CS 432/532: Web Science

Spring 2017

Assignment 8

Michael Micros

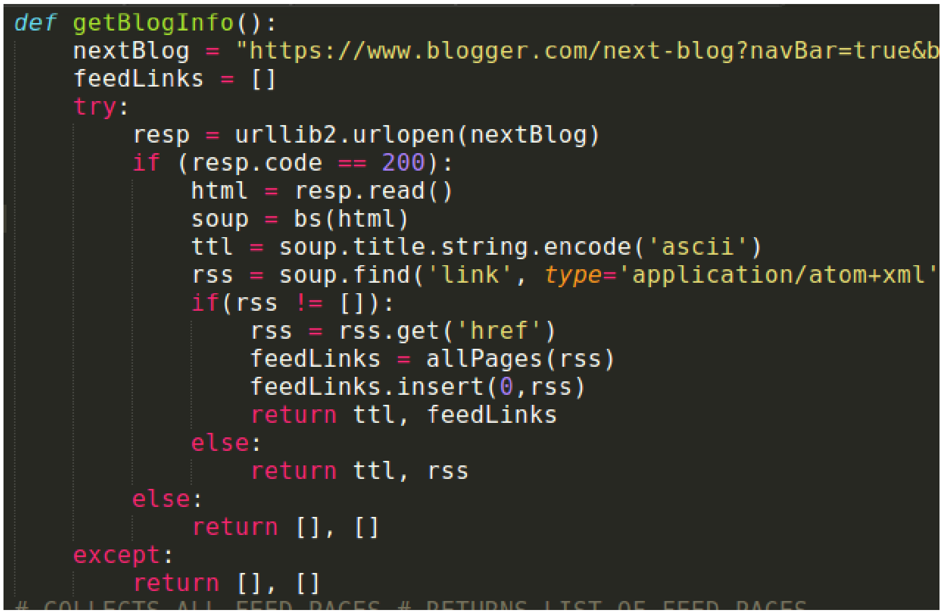
Instructor: Michael L. Nelson

Old Dominon University
Norfolk, Virginia

April 13, 2017

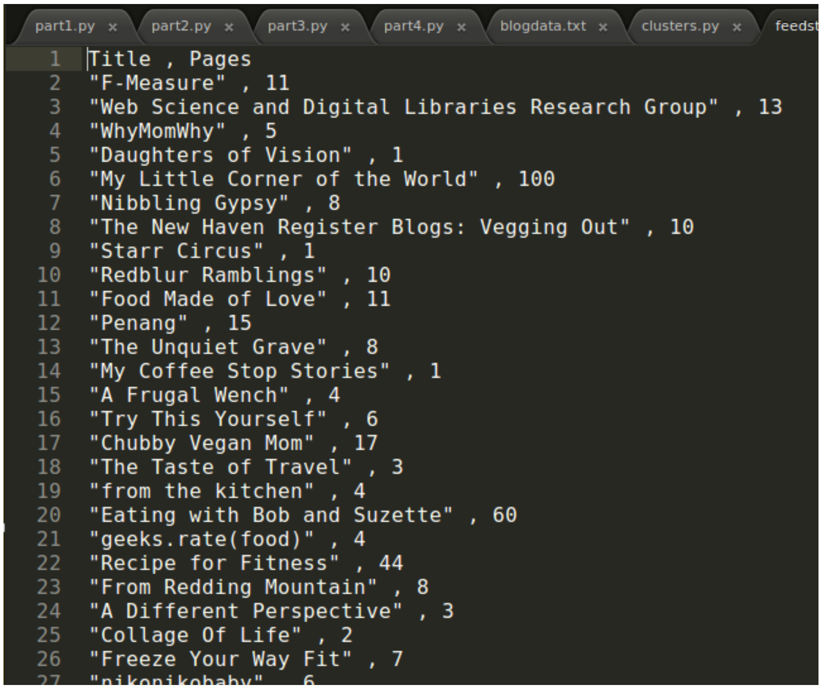
Problem 1: Generating blog-term matrix

In order to generate the blog term matrix that is used throughout this assignment, it was first necessary to get 100 unique blogs. The "part1.py" script accomplishes this by making use of "urllib2" and "BeautifulSoup". The way it works is we make a request for the "next blog" and once we get a "200 OK", we iterate through all of the blog's pages and add them to the "feedlist.txt" which will later be given to the feedparser script. A summary of the blog title and number of pages is saved in "blogstats.txt".



```
def getBlogInfo():
    nextBlog = "https://www.blogger.com/next-blog?navBar=true&b
    feedLinks = []
    try:
        resp = urllib2.urlopen(nextBlog)
        if (resp.code == 200):
            html = resp.read()
            soup = bs(html)
            ttl = soup.title.string.encode('ascii')
            rss = soup.find('link', type='application/atom+xml'
            if(rss != []):
                rss = rss.get('href')
                feedLinks = allPages(rss)
                feedLinks.insert(0,rss)
                return ttl, feedLinks
            else:
                return ttl, rss
        else:
            return [], []
    except:
        return [], []
# COLLECTS ALL FEED PAGES # RETURNS LIST OF FEED PAGES
```

