CS532S19: Assignment #7

Due on Sunday, April 07, 2019

Nwala, Alexander C

Giridharan Ganeshkumar

Question 1

1. Create a blog-term matrix. Start by grabbing 100 blogs hosted on https://www.blogger.com. Include: http://f-measure.blogspot.com/ and http://ws-dl.blogspot.com/

1. The first step is to get the list of blogs from blogspot.com. The google search library is used for this purpose.
2. A query with site:blogspot.com is made. This restricts the search results to bolgspot.com
3. The results are added to the uniqueUrls set but checked to be unique using the canonicalizeURI function
4. The results are written add stored in the UriListBlogspot file

Listing 1: Python Script

```
from googlesearch import search
import csv
from urllib.parse import urlparse
import hashlib

def canonicalizeURI(uri):
    uri = uri.strip()
    if( len(uri) == 0 ):
        return ''
    exceptionDomains = ['www.youtube.com']
    try:
        scheme, netloc, path, params, query, fragment = urlparse( uri )
        netloc = netloc.strip()
        path = path.strip()
        optionalQuery = ''

        if( len(path) != 0 ):
            if( path[-1] != '/' ):
                path = path + '/'

        if( netloc in exceptionDomains ):
            optionalQuery = query.strip()

        return netloc
    except:
        print('Error uri:', uri)

    return ''


query = "site:blogspot.com  technology"
uniqueUrls = set()
with open("C:\\UriListBlogspot.csv", "w", newline='') as csvFile:
    writer = csv.writer(csvFile)
    currentData = ['http://f-measure.blogspot.com/']
    writer.writerow(currentData)
    currentData = ['http://ws-dl.blogspot.com/']
    writer.writerow(currentData)
    for j in search(query, tld="co.in", num=100, stop=400, pause=2):
```

Question 1 continued on next page...

```
        currentCanonicalizedURI = canonicalizeURI(j)
        if currentCanonicalizedURI not in uniqueUrls:
            currentData = [j]
            writer.writerow(currentData)
            uniqueUrls.add(currentCanonicalizedURI)
            print(j)
```

1. The findfeed method is utilized to get the RSS feed urls
2. The next step is to utilize the getwordcounts for every blog feed found.
3. The method is used to get the title of the blog and the word count
4. Then for each word in the wordlist for each of the blogs the method forms BlogData matrix and is written to the BlogData text file.
5. This is the matrix data that will be used for the rest of the calculation in the next three problems. This BlogData text file is uploaded to git hub for reference.

Listing 2: Python Script

```
import feedparser
import re
import urllib.parse
import requests
from bs4 import BeautifulSoup as bs4

def getwordcounts(url):
    '''
    Returns title and dictionary of word counts for an RSS feed
    '''
    # Parse the feed
    d = feedparser.parse(url)
    wc = {}
    # Loop over all the entries
    for e in d.entries:
        if 'summary' in e:
            summary = e.summary
        else:
            summary = e.description
        # Extract a list of words
        words = getwords(e.title + ' ' + summary)
        for word in words:
            wc.setdefault(word, 0)
            wc[word] += 1
    return (d.feed.title , wc)


def getwords(html):
    # Remove all the HTML tags
    txt = re.compile(r'<[^>]+>').sub('', html)

    # Split words by all non-alpha characters
    words = re.compile(r'[^A-Z^a-z]+').split(txt)

    # Convert to lowercase
```

Question 1 continued on next page...

```
    return [word.lower() for word in words if word != '']

def findfeed(site):
    raw = requests.get(site).text
    result = []
    possible_feeds = []
    html = bs4(raw)
    feed_urls = html.findAll("link", rel="alternate")
    for f in feed_urls:
        t = f.get("type",None)
        if t:
            if "rss" in t or "xml" in t:
                href = f.get("href",None)
                if href:
                    possible_feeds.append(href)
    parsed_url = urllib.parse.urlparse(site)
    base = "parsed.scheme"+"://"+parsed_url.hostname
    atags = html.findAll("a")
    for a in atags:
        href = a.get("href",None)
        if href:
            if "xml" in href or "rss" in href or "feed" in href:
                possible_feeds.append(base+href)
    for url in list(set(possible_feeds)):
        f = feedparser.parse(url)
        if len(f.entries) > 0:
            if url not in result:
                result.append(url)
    return(result)

apcount = {}
wordcounts = {}
feedlist = [line for line in open("C:\\UriListBlogspot.txt")]
#feedlistLen = 1
for feedurl in feedlist:
    try:
        rssFeedUrl = findfeed(feedurl)
        (title , wc) = getwordcounts(rssFeedUrl[0])
        #(title , wc) = getwordcounts(feedurl)
        wordcounts[title] = wc
        for (word, count) in wc.items():
            apcount.setdefault(word, 0)
            if count > 1:
                apcount[word] += 1
    except:
        print('Failed to parse feed %s' % feedurl)

wordlist = []
for (w, bc) in apcount.items():
    frac = float(bc) / len(feedlist)
    #frac = float(bc) / feedlistLen
    if frac > 0.1 and frac < 0.5:
        wordlist.append(w)
out = open("C:\\BlogData.txt", 'w')
out.write('Blog')
for word in wordlist:
```

Question 1 continued on next page...

