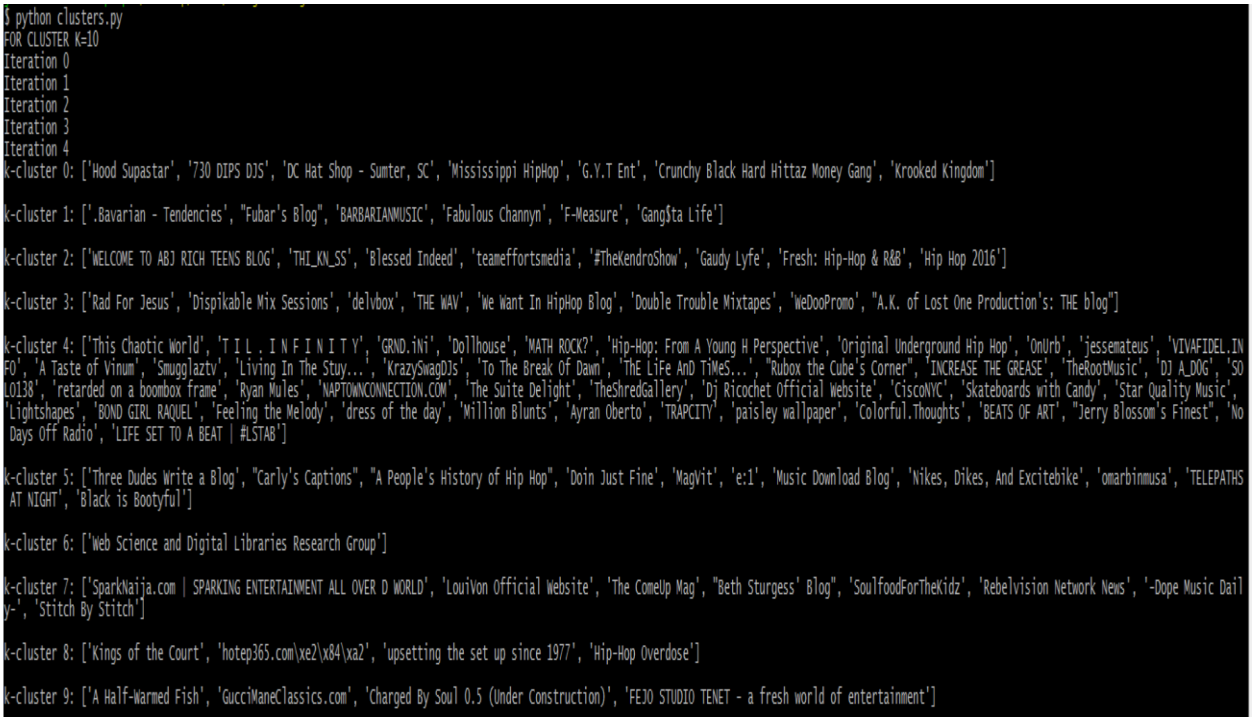
```
$ python clusters.py
FOR CLUSTER K=5
Iteration 0
Iteration 1
Iteration 2
Iteration 3
Iteration 4
k-cluster 0: ['Three Dudes Write a Blog', 'MATH ROCK?', 'Kings of the Court', '.Bavarian - Tendencies', 'We Want In HipHop Blog', 'Living In The Stuy...', 'INCREASE THE GREASE', 'retarded on a boombox frame', '
DC Hat Shop - Sumter, SC', 'SoulfoodForTheKidz', 'e:1', 'paisley wallpaper', 'Music Download Blog', 'F-Measure', '-Dope Music Daily-', 'TELEPATHS AT NIGHT', 'Black is Bootyful']

k-cluster 1: ['Rad For Jesus', 'This Chaotic World', 'T I L . I N F I N I T Y', 'WELCOME TO ABJ RICH TEENS BLOG', 'GRND.iNi', 'Hood Supastar', 'delvbox', 'LouiVon Official Website', 'Dollhouse', 'Hip-Hop: From
A Young H Perspective', 'Original Underground Hip Hop', 'OnUrb', 'THE WAV', 'jessemateus', 'The ComeUp Mag', 'A Taste of Vinum', 'Double Trouble Mixtapes', '730 DIPS DJS', 'THI_KN_SS', 'Smugglaztv', 'KrazySwagD
Js', 'WeDooPromo', 'ThE LiFe AnD TiMeS...', 'hotep365.com\xe2\x84\xa2', 'TheRootMusic', 'DJ A_DOG', 'Blessed Indeed', "Carly's Captions", 'Ryan Mules', 'NAPTOWNCONNECTION.COM', 'The Suite Delight', "A People's
History of Hip Hop", 'A Half-Warmed Fish', 'teameffortsmedia', 'Mississippi HipHop', 'Star Quality Music', 'Doin Just Fine', '#TheKendroShow', 'Gaudy Lyfe', 'GucciManeClassics.com', 'BOND GIRL RAQUEL', 'Feeling
 the Melody', 'upsetting the set up since 1977', 'Hip-Hop Overdose', 'MagVit', 'dress of the day', 'G.Y.T Ent', 'Million Blunts', 'Fresh: Hip-Hop & R&B', 'Ayran Oberto', 'Fabulous Channyn', 'Crunchy Black Hard
Hittaz Money Gang', 'Charged By Soul 0.5 (Under Construction)', 'FEJO STUDIO TENET - a fresh world of entertainment', 'Rebelvision Network News', 'Nikes, Dikes, And Excitebike', 'Colorful.Thoughts', 'Krooked Ki
ngdom', 'No Days Off Radio', 'Hip Hop 2016', 'LIFE SET TO A BEAT | #LSTAB']

k-cluster 2: ['SparkNaija.com | SPARKING ENTERTAINMENT ALL OVER D WORLD', 'VIVAFIDEL.INFO', 'To The Break Of Dawn', "Rubox the Cube's Corner", 'SOLO138', 'TheShredGallery', 'Dj Ricochet Official Website', 'Cisc
oNYC', 'Lightshapes', 'TRAPCITY', 'omarbinmusa', 'BEATS OF ART']

k-cluster 3: ['Dispikable Mix Sessions', "A.K. of Lost One Production's: THE blog", 'Skateboards with Candy', "Fubar's Blog", 'BARBARIANMUSIC', 'Stitch By Stitch', 'Gang$ta Life']

k-cluster 4: ['Web Science and Digital Libraries Research Group', "Beth Sturgess' Blog", "Jerry Blossom's Finest"]
```