Figure 1: Section of "part1.py"



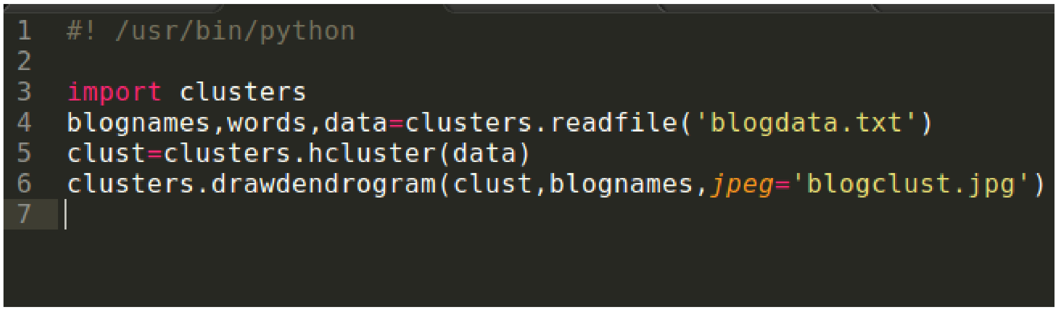
```
Title , Pages
"F-Measure" , 11
"Web Science and Digital Libraries Research Group" , 13
"WhyMomWhy" , 5
"Daughters of Vision" , 1
"My Little Corner of the World" , 100
"Nibbling Gypsy" , 8
"The New Haven Register Blogs: Vegging Out" , 10
"Starr Circus" , 1
"Redblur Ramblings" , 10
"Food Made of Love" , 11
"Penang" , 15
"The Unquiet Grave" , 8
"My Coffee Stop Stories" , 1
"A Frugal Wench" , 4
"Try This Yourself" , 6
"Chubby Vegan Mom" , 17
"The Taste of Travel" , 3
"from the kitchen" , 4
"Eating with Bob and Suzette" , 60
"geeks.rate(food)" , 4
"Recipe for Fitness" , 44
"From Redding Mountain" , 8
"A Different Perspective" , 3
"Collage Of Life" , 2
"Freeze Your Way Fit" , 7
"nikonikobaby" , 6
```

Figure 2: Part of the blog data collected

Once all the blog data is collected, it is fed to the feedparser introduced in Chapter 3 of the “Collective Intelligence Programming”, which created the blog-term matrix where every row represents a blog and every column represents 1 of the 1000 most common words across all blogs.

Problem 2: Creating a dendrogram

The rest of this assignments makes heavy use of the “clusters.py” script and its functions provided in the “Collective Intelligence Programming” textbook.



```
#! /usr/bin/python

import clusters
blognames,words,data=clusters.readfile('blogdata.txt')
clust=clusters.hcluster(data)
clusters.drawdendrogram(clust,blognames,jpeg='blogclust.jpg')

```

Figure 3: “part2.py””

The dendrogram of the collected blogs is displayed on the next page. Obviously it is not easily readable, but the top half of the dendrogram represents two big glusters for lifestyle and food , which are all pretty similar to each other which explains the layout. Additionally, there are smaller clusters toward the bottom having to do with coffee, art, travel, education/programming, etc.