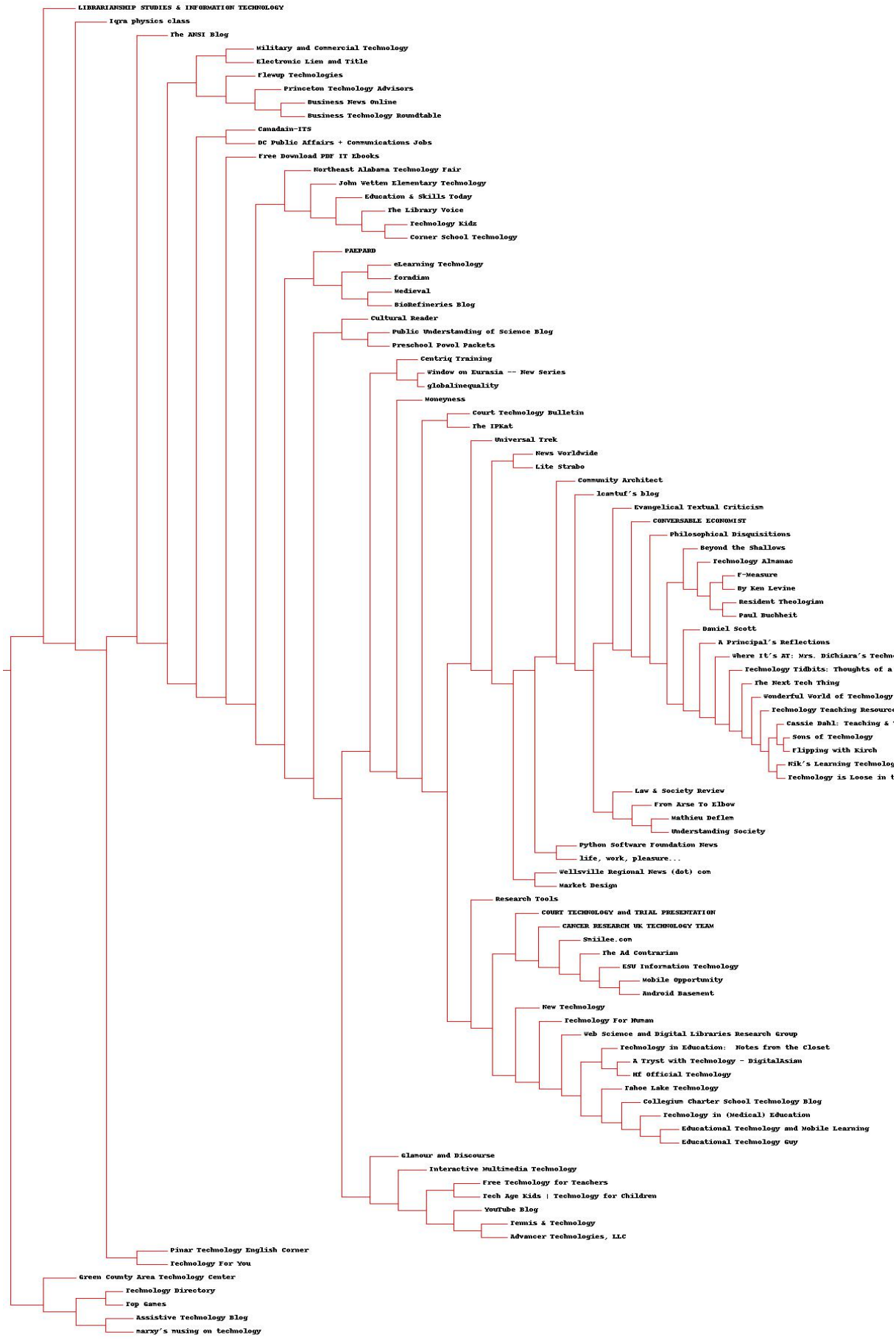
```
92      out.write('\t%s' % word)
93  out.write('\n')
94  for (blog, wc) in wordcounts.items():
95      print(blog)
96      out.write(blog)
97      for word in wordlist:
98          if word in wc:
99              out.write('\t%d' % wc[word])
100         else:
101             out.write('\t0')
102     out.write('\n')
103 out.close()
```

Question 2

2. Create an ASCII and JPEG dendrogram that clusters the most similar blogs.

1. In order to form the cluster the hcluster method is called by passing the matrix formed in the previous question
2. Followed by the hcluster method we call the drawdendrogram method by passing the blog names and the clusters identified in the previous step
3. The below figure represents the dendrogram of clusters of blogs that are similar.

Question 2 continued on next page...



Question 2 continued on next page...

Question 3

3. Cluster the blogs using K-Means, using k is equal to 5,10,20. Print the values in each centroid, for each value of k. How many iterations were required for each value of k?

1. In order to provide a Cluster the blogs using K-Mean and using k is equal to 5 the kcluster method is used
2. To start, the data received in question 1 is utilized and passed to the kcluster method
3. The second parameter used is the k=5
4. It took 5 iterations to for k value being 5
5. The values in each centroid is printed and shown in the below figure