Output shown in 'ClusterFive.txt'

For k = 10: Completed in 5 total iterations.



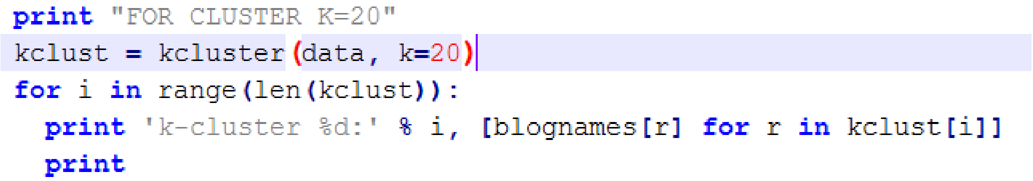
```
print "FOR CLUSTER K=10"
kclust = kcluster(data, k=10)
for i in range(len(kclust)):
  print 'k-cluster %d:' % i, [blognames[r] for r in kclust[i]]
  print
```



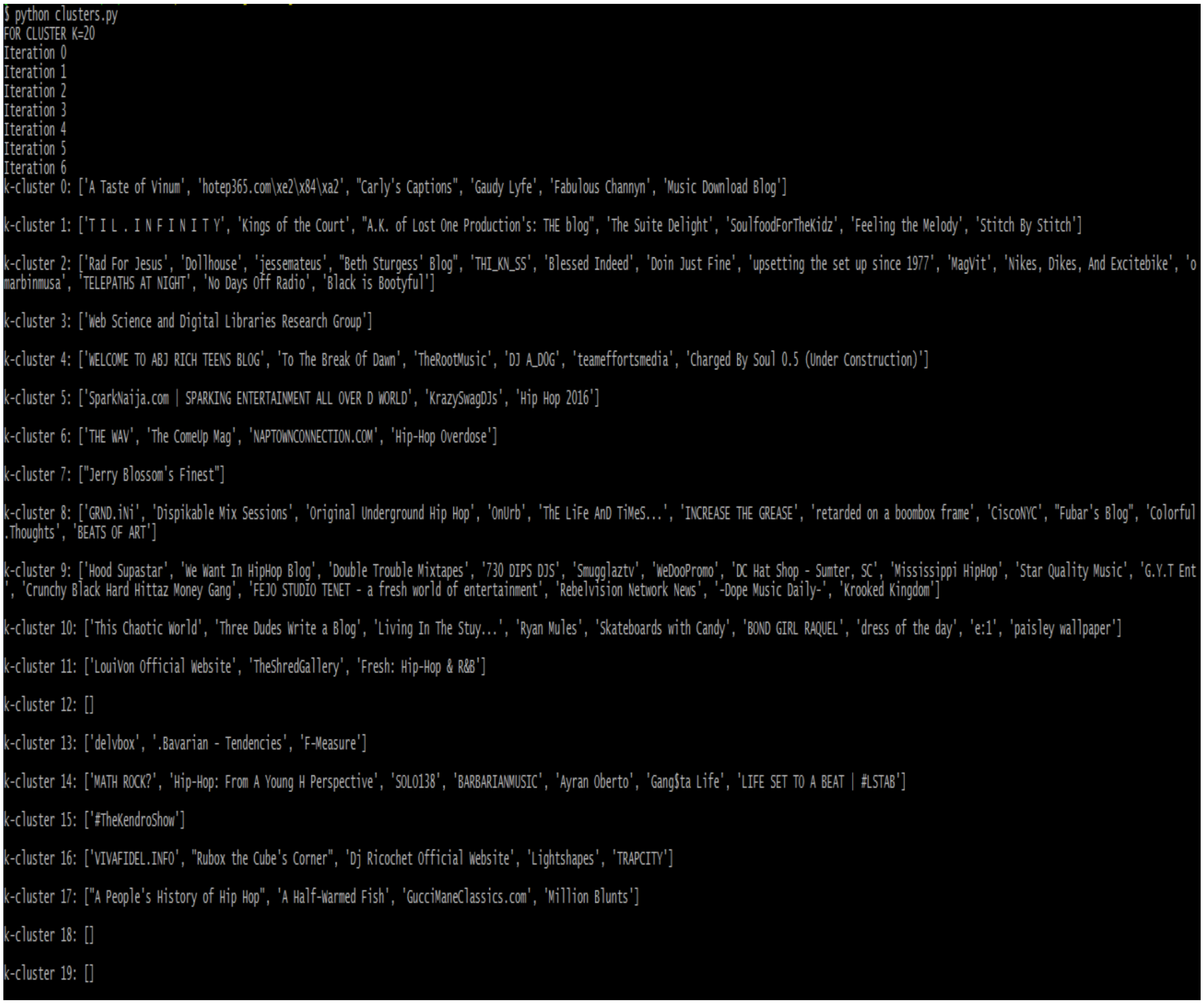
```
$ python clusters.py
FOR CLUSTER K=10
Iteration 0
Iteration 1
Iteration 2
Iteration 3
Iteration 4
k-cluster 0: ['Hood Supastar', '730 DIPS DJS', 'DC Hat Shop - Sumter, SC', 'Mississippi HipHop', 'G.Y.T Ent', 'Crunchy Black Hard Hittaz Money Gang', 'Krooked Kingdom']

k-cluster 1: ['.Bavarian - Tendencies', "Fubar's Blog", 'BARBARIANMUSIC', 'Fabulous Channyn', 'F-Measure', 'Gang$ta Life']

k-cluster 2: ['WELCOME TO ABJ RICH TEENS BLOG', 'THI_KN_SS', 'Blessed Indeed', 'teameffortsmedia', '#TheKendroShow', 'Gaudy Lyfe', 'Fresh: Hip-Hop & R&B', 'Hip Hop 2016']

k-cluster 3: ['Rad For Jesus', 'Dispikable Mix Sessions', 'delvbox', 'THE WAV', 'We Want In HipHop Blog', 'Double Trouble Mixtapes', 'WeDooPromo', "A.K. of Lost One Production's: THE blog"]

k-cluster 4: ['This Chaotic World', 'T I L . I N F I N I T Y', 'GRND.iNi', 'Dollhouse', 'MATH ROCK?', 'Hip-Hop: From A Young H Perspective', 'Original Underground Hip Hop', 'OnUrb', 'jessemateus', 'VIVAFIDEL.IN
FO', 'A Taste of Vinum', 'Smugglaztv', 'Living In The Stuy...', 'KrazySwagDJs', 'To The Break Of Dawn', 'ThE LiFe AnD TiMeS...', "Rubox the Cube's Corner", 'INCREASE THE GREASE', 'TheRootMusic', 'DJ A_DOG', 'SO