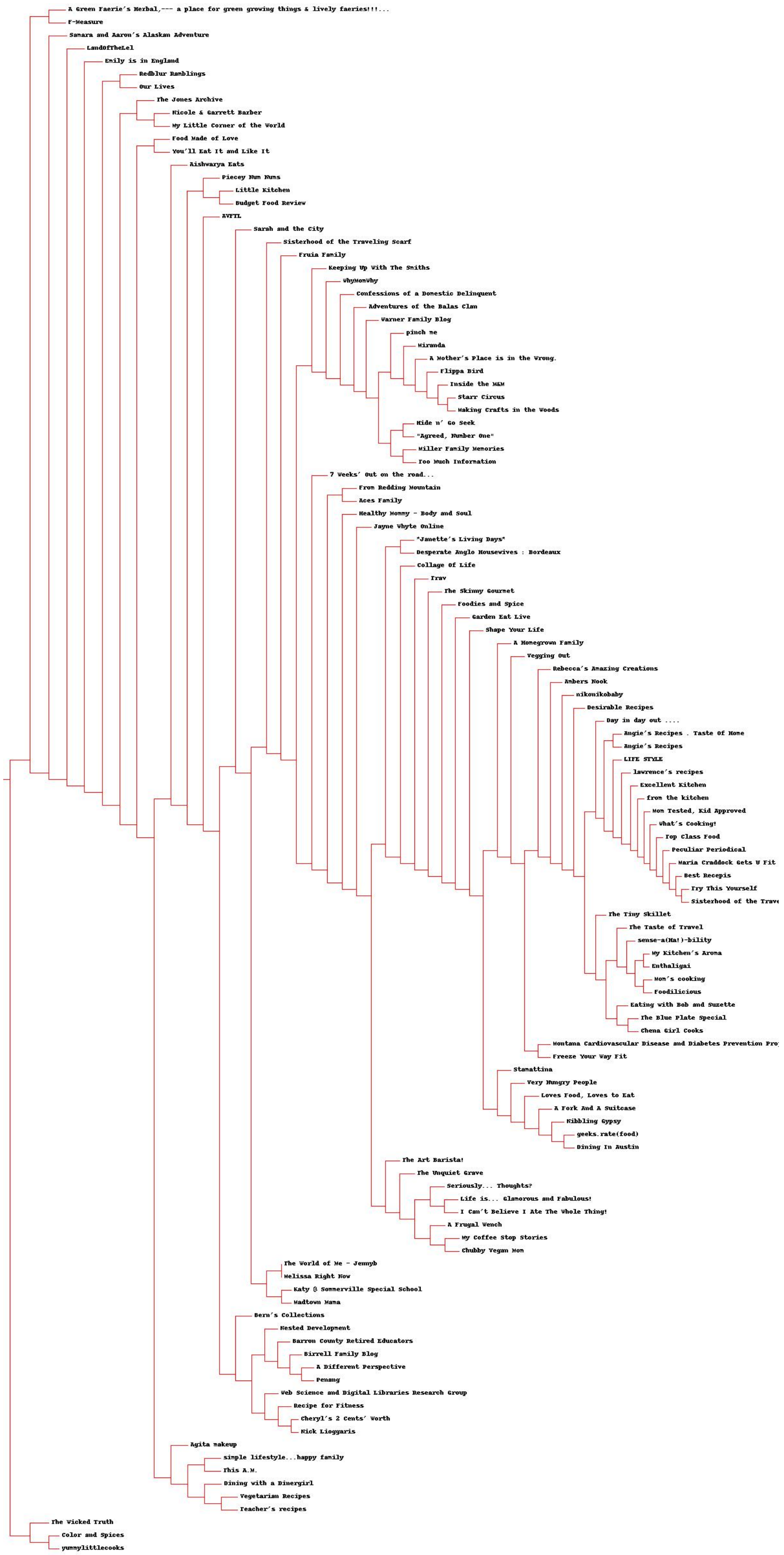
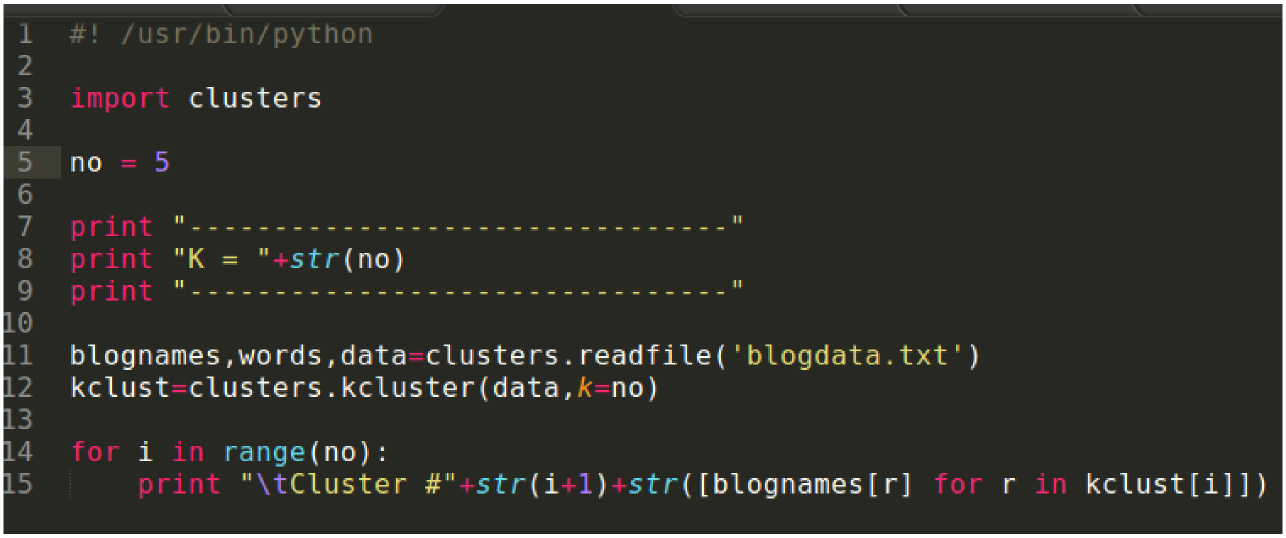