Listing 3: K is 5

```
Iteration 0
Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 5
['News Worldwide', 'Mathieu Deflem', 'From Arse To Elbow', 'Technology
    For You', 'Law & Society Review', 'Python Software Foundation News
   ', 'Understanding Society', 'Northeast Alabama Technology Fair', '
   Window on Eurasia -- New Series', 'Wellsville Regional News (dot)
   com', 'CONVERSABLE ECONOMIST', 'Universal Trek', 'BioRefineries
   Blog', 'life , work, pleasure...', "lcamtuf's blog", 'Evangelical
   Textual Criticism', 'Cultural Reader', 'globalinequality', 'Lite
   Strabo']
======================================================================
['A Tryst with Technology - DigitalAsian', 'Hf Official Technology', '
   Medieval', 'Canadain-ITS', 'eLearning Technology', 'foradian', '
   Free Download PDF IT Ebooks']
======================================================================
['F-Measure', 'Web Science and Digital Libraries Research Group', '
   Educational Technology and Mobile Learning', 'Smiilee.com', 'Free
   Technology for Teachers', 'Technology Teaching Resources with
   Brittany Washburn', 'Tahoe Lake Technology', 'The Next Tech Thing',
    'Cassie Dahl: Teaching & Technology', 'Sons of Technology', '
   Wonderful World of Technology', "Nik's Learning Technology Blog", '
   Technology Tidbits: Thoughts of a Cyber Hero', 'Technology Kidz', '
   Daniel Scott', 'Assistive Technology Blog', 'Centriq Training', '
   Technology is Loose in the Library                    &
   Around the School!!', 'Mobile Opportunity', 'COURT TECHNOLOGY and
   TRIAL PRESENTATION', 'Collegium Charter School Technology Blog', '
   Interactive Multimedia Technology', 'Advancer Technologies, LLC', '
   Educational Technology Guy', 'CANCER RESEARCH UK TECHNOLOGY TEAM',
   'Technology in (Medical) Education', 'Technology Almanac', '
   Resident Theologian', 'John Wetten Elementary Technology', "Where
   It's AT: Mrs. DiChiara's Technology Blog", 'Technology in Education
   :  Notes from the Closet', 'Android Basement', "marxy's musing on
   technology", 'The Library Voice', 'ESU Information Technology', '
   Flipping with Kirch', 'Paul Buchheit', "A Principal's Reflections",
```

Question 3 continued on next page...

```
    'Philosophical Disquisitions', 'The Ad Contrarian', 'By Ken Levine
    ', 'Glamour and Discourse', 'Beyond the Shallows', 'YouTube Blog']
================================================================
['Business News Online', 'Court Technology Bulletin', 'Corner School
    Technology', 'Tech Age Kids | Technology for Children', 'Public
    Understanding of Science Blog', 'Technology Directory', 'Top Games'
    , 'The ANSI Blog', 'Preschool Powol Packets']
================================================================
['Business Technology Roundtable', 'LIBRARIANSHIP STUDIES &
    INFORMATION TECHNOLOGY', 'Pinar Technology English Corner', 'Tennis
     & Technology', 'Military and Commercial Technology', 'Princeton
    Technology Advisors', 'Technology For Human', 'Education & Skills
    Today', 'Green County Area Technology Center', 'DC Public Affairs +
     Communications Jobs', 'Research Tools', 'Electronic Lien and Title
    ', 'PAEPARD', 'Flewup Technologies', 'The IPKat', 'Community
    Architect', 'Iqra physics class', 'Moneyness', 'Market Design', '
    New Technology']
================================================================
```

Listing 4: K is 10

```
Iteration 0
Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 5
Iteration 6
['LIBRARIANSHIP STUDIES & INFORMATION TECHNOLOGY', 'From Arse To Elbow
    ', 'Android Basement']
================================================================
['Python Software Foundation News', 'Education & Skills Today', '
    Public Understanding of Science Blog', 'DC Public Affairs +
    Communications Jobs', 'PAEPARD', 'life , work, pleasure...', '
    YouTube Blog']
================================================================
['F-Measure', 'News Worldwide', 'Smiilee.com', 'Mobile Opportunity', '
    Tech Age Kids | Technology for Children', 'Canadain-ITS', '
    Technology Almanac', 'Resident Theologian', "marxy's musing on
    technology", 'Paul Buchheit', 'Technology Directory', 'Top Games',
    'The Ad Contrarian', 'By Ken Levine', 'Glamour and Discourse', '
    Beyond the Shallows', 'Evangelical Textual Criticism']
================================================================
['Pinar Technology English Corner', 'A Tryst with Technology -
    DigitalAsian', 'Hf Official Technology', 'Medieval', 'BioRefineries
     Blog']
================================================================
['eLearning Technology', 'foradian']
================================================================
['Business News Online', 'Mathieu Deflem', 'Court Technology Bulletin'
    , 'Business Technology Roundtable', 'Military and Commercial
    Technology', 'Law & Society Review', 'Understanding Society', '
    Window on Eurasia -- New Series', 'Philosophical Disquisitions', '
    Wellsville Regional News (dot) com', 'The ANSI Blog', 'CONVERSABLE
    ECONOMIST', 'Universal Trek', 'The IPKat', 'Community Architect', '
```

Question 3 continued on next page...

```
    Iqra physics class', 'Moneyness', 'Market Design', "Icamtuf's blog"
    , 'globalinequality', 'Lite Strabo']
==================================================================
['Web Science and Digital Libraries Research Group', 'Educational
    Technology and Mobile Learning', 'Free Technology for Teachers', '
    Technology Teaching Resources with Brittany Washburn', 'Tahoe Lake
    Technology', 'The Next Tech Thing', 'Cassie Dahl: Teaching &
    Technology', 'Sons of Technology', 'Wonderful World of Technology',
     "Nik's Learning Technology Blog", 'Technology Tidbits: Thoughts of
     a Cyber Hero', 'Technology Kidz', 'Corner School Technology', '
    Technology is Loose in the Library                              &
    Around the School!!', 'Collegium Charter School Technology Blog', '
    Interactive Multimedia Technology', 'Educational Technology Guy', '
    Technology in (Medical) Education', 'Northeast Alabama Technology
    Fair', 'John Wetten Elementary Technology', 'Research Tools', "
    Where It's AT: Mrs. DiChiara's Technology Blog", 'Technology in
    Education:  Notes from the Closet', 'The Library Voice', 'Flipping
    with Kirch', "A Principal's Reflections", 'Preschool Powol Packets'
    ]
==================================================================
['Technology For You', 'Daniel Scott', 'Assistive Technology Blog', '
    CANCER RESEARCH UK TECHNOLOGY TEAM', 'Free Download PDF IT Ebooks',
     'Cultural Reader']
==================================================================
['Tennis & Technology', 'Princeton Technology Advisors', 'Centriq
    Training', 'COURT TECHNOLOGY and TRIAL PRESENTATION', 'Technology
    For Human', 'Advancer Technologies, LLC', 'Green County Area
    Technology Center', 'ESU Information Technology', 'New Technology']
==================================================================
['Electronic Lien and Title', 'Flewup Technologies']
==================================================================
```