LO138', 'retarded on a boombox frame', 'Ryan Mules', 'NAPTOWNCONNECTION.COM', 'The Suite Delight', 'TheShredGallery', 'Dj Ricochet Official Website', 'CiscoNYC', 'Skateboards with Candy', 'Star Quality Music',
'Lightshapes', 'BOND GIRL RAQUEL', 'Feeling the Melody', 'dress of the day', 'Million Blunts', 'Ayran Oberto', 'TRAPCITY', 'paisley wallpaper', 'Colorful.Thoughts', 'BEATS OF ART', "Jerry Blossom's Finest", 'No
 Days Off Radio', 'LIFE SET TO A BEAT | #LSTAB']

k-cluster 5: ['Three Dudes Write a Blog', "Carly's Captions", "A People's History of Hip Hop", 'Doin Just Fine', 'MagVit', 'e:1', 'Music Download Blog', 'Nikes, Dikes, And Excitebike', 'omarbinmusa', 'TELEPATHS
 AT NIGHT', 'Black is Bootyful']

k-cluster 6: ['Web Science and Digital Libraries Research Group']

k-cluster 7: ['SparkNaija.com | SPARKING ENTERTAINMENT ALL OVER D WORLD', 'LouiVon Official Website', 'The ComeUp Mag', "Beth Sturgess' Blog", 'SoulfoodForTheKidz', 'Rebelvision Network News', '-Dope Music Dail
y-', 'Stitch By Stitch']

k-cluster 8: ['Kings of the Court', 'hotep365.com\xe2\x84\xa2', 'upsetting the set up since 1977', 'Hip-Hop Overdose']

k-cluster 9: ['A Half-Warmed Fish', 'GucciManeClassics.com', 'Charged By Soul 0.5 (Under Construction)', 'FEJO STUDIO TENET - a fresh world of entertainment']
```

