CS432 Spring 2018

Assignment 7

Jonathan Kruszewski

04/28/18

CS 432/532 Web Science
Spring 2018
http://anwala.github.io/lectures/cs532-s18/

Assignment #7
Due: 11:59pm March 31

(10 points; 2 points for each question and 2 points for aesthetics)

Support your answer: include all relevant discussion, assumptions, examples, etc.

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) 1000 terms, this is *after* the criteria on p. 32 (slide 8) has been satisfied. Remember that blogs are paginated.

2. Create an ASCII and JPEG dendrogram that clusters (i.e., HAC) the most similar blogs (see slides 13 & 14). Include the JPEG in your report and upload the ascii file to github (it will be too unwieldy for inclusion in the report).

3. Cluster the blogs using K-Means, using k=5,10,20. (see slide 25). Print the values in each centroid, for each value of k. How many iterations were required for each value of k?

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

Part 1:

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) 1000 terms, this is *after* the criteria on p. 32 (slide 8) has been satisfied. Remember that blogs are paginated.

To accomplish this task I used a bash file to issue the curl command found in file “CurlBlogs.sh”, the curl command was repeated 498 times and saves the each result into an individual “curlFile[n].txt” into the folder “CurlFiles” and was then used on the two urls specified in the instructions.

```
filename='CurlFiles/curlFile'${i}'.txt'
curl -L -o /dev/null -w %{url_effective} 'http://www.blogger.com/next-blog?navBar=true&blogID=3471633091411211117' > ${filename}
```

After this step “GetUrl.py” is run which reads each curlFile from the previous step and removes “/?expref=next-blog” then adds “/feeds/posts/default” to the end of each. Two txt files are then output, the first file: “unmodifiedURLS.txt” contains a complete list of blog urls, the second file: “feedlist.txt” contains the same list of urls with all duplicates removed.

```
urls = []
for x in inF:
    tempStr = ''
    tempStr = x.rstrip('/?expref=next-blog')
    tempStr = tempStr + '/feeds/posts/default'
    urls.append(tempStr)
    print(tempStr)
```

After this step “generatefeedvector.py” from https://github.com/arthur-e/Programming-Collective-Intelligence/tree/master/chapter3 is run with a small modification. The modified version functions the same as the version provided except that it outputs a file with a list of any feeds that failed to parse.

The blog data can be found in “blogdata1.txt” or the blogdata1 excel sheets in the containing folder.

Part 2:

2. Create an ASCII and JPEG dendrogram that clusters (i.e., HAC) the most similar blogs (see slides 13 & 14). Include the JPEG in your report and upload the ascii file to github (it will be too unwieldy for inclusion in the report).

This task was accomplished by using “Dend.py” which used a slightly modified version of the “cluster.py” python code provided at https://github.com/arthur-e/Programming-Collective-Intelligence/tree/master/chapter3 .“cluster.py” was modified by changing the printclust funtion to make it output a file “cluserASCII.txt” with the ASCII dendrogram. “Dend.py” calls the functions in “cluster.py” to get both the ASCII and JPEG dendrograms. The JPEG can be found at the end of this report.

<u>Part 3</u>

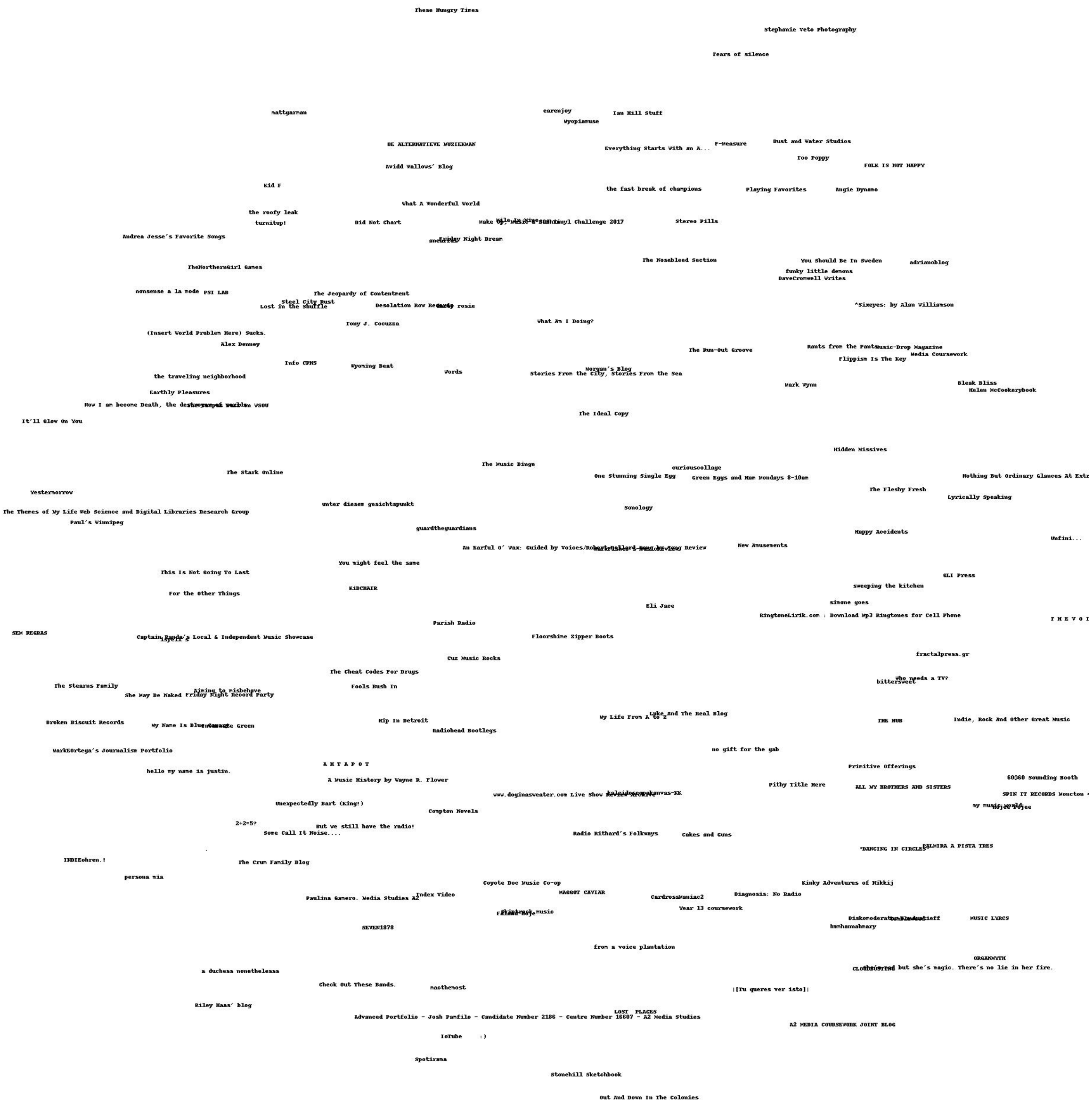
3. Cluster the blogs using K-Means, using k=5,10,20. (see slide 25). Print the values in each centroid, for each value of k. How many iterations were required for each value of k?