Listing 5: K is 20

```
Iteration 0
Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 5
Iteration 6
Iteration 7
['A Tryst with Technology – DigitalAsian', 'Hf Official Technology', '
    Medieval', 'BioRefineries Blog']
==================================================================
['Educational Technology and Mobile Learning', 'Free Technology for
    Teachers', 'Technology Teaching Resources with Brittany Washburn',
    'Tahoe Lake Technology', 'The Next Tech Thing', 'Cassie Dahl:
    Teaching & Technology', 'Sons of Technology', 'Wonderful World of
    Technology', "Nik's Learning Technology Blog", 'Technology Tidbits:
     Thoughts of a Cyber Hero', 'Technology Kidz', 'Corner School
    Technology', 'Assistive Technology Blog', 'Technology is Loose in
    the Library                              & Around the School!!', '
    Collegium Charter School Technology Blog', 'Interactive Multimedia
    Technology', 'Educational Technology Guy', 'Technology in (Medical)
```

Question 3 continued on next page...

```
    Education', 'Northeast Alabama Technology Fair', 'John Wetten
    Elementary Technology', 'Green County Area Technology Center', "
    Where It's AT: Mrs. DiChiara's Technology Blog", 'Technology in
    Education:  Notes from the Closet', 'The Library Voice', 'Flipping
    with Kirch', "A Principal's Reflections"]
================================================================
['Canadain-ITS', 'foradian']
================================================================
['News Worldwide', 'Mathieu Deflem', 'From Arse To Elbow', 'Law &
    Society Review', 'Centriq Training', 'Understanding Society', '
    Window on Eurasia -- New Series', 'DC Public Affairs +
    Communications Jobs', 'Electronic Lien and Title', 'Wellsville
    Regional News (dot) com', 'The IPKat', 'Cultural Reader', '
    globalinequality', 'Lite Strabo']
================================================================
['Web Science and Digital Libraries Research Group', 'Business News
    Online', 'Business Technology Roundtable', 'Military and Commercial
     Technology', 'Princeton Technology Advisors', 'Research Tools', '
    PAEPARD', 'Flewup Technologies', 'life , work, pleasure...', '
    YouTube Blog']
================================================================
['Technology Directory', 'Top Games']
================================================================
['Python Software Foundation News']
================================================================
[]
================================================================
['Glamour and Discourse']
================================================================
['Tennis & Technology', 'COURT TECHNOLOGY and TRIAL PRESENTATION']
================================================================
['Court Technology Bulletin', 'Advancer Technologies, LLC', 'Free
    Download PDF IT Ebooks']
================================================================
['Education & Skills Today', 'The ANSI Blog', 'Universal Trek', '
    Community Architect']
================================================================
[]
================================================================
[]
================================================================
['Pinar Technology English Corner', 'Technology For You']
================================================================
['eLearning Technology']
================================================================
['LIBRARIANSHIP STUDIES & INFORMATION TECHNOLOGY', 'Tech Age Kids |
    Technology for Children', 'Public Understanding of Science Blog', '
    Preschool Powol Packets']
================================================================
[]
================================================================
['F-Measure', 'Smiilee.com', 'Daniel Scott', 'Mobile Opportunity', '
    Technology For Human', 'CANCER RESEARCH UK TECHNOLOGY TEAM', '
    Technology Almanac', 'Resident Theologian', 'Android Basement', "
```

Question 3 continued on next page...

```
    marxy's musing on technology", 'ESU Information Technology', 'Paul
    Buchheit', 'Philosophical Disquisitions', 'CONVERSABLE ECONOMIST',
    'The Ad Contrarian', 'By Ken Levine', 'Iqra physics class', '
    Moneyness', 'Beyond the Shallows', 'Market Design', "lcamtuf's blog
    ", 'Evangelical Textual Criticism', 'New Technology']
==========================================================================
```

Question 4

Use MDS to create a JPEG of the blogs similar to slide 29 of the week 11 lecture. How many iterations were required?

Question 4 continued on next page...

1. To create MDS the first step is to read the BlogData matrix formed in question 1.
2. The blognames, words and data are read from the BlogData text file
3. The scaledown method is called by passing the data as parameter
4. The final step is to call the draw2d method and pass coordinates returned from the scaledown method.

Question 5

Re-run question 2, but this time with proper TFIDF calculations instead of the hack

Not Attempted.