Output shown in ‘kclustten.txt’

For k = 20: Completed in 7 total iterations.



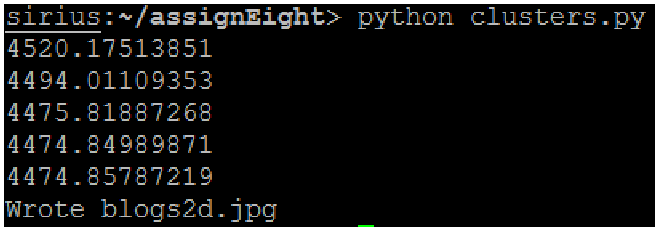
```
print "FOR CLUSTER K=20"
kclust = kcluster(data, k=20)
for i in range(len(kclust)):
  print 'k-cluster %d:' % i, [blognames[r] for r in kclust[i]]
  print
```



```
$ python clusters.py
FOR CLUSTER K=20
Iteration 0
Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 5
Iteration 6
k-cluster 0: ['A Taste of Vinum', 'hotep365.com\xe2\x84\xa2', "Carly's Captions", 'Gaudy Lyfe', 'Fabulous Channyn', 'Music Download Blog']

k-cluster 1: ['T I L . I N F I N I T Y', 'Kings of the Court', "A.K. of Lost One Production's: THE blog", 'The Suite Delight', 'SoulfoodForTheKidz', 'Feeling the Melody', 'Stitch By Stitch']

k-cluster 2: ['Rad For Jesus', 'Dollhouse', 'jessemateus', "Beth Sturgess' Blog", 'THI_KN_SS', 'Blessed Indeed', 'Doin Just Fine', 'upsetting the set up since 1977', 'MagVit', 'Nikes, Dikes, And Excitebike', 'o
marbinmusa', 'TELEPATHS AT NIGHT', 'No Days Off Radio', 'Black is Bootyful']

k-cluster 3: ['Web Science and Digital Libraries Research Group']

k-cluster 4: ['WELCOME TO ABJ RICH TEENS BLOG', 'To The Break Of Dawn', 'TheRootMusic', 'DJ A_DOG', 'teameffortsmedia', 'Charged By Soul 0.5 (Under Construction)']

k-cluster 5: ['SparkNaija.com | SPARKING ENTERTAINMENT ALL OVER D WORLD', 'KrazySwagDJs', 'Hip Hop 2016']

k-cluster 6: ['THE WAV', 'The ComeUp Mag', 'NAPTOWNCONNECTION.COM', 'Hip-Hop Overdose']

k-cluster 7: ["Jerry Blossom's Finest"]

k-cluster 8: ['GRND.iNi', 'Dispikable Mix Sessions', 'Original Underground Hip Hop', 'OnUrb', 'ThE LiFe AnD TiMeS...', 'INCREASE THE GREASE', 'retarded on a boombox frame', 'CiscoNYC', "Fubar's Blog", 'Colorful
.Thoughts', 'BEATS OF ART']

k-cluster 9: ['Hood Supastar', 'We Want In HipHop Blog', 'Double Trouble Mixtapes', '730 DIPS DJS', 'Smugglaztv', 'WeDooPromo', 'DC Hat Shop - Sumter, SC', 'Mississippi HipHop', 'Star Quality Music', 'G.Y.T Ent
', 'Crunchy Black Hard Hittaz Money Gang', 'FEJO STUDIO TENET - a fresh world of entertainment', 'Rebelvision Network News', '-Dope Music Daily-', 'Krooked Kingdom']

k-cluster 10: ['This Chaotic World', 'Three Dudes Write a Blog', 'Living In The Stuy...', 'Ryan Mules', 'Skateboards with Candy', 'BOND GIRL RAQUEL', 'dress of the day', 'e:1', 'paisley wallpaper']

k-cluster 11: ['LouiVon Official Website', 'TheShredGallery', 'Fresh: Hip-Hop & R&B']

k-cluster 12: []

k-cluster 13: ['delvbox', '.Bavarian - Tendencies', 'F-Measure']

k-cluster 14: ['MATH ROCK?', 'Hip-Hop: From A Young H Perspective', 'SOLO138', 'BARBARIANMUSIC', 'Ayran Oberto', 'Gang$ta Life', 'LIFE SET TO A BEAT | #LSTAB']

k-cluster 15: ['#TheKendroShow']

k-cluster 16: ['VIVAFIDEL.INFO', "Rubox the Cube's Corner", 'Dj Ricochet Official Website', 'Lightshapes', 'TRAPCITY']

k-cluster 17: ["A People's History of Hip Hop", 'A Half-Warmed Fish', 'GucciManeClassics.com', 'Million Blunts']

k-cluster 18: []

k-cluster 19: []
```