Figure 4: Dendrogram of collected blogs

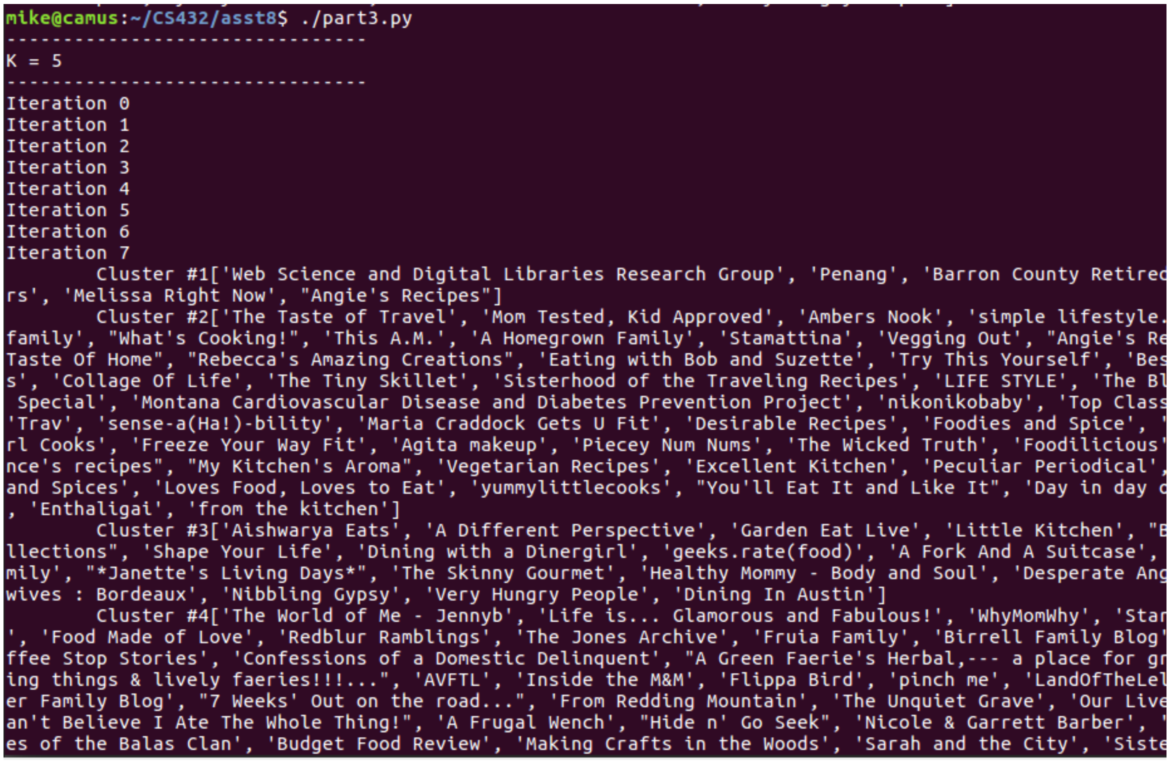
Problem 3: Clustering using K-Means

In order to get the desired results the "kclusters" function in the "clusters.py script" was used 3 times for k=5,10 and 20. The number of iterations needed for 5 clusters was 8. 6 iterations were needed to generate 10 clusters, and 4 iterations to get 20 clusters.



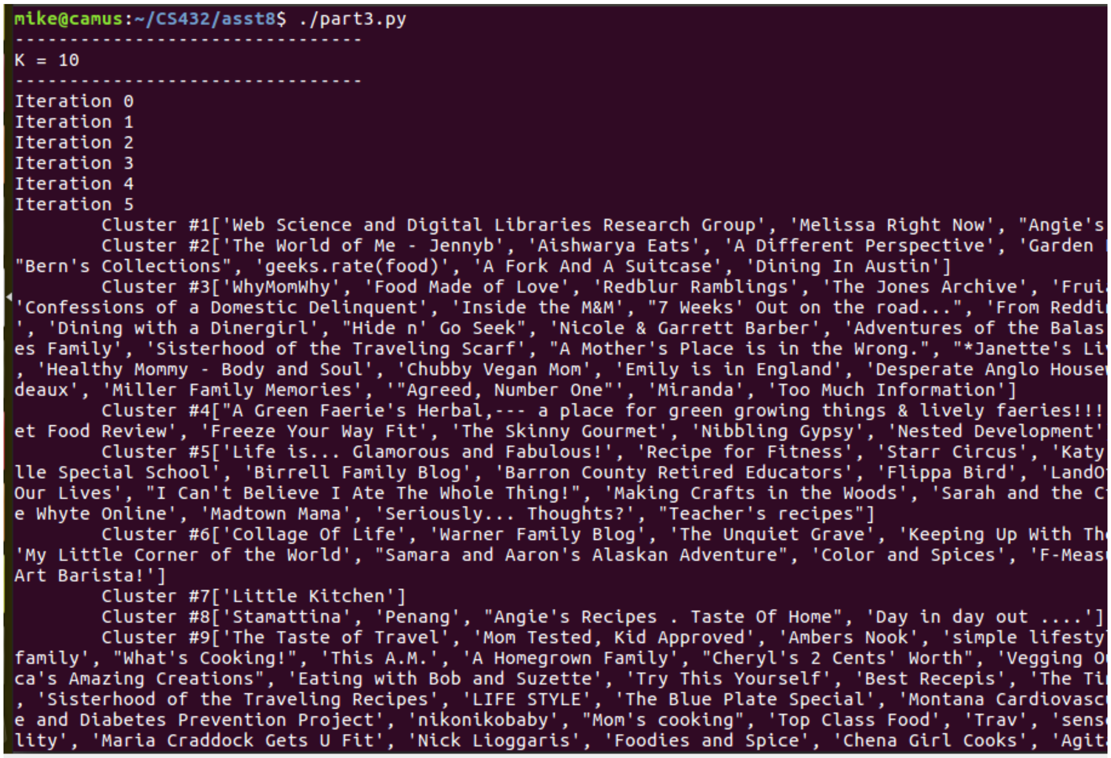
```python
#! /usr/bin/python

import clusters

no = 5

print "------------------------------"
print "K = "+str(no)
print "------------------------------"

blognames,words,data=clusters.readfile('blogdata.txt')
kclust=clusters.kcluster(data,k=no)

for i in range(no):
    print "\tCluster #"+str(i+1)+str([blognames[r] for r in kclust[i]])
```