Question 6

1. Create a blog-term matrix. Start by grabbing 100 blogs hosted on https://www.blogger.com. Include: http://f-measure.blogspot.com/ and http://ws-dl.blogspot.com/

1. The first step is to get the list of blogs from blogspot.com. The google search library is used for this purpose. Although this time the query is altered to get the related blogs.
2. A query with site:blogspot.com is made. This restricts the search results to bolgspot.com and related blogs
3. The results are added to the uniqueUrls set but checked to be unique using the canonicalizeURI function
4. The results are written add stored in the UriListBlogspot file

Listing 6: Python Script

```
from googlesearch import search
import csv
from urllib.parse import urlparse
import hashlib

def canonicalizeURI(uri):
    uri = uri.strip()
    if( len(uri) == 0 ):
        return ''
    exceptionDomains = ['www.youtube.com']
    try:
        scheme, netloc, path, params, query, fragment = urlparse( uri )
        netloc = netloc.strip()
        path = path.strip()
        optionalQuery = ''

        if( len(path) != 0 ):
            if( path[-1] != '/' ):
                path = path + '/'

```

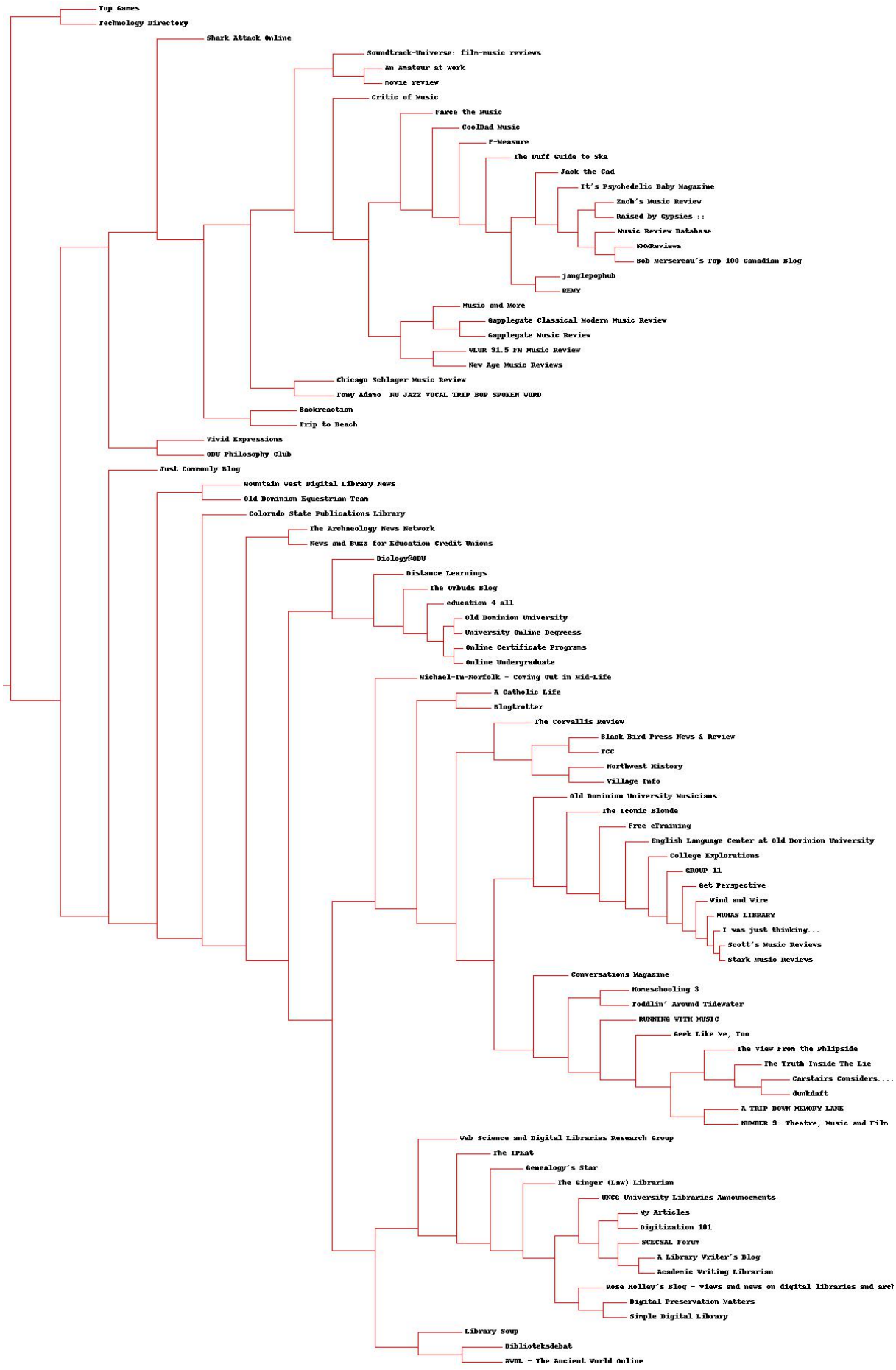
Question 6 continued on next page...

```
        if( netloc in exceptionDomains ):
            optionalQuery = query.strip()

        return netloc
    except:
        print('Error uri:', uri)

    return ''

query = "site:blogspot.com  music review"
uniqueUrls = set()
counter1 = 0
with open("C:\\Q61UriListBlogspot.csv", "w", newline='') as csvFile:
    writer = csv.writer(csvFile)
    currentData = ['http://f-measure.blogspot.com/']
    #print(currentData)
    writer.writerow(currentData)
    for j in search(query, tld="com", num=49, stop=200, pause=3):
        currentCanonicalizedURI = canonicalizeURI(j)
        if currentCanonicalizedURI not in uniqueUrls:
            currentData = [j]
            writer.writerow(currentData)
            uniqueUrls.add(currentCanonicalizedURI)
            print(j)
            counter1 = counter1 + 1
            if(counter1 > 49):
                break
```

1. The findfeed method is utilized to get the RSS feed urls
2. The next step is to utilize the getwordcounts for every blog feed found.
3. The method is used to get the title of the blog and the word count
4. Then for each word in the wordlist for each of the blogs the method forms BlogData matrix and is written to the BlogData text file.
5. This is the matrix data that will be used for the rest of the calculation in the next three problems. This BlogData text file is uploaded to git hub for reference.