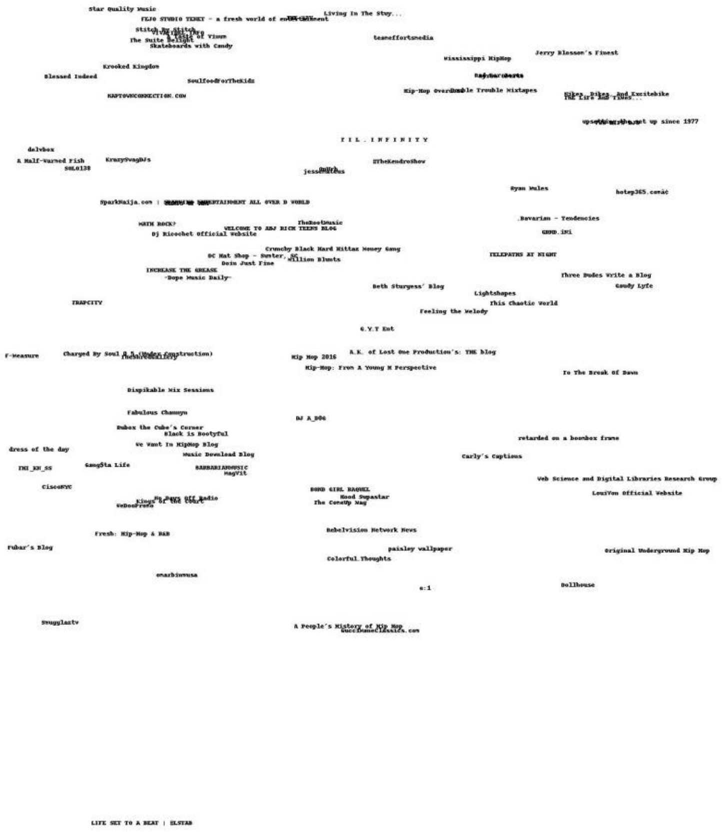
Output shown in 'kclusttwenty.txt'

4. Use MDS to create a JPEG of the blogs similar to slide 29. How many iterations were required?



```
sirius:~/assignEight> python clusters.py
4520.17513851
4494.01109353
4475.81887268
4474.84989871
4474.85787219
Wrote blogs2d.jpg
```

In 'clusters.py' (PCI) I used the 'drawclust.draw2d' function and the data from question 1 to create the 2d graph. Five total iterations were required in this compilation. The picture is called 'blogs2d.jpg.'



References:

https://github.com/shawnmjones/cs595-f13/blob/master/assignment9/q1/fetchFeeds.py

https://github.com/nico/collectiveintelligence-book