Figure 5: "part3.py""



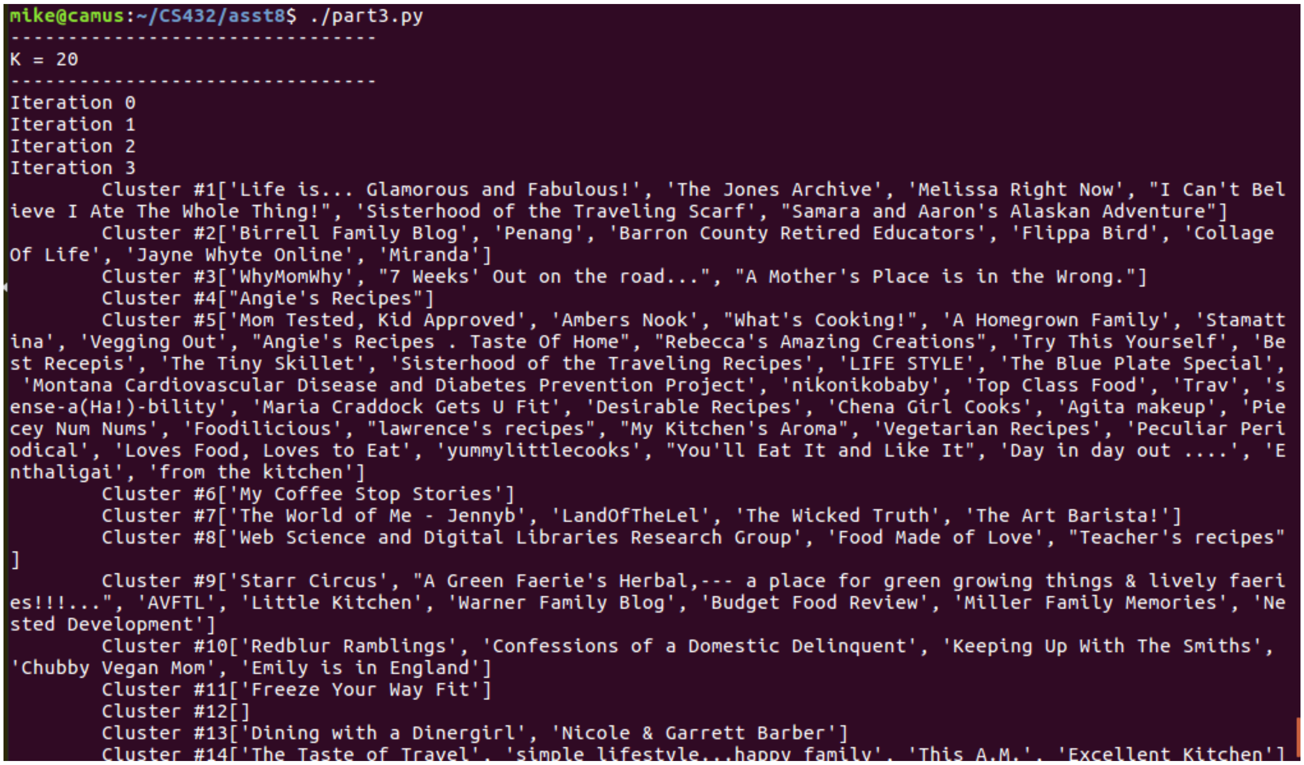
```
mike@camus:~/CS432/asst8$ ./part3.py
------------------------------
K = 5
------------------------------
Iteration 0
Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 5
Iteration 6
Iteration 7
        Cluster #1['Web Science and Digital Libraries Research Group', 'Penang', 'Barron County Retired
rs', 'Melissa Right Now', "Angie's Recipes"]
        Cluster #2['The Taste of Travel', 'Mom Tested, Kid Approved', 'Ambers Nook', 'simple lifestyle.
family', "What's Cooking!", 'This A.M.', 'A Homegrown Family', 'Stamattina', 'Vegging Out', "Angie's Re
Taste Of Home", "Rebecca's Amazing Creations", 'Eating with Bob and Suzette', 'Try This Yourself', 'Bes
s', 'Collage Of Life', 'The Tiny Skillet', 'Sisterhood of the Traveling Recipes', 'LIFE STYLE', 'The Bl
 Special', 'Montana Cardiovascular Disease and Diabetes Prevention Project', 'nikonikobaby', 'Top Class
'Trav', 'sense-a(Ha!)-bility', 'Maria Craddock Gets U Fit', 'Desirable Recipes', 'Foodies and Spice', '
rl Cooks', 'Freeze Your Way Fit', 'Agita makeup', 'Piecey Num Nums', 'The Wicked Truth', 'Foodilicious'
nce's recipes", "My Kitchen's Aroma", 'Vegetarian Recipes', 'Excellent Kitchen', 'Peculiar Periodical',
and Spices', 'Loves Food, Loves to Eat', 'yummylittlecooks', "You'll Eat It and Like It", 'Day in day c
, 'Enthaligai', 'from the kitchen']
        Cluster #3['Aishwarya Eats', 'A Different Perspective', 'Garden Eat Live', 'Little Kitchen', "E
llections", 'Shape Your Life', 'Dining with a Dinergirl', 'geeks.rate(food)', 'A Fork And A Suitcase',
mily', "*Janette's Living Days*", 'The Skinny Gourmet', 'Healthy Mommy - Body and Soul', 'Desperate Ang
wives : Bordeaux', 'Nibbling Gypsy', 'Very Hungry People', 'Dining In Austin']
        Cluster #4['The World of Me - Jennyb', 'Life is... Glamorous and Fabulous!', 'WhyMomWhy', 'Star
', 'Food Made of Love', 'Redblur Ramblings', 'The Jones Archive', 'Fruia Family', 'Birrell Family Blog'
ffee Stop Stories', 'Confessions of a Domestic Delinquent', "A Green Faerie's Herbal,--- a place for gr
ing things & lively faeries!!!...", 'AVFTL', 'Inside the M&M', 'Flippa Bird', 'pinch me', 'LandOfTheLel
er Family Blog', "7 Weeks' Out on the road...", 'From Redding Mountain', 'The Unquiet Grave', 'Our Live
an't Believe I Ate The Whole Thing!", 'A Frugal Wench', "Hide n' Go Seek", 'Nicole & Garrett Barber', '
es of the Balas Clan', 'Budget Food Review', 'Making Crafts in the Woods', 'Sarah and the City', 'Siste
```