1. In order to form the cluster the hcluster method is called by passing the matrix formed in the previous question
2. Followed by the hcluster method we call the drawdendrogram method by passing the blog names and the clusters identified in the previous step
3. The below figure represents the dendrogram of clusters of blogs that are similar.

Question 6 continued on next page...



Question 6 continued on next page...

1. In order to provide a Cluster the blogs using K-Mean and using k is equal to 5 the kcluster method is used
2. To start, the data received in question 1 is utilized and passed to the kcluster method
3. The second parameter used is the k=5
4. It took 5 iterations to for k value being 5
5. The values in each centroid is printed and shown in the below figure

Listing 7: K is 5

```
Iteration 0
Iteration 1
Iteration 2
Iteration 3
Iteration 4
['UNCG University Libraries Announcements', 'The Archaeology News
    Network', 'education 4 all', 'Distance Learnings', 'Online
    Certificate Programs', 'Online Undergraduate', 'Old Dominion
    Equestrian Team', 'Old Dominion University', 'University Online
    Degreess', 'The Ombuds Blog', 'News and Buzz for Education Credit
    Unions', 'Biology@ODU']
================================================================
['F-Measure', 'Gapplegate Classical-Modern Music Review', 'Gapplegate
    Music Review', 'KMMReviews', 'WLUR 91.5 FM Music Review', 'New Age
    Music Reviews', "Zach's Music Review", 'janglepophub', 'Raised by
    Gypsies ::', "Bob Mersereau's Top 100 Canadian Blog", 'Music Review
     Database', 'CoolDad Music', 'The Duff Guide to Ska', 'REMY', 'Jack
     the Cad', 'Critic of Music', 'Farce the Music', 'Music and More',
    "It's Psychedelic Baby Magazine"]
================================================================
["Scott's Music Reviews", 'I was just thinking...', 'Stark Music
    Reviews', 'Wind and Wire', 'Get Perspective', 'The Iconic Blonde',
    'Web Science and Digital Libraries Research Group', 'Free eTraining
    ', 'GROUP 11', 'Mountain West Digital Library News', 'AWOL - The
    Ancient World Online', 'Simple Digital Library', 'MUHAS LIBRARY', '
    Old Dominion University Musicians', 'English Language Center at Old
     Dominion University', "Toddlin' Around Tidewater", 'College
    Explorations']
================================================================
['The Corvallis Review', 'Chicago Schlager Music Review', 'A Catholic
    Life', 'Carstairs Considers....', 'Just Commonly Blog', 'A TRIP
    DOWN MEMORY LANE', 'Shark Attack Online', 'Tony Adamo  NU JAZZ
    VOCAL TRIP BOP SPOKEN WORD', 'NUMBER 9: Theatre, Music and Film', '
    An Amateur at work', 'The View From the Phlipside', 'Top Games', '
    Technology Directory', 'Soundtrack-Universe: film-music reviews', '
    dunkdaft', 'Blogtrotter', 'movie review', 'Black Bird Press News &
    Review', 'Conversations Magazine', "Genealogy's Star", '
    Biblioteksdebat', 'The IPKat', 'Northwest History', 'Vivid
    Expressions', 'TCC', 'ODU Philosophy Club', 'Trip to Beach', '
    Village Info', 'Michael-In-Norfolk - Coming Out in Mid-Life']
================================================================
```

Question 6 continued on next page...

```
['RUNNING WITH MUSIC', 'The Truth Inside The Lie', 'Geek Like Me, Too'
    , 'Homeschooling 3', 'My Articles', "A Library Writer's Blog", '
    Digitization 101', "Rose Holley's Blog - views and news on digital
    libraries and archives", 'Digital Preservation Matters', 'Library
    Soup', 'Colorado State Publications Library', 'Academic Writing
    Librarian', 'The Ginger (Law) Librarian', 'SCECSAL Forum', '
    Backreaction']
================================================================
```

Listing 8: K is 10

```
Iteration 0
Iteration 1
Iteration 2
Iteration 3
Iteration 4
['Just Commonly Blog', 'A TRIP DOWN MEMORY LANE', 'Shark Attack Online
    ', 'Tony Adamo  NU JAZZ VOCAL TRIP BOP SPOKEN WORD', 'Black Bird
    Press News & Review', 'TCC']
================================================================
['RUNNING WITH MUSIC', 'Geek Like Me, Too', 'Homeschooling 3', '
    Backreaction', 'Vivid Expressions', "Toddlin' Around Tidewater"]
================================================================
['Carstairs Considers....', 'The View From the Phlipside', 'dunkdaft',
     'The Truth Inside The Lie', 'Conversations Magazine']
================================================================
["Scott's Music Reviews", 'I was just thinking...', 'Stark Music
    Reviews', 'Wind and Wire', 'Get Perspective', 'The Iconic Blonde',
    'Free eTraining', 'GROUP 11', 'MUHAS LIBRARY', 'Old Dominion
    University Musicians', 'English Language Center at Old Dominion
    University', 'College Explorations']
================================================================
[]
================================================================
['Technology Directory', 'Web Science and Digital Libraries Research
    Group', 'My Articles', 'Digitization 101', "Rose Holley's Blog -
    views and news on digital libraries and archives", 'Digital
    Preservation Matters', "Genealogy's Star", 'Library Soup', '
    Biblioteksdebat', 'The Ginger (Law) Librarian', 'The IPKat']
================================================================
['F-Measure', 'Gapplegate Classical-Modern Music Review', 'Gapplegate
    Music Review', 'KMMReviews', 'WLUR 91.5 FM Music Review', 'Chicago
    Schlager Music Review', 'New Age Music Reviews', "Zach's Music
    Review", 'janglepophub', 'Raised by Gypsies ::', "Bob Mersereau's
    Top 100 Canadian Blog", 'Music Review Database', 'CoolDad Music', '
    The Duff Guide to Ska', 'NUMBER 9: Theatre, Music and Film', 'REMY'
    , 'An Amateur at work', 'Jack the Cad', 'Critic of Music', 'Farce
    the Music', 'Soundtrack-Universe: film-music reviews', 'Music and
    More', "It's Psychedelic Baby Magazine", 'movie review']
================================================================
['Top Games', "A Library Writer's Blog", 'Mountain West Digital
    Library News', 'AWOL - The Ancient World Online', 'Colorado State
    Publications Library', 'Simple Digital Library', 'Academic Writing
    Librarian', 'UNCG University Libraries Announcements', 'SCECSAL
    Forum']
```