Figure 6: Results for K-Means with k=5



```
mike@camus:~/CS432/asst8$ ./part3.py
-------------------------------
K = 10
-------------------------------
Iteration 0
Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 5
        Cluster #1['Web Science and Digital Libraries Research Group', 'Melissa Right Now', "Angie's
        Cluster #2['The World of Me - Jennyb', 'Aishwarya Eats', 'A Different Perspective', 'Garden 
"Bern's Collections", 'geeks.rate(food)', 'A Fork And A Suitcase', 'Dining In Austin']
        Cluster #3['WhyMomWhy', 'Food Made of Love', 'Redblur Ramblings', 'The Jones Archive', 'Frui
'Confessions of a Domestic Delinquent', 'Inside the M&M', "7 Weeks' Out on the road...", 'From Reddi
', 'Dining with a Dinergirl', "Hide n' Go Seek", 'Nicole & Garrett Barber', 'Adventures of the Balas
es Family', 'Sisterhood of the Traveling Scarf', "A Mother's Place is in the Wrong.", "*Janette's Li
, 'Healthy Mommy - Body and Soul', 'Chubby Vegan Mom', 'Emily is in England', 'Desperate Anglo House
deaux', 'Miller Family Memories', '"Agreed, Number One"', 'Miranda', 'Too Much Information']
        Cluster #4["A Green Faerie's Herbal,--- a place for green growing things & lively faeries!!!
et Food Review', 'Freeze Your Way Fit', 'The Skinny Gourmet', 'Nibbling Gypsy', 'Nested Development'
        Cluster #5['Life is... Glamorous and Fabulous!', 'Recipe for Fitness', 'Starr Circus', 'Katy
lle Special School', 'Birrell Family Blog', 'Barron County Retired Educators', 'Flippa Bird', 'LandO
Our Lives', "I Can't Believe I Ate The Whole Thing!", 'Making Crafts in the Woods', 'Sarah and the C
e Whyte Online', 'Madtown Mama', 'Seriously... Thoughts?', "Teacher's recipes"]
        Cluster #6['Collage Of Life', 'Warner Family Blog', 'The Unquiet Grave', 'Keeping Up With Th
'My Little Corner of the World', "Samara and Aaron's Alaskan Adventure", 'Color and Spices', 'F-Meas
Art Barista!']
        Cluster #7['Little Kitchen']
        Cluster #8['Stamattina', 'Penang', "Angie's Recipes . Taste Of Home", 'Day in day out ....']
        Cluster #9['The Taste of Travel', 'Mom Tested, Kid Approved', 'Ambers Nook', 'simple lifesty
family', "What's Cooking!", 'This A.M.', 'A Homegrown Family', "Cheryl's 2 Cents' Worth", 'Vegging O
ca's Amazing Creations", 'Eating with Bob and Suzette', 'Try This Yourself', 'Best Recepis', 'The Ti
, 'Sisterhood of the Traveling Recipes', 'LIFE STYLE', 'The Blue Plate Special', 'Montana Cardiovasc
e and Diabetes Prevention Project', 'nikonikobaby', "Mom's cooking", 'Top Class Food', 'Trav', 'sens
lity', 'Maria Craddock Gets U Fit', 'Nick Lioggaris', 'Foodies and Spice', 'Chena Girl Cooks', 'Agit
```