Question 6 continued on next page...

```
================================================================
['The Corvallis Review', 'A Catholic Life', 'Blogtrotter', 'The
    Archaeology News Network', 'Northwest History', 'ODU Philosophy
    Club', 'Trip to Beach', 'Village Info', 'Michael-In-Norfolk -
    Coming Out in Mid-Life']
================================================================
['education 4 all', 'Distance Learnings', 'Online Certificate Programs
    ', 'Online Undergraduate', 'Old Dominion Equestrian Team', 'Old
    Dominion University', 'University Online Degreess', 'The Ombuds
    Blog', 'News and Buzz for Education Credit Unions', 'Biology@ODU']
================================================================
```

Listing 9: K is 20

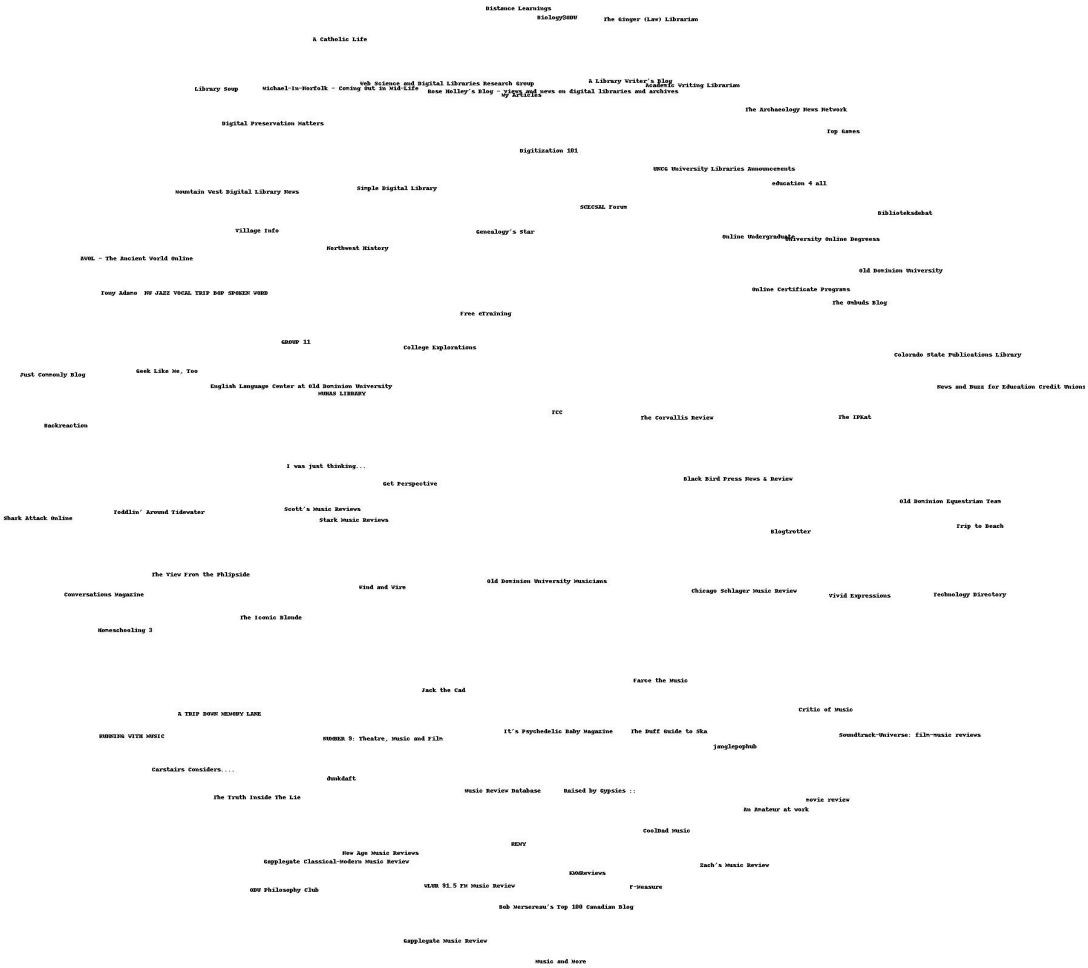
```
Iteration 0
Iteration 1
Iteration 2
Iteration 3
Iteration 4
["A Library Writer's Blog", 'Academic Writing Librarian', 'Trip to
    Beach', 'Biology@ODU']
================================================================
['Online Certificate Programs']
================================================================
['Online Undergraduate']
================================================================
['An Amateur at work', 'Soundtrack-Universe: film-music reviews', '
    movie review']
================================================================
['Tony Adamo  NU JAZZ VOCAL TRIP BOP SPOKEN WORD', 'Technology
    Directory', 'Web Science and Digital Libraries Research Group']
================================================================
['My Articles', 'Digitization 101', "Rose Holley's Blog - views and
    news on digital libraries and archives", 'Digital Preservation
    Matters', 'Simple Digital Library']
================================================================
['Chicago Schlager Music Review', 'Just Commonly Blog', 'CoolDad Music
    ', 'The Duff Guide to Ska', 'NUMBER 9: Theatre, Music and Film', "
    It's Psychedelic Baby Magazine", 'Black Bird Press News & Review',
    'Northwest History', 'Village Info']
================================================================
['Top Games', 'The Archaeology News Network', 'News and Buzz for
    Education Credit Unions']
================================================================
['Biblioteksdebat']
================================================================
['Carstairs Considers....', 'A TRIP DOWN MEMORY LANE', 'The View From
    the Phlipside', 'dunkdaft', 'RUNNING WITH MUSIC', 'The Truth Inside
     The Lie', 'Geek Like Me, Too', 'Homeschooling 3', 'Conversations
    Magazine', 'Backreaction', 'Vivid Expressions']
================================================================
['Blogtrotter']
================================================================
['Distance Learnings']
================================================================
```