Figure 7: Results for K-Means with k=10

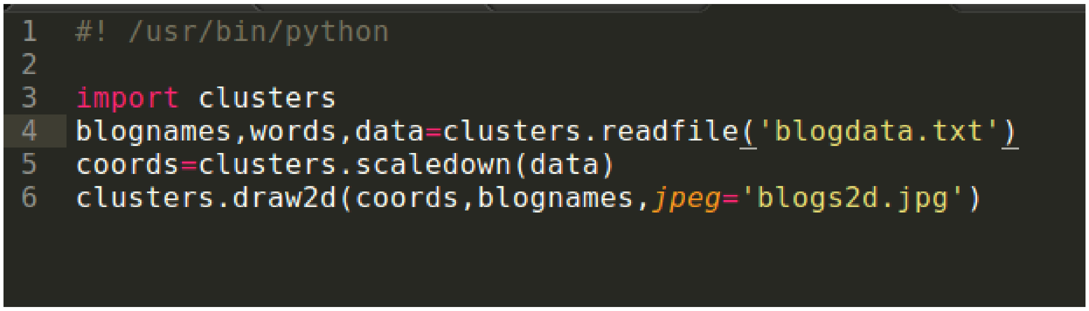


```
mike@camus:~/CS432/asst8$ ./part3.py
-------------------------------
K = 20
-------------------------------
Iteration 0
Iteration 1
Iteration 2
Iteration 3
        Cluster #1['Life is... Glamorous and Fabulous!', 'The Jones Archive', 'Melissa Right Now', "I Can't Bel
ieve I Ate The Whole Thing!", 'Sisterhood of the Traveling Scarf', "Samara and Aaron's Alaskan Adventure"]
        Cluster #2['Birrell Family Blog', 'Penang', 'Barron County Retired Educators', 'Flippa Bird', 'Collage
Of Life', 'Jayne Whyte Online', 'Miranda']
        Cluster #3['WhyMomWhy', "7 Weeks' Out on the road...", "A Mother's Place is in the Wrong."]
        Cluster #4["Angie's Recipes"]
        Cluster #5['Mom Tested, Kid Approved', 'Ambers Nook', "What's Cooking!", 'A Homegrown Family', 'Stamatt
ina', 'Vegging Out', "Angie's Recipes . Taste Of Home", "Rebecca's Amazing Creations", 'Try This Yourself', 'Be
st Recepis', 'The Tiny Skillet', 'Sisterhood of the Traveling Recipes', 'LIFE STYLE', 'The Blue Plate Special',
 'Montana Cardiovascular Disease and Diabetes Prevention Project', 'nikonikobaby', 'Top Class Food', 'Trav', 's
ense-a(Ha!)-bility', 'Maria Craddock Gets U Fit', 'Desirable Recipes', 'Chena Girl Cooks', 'Agita makeup', 'Pie
cey Num Nums', 'Foodilicious', "lawrence's recipes", "My Kitchen's Aroma", 'Vegetarian Recipes', 'Peculiar Peri
odical', 'Loves Food, Loves to Eat', 'yummylittlecooks', "You'll Eat It and Like It", 'Day in day out ....', 'E
nthaligai', 'from the kitchen']
        Cluster #6['My Coffee Stop Stories']
        Cluster #7['The World of Me - Jennyb', 'LandOfTheLel', 'The Wicked Truth', 'The Art Barista!']
        Cluster #8['Web Science and Digital Libraries Research Group', 'Food Made of Love', "Teacher's recipes"
]
        Cluster #9['Starr Circus', "A Green Faerie's Herbal,--- a place for green growing things & lively faeri
es!!!...", 'AVFTL', 'Little Kitchen', 'Warner Family Blog', 'Budget Food Review', 'Miller Family Memories', 'Ne
sted Development']
        Cluster #10['Redblur Ramblings', 'Confessions of a Domestic Delinquent', 'Keeping Up With The Smiths',
'Chubby Vegan Mom', 'Emily is in England']
        Cluster #11['Freeze Your Way Fit']
        Cluster #12[]
        Cluster #13['Dining with a Dinergirl', 'Nicole & Garrett Barber']
        Cluster #14['The Taste of Travel', 'simple lifestyle...happy family', 'This A.M.', 'Excellent Kitchen']
```