Question 6 continued on next page...

```
['A Catholic Life', 'Mountain West Digital Library News', 'Library
    Soup', 'UNCG University Libraries Announcements', 'SCECSAL Forum',
    'TCC', 'ODU Philosophy Club']
================================================================
['F-Measure', 'Gapplegate Classical-Modern Music Review', 'Gapplegate
    Music Review', 'KMMReviews', 'WLUR 91.5 FM Music Review', 'New Age
    Music Reviews', "Zach's Music Review", 'janglepophub', 'Raised by
    Gypsies ::', "Bob Mersereau's Top 100 Canadian Blog", 'Music Review
     Database', 'Shark Attack Online', 'REMY', 'Jack the Cad', 'Critic
    of Music', 'Farce the Music', 'Music and More']
================================================================
["Scott's Music Reviews", 'I was just thinking...', 'Stark Music
    Reviews', 'Wind and Wire', 'Get Perspective', 'The Iconic Blonde',
    'Free eTraining', 'GROUP 11', 'MUHAS LIBRARY', 'Old Dominion
    University Musicians', 'English Language Center at Old Dominion
    University', "Toddlin' Around Tidewater", 'College Explorations']
================================================================
["Genealogy's Star", 'Colorado State Publications Library', 'The
    Ginger (Law) Librarian', 'The IPKat', 'Michael-In-Norfolk - Coming
    Out in Mid-Life']
================================================================
['AWOL - The Ancient World Online']
================================================================
['The Corvallis Review', 'education 4 all', 'Old Dominion Equestrian
    Team', 'Old Dominion University', 'University Online Degreess', '
    The Ombuds Blog']
================================================================
```

1. To create MDS the first step is to read the BlogData matrix formed in question 1.
2. The blognames, words and data are read from the BlogData text file
3. The scaledown method is called by passing the data as parameter
4. The final step is to call the draw2d method and pass coordinates returned from the scaledown method.
5. To compare the random cluster vs cluster made up from the related blogs they both do result in clusters.
6. The cluster although formed with the related blogs are much closer than those of the random blogs but we could still make some meaning out the of the blogs relationship with the random blogs collected and its results.

Question 6 continued on next page...


Distance Learnings
BiologyDODU
The Ginger (Law) Librarian
A Catholic Life
Web Science and Digital Libraries Research Group
A Library Writer's Blog
Academic Writing Librarian
Library Soup
Michael-in-Norfolk - Coming Out in Mid-Life
Rose Holley's Blog - views and news on digital libraries and archives
My Articles
The Archaeology News Network
Digital Preservation Matters
Top Games
Digitization 101
UNCG University Libraries Announcements
education 4 all
Mountain West Digital Library News
Simple Digital Library
SCECSAL Forum
Bibliotekskebat
Village Info
Genealogy's Star
Online Undergraduate
University Online Degreess
Northwest History
AWOL - The Ancient World Online
Old Dominion University
Tony Adams PHI JAZZ VOCAL TRIP DOP SPOKEN WORD
Online Certificate Programs
The Ombuds Blog
Free eTraining
GROUP 11
College Explorations
Colorado State Publications Library
Just Comically Blog
Geek Like Me, Too
English Language Center at Old Dominion University
HONAS LIBRARY
News and Buzz for Education Credit Unions
Fcc
The Corvallis Review
The IPKat
Backreaction
I was just thinking...
Get Perspective
Black Bird Press News & Review
Old Dominion Equestrian Team
Shark Attack Online
Toddlin' Around Tidewater
Scott's Music Reviews
Stark Music Reviews
Trip to Beach
Blogtrotter
The View From the Philpside
Old Dominion University Musicians
Wind and Wire
Chicago Schlager Music Review
Vivid Expressions
Technology Directory
Conversations Magazine
The Iconic Blonde
Homeschooling 3
Farce the Music
Jack the Cad
Critic of Music
A TRIP DOWN MEMORY LANE
It's Psychedelic Baby Magazine
The Duff Guide to Ska
Soundtrack-Universe: film-music reviews
RUMBLING WITH MUSIC
NUMBER 9: Theatre, Music and Film
janglepophub
Carstairs Considers....
dunkdaft
Music Review Database
Raised by Gypsies ::
The Truth Inside The Lie
movie review
An Amateur at work
Coolbad Music
NENV
New Age Music Reviews
Applegate Classical-Modern Music Review
Zach's Music Review
KWARreviews
ODU Philosophy Club
WLNR 91.5 FM Music Review
F-Measure
Bob Mersereau's Top 100 Canadian Blog
Applegate Music Review
Music and More