Figure 8: Results for K-Means with k=20

Problem 4: 2D display using Multidimentional Scaling

In order to perform multidimensional scaling the “scaledown” and “draw2d” functions were used from the “clusters.py script”, as is seen below.



```
#! /usr/bin/python

import clusters
blognames,words,data=clusters.readfile('blogdata.txt')
coords=clusters.scaledown(data)
clusters.draw2d(coords,blognames,jpeg='blogs2d.jpg')
```

Figure 9: “part4.py””

Again the results are not easily visible but the graph upon analysis has 2 easily distinguishable groups. The blogs in the bottom left correspond to the “lifestyle” blogs whereas the blogs on the top half and also the right mainly are “food” blogs. In total, 136 iterations were required going from an average total difference of 6733.3 (potatos?) to 4257.7.

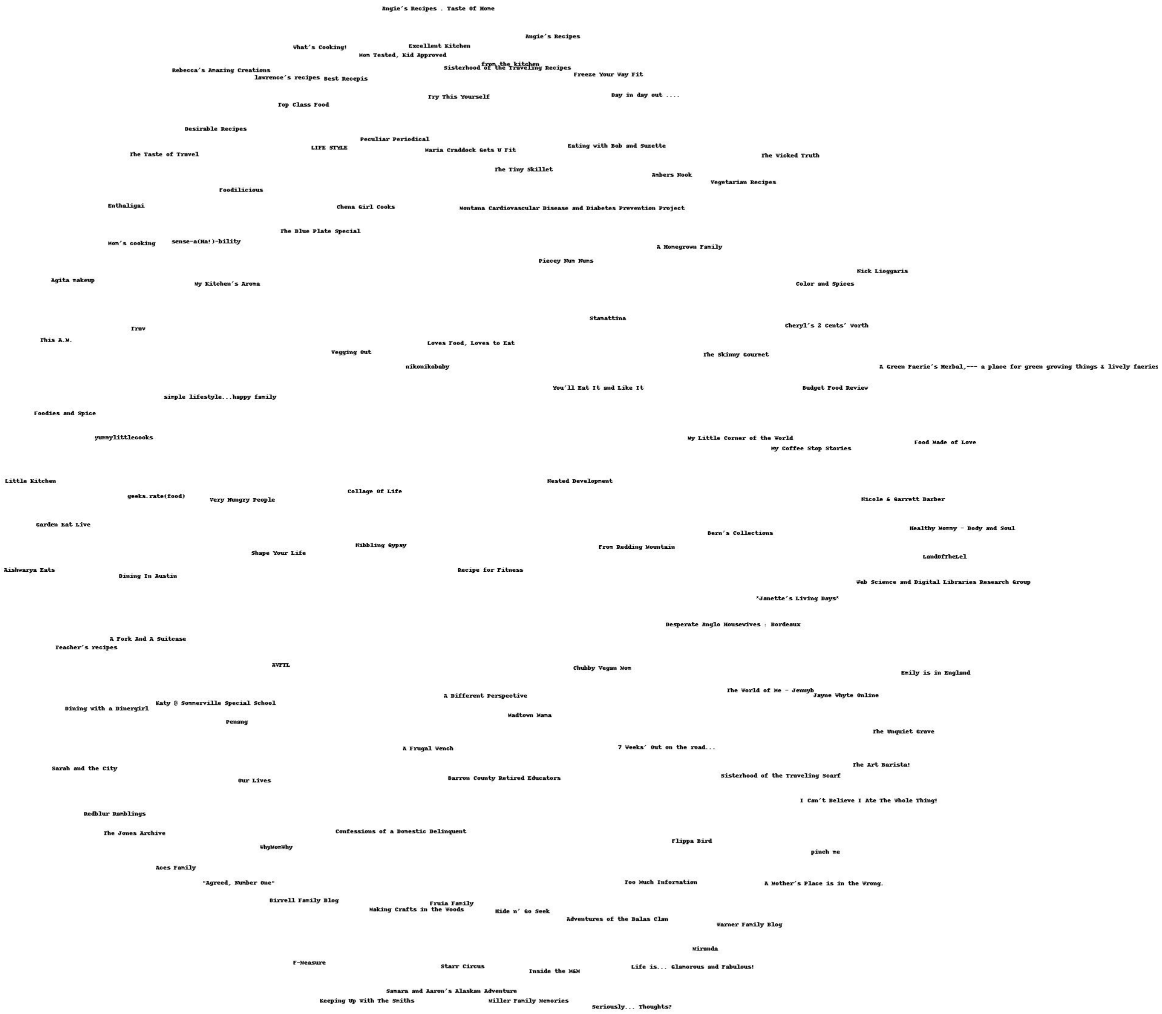


Figure 10: 2D display of our blog data