To accomish these tasks a python program “kmod.py” was used along with “cluster.py”. “kmod.py” called the kcluster function from “cluster.py” for k = 5, 10, 20. “cluster.py” was modified for this task to output to a file “kcluster[k].txt” after each iteration after which “kmod.py” would output the centroid for each to its corresponding file. K = 5 took 9 iterations, K = 10 took 8 iterations, and K = 20 took 7 iterations.

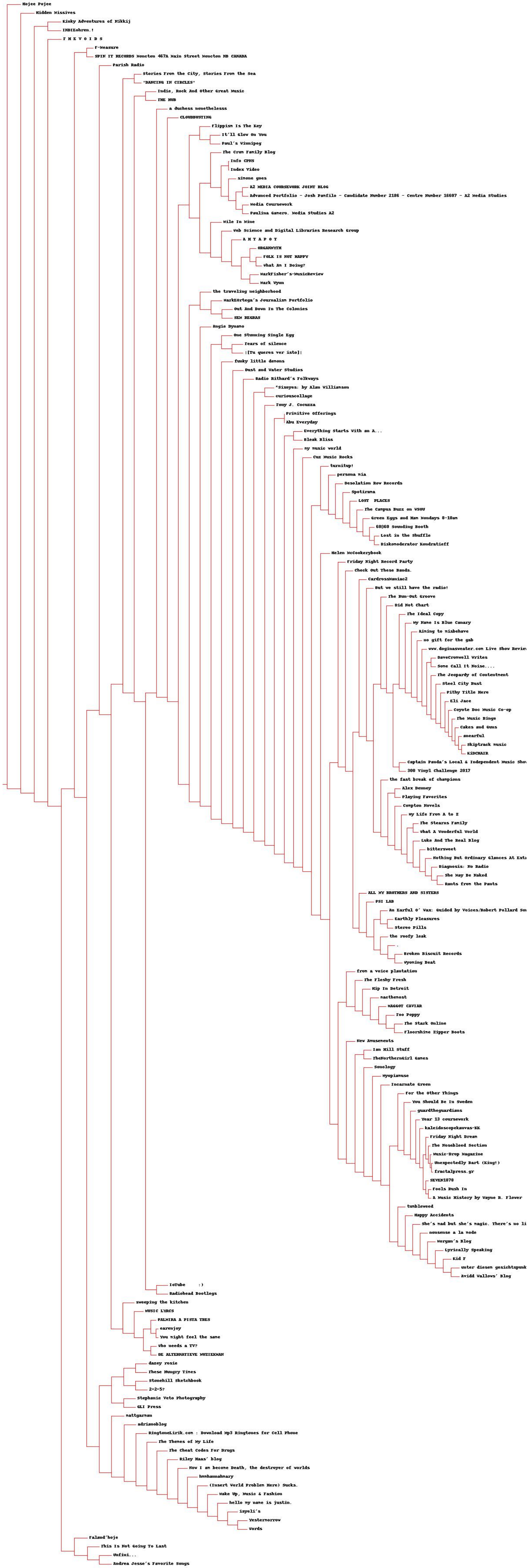
Part 4

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

To accomish these tasks a python program “mds.py” was used along with “cluster.py”. It took 2 iterations. The mds image can be viewed below or found sepretly as “mds.jpg”



Dendrogram:
“blogClusetr.jpg”


Mojee Pojee
Hidden Missives
Kinky Adventures of Nikkij
INDIEohren.!
T H E V O I D S
F-Measure
SPIN IT RECORDS Moncton 467A Main Street Moncton NB CANADA
Parish Radio
Stories From the City, Stories From the Sea
"DANCING IN CIRCLES"
Indie, Rock And Other Great Music
THE HUB
a duchess nonethelesss
CLOUDBUSTING
Flippism Is The Key
It’ll Glow On You
Paul’s Winnipeg
The Crum Family Blog
Info CPNS
Index Video
simone goes
A2 MEDIA COURSEWORK JOINT BLOG
Advanced Portfolio - Josh Pamfilo - Candidate Number 2186 - Centre Number 16607 - A2 Media Studies
Media Coursework
Paulina Gamero. Media Studies A2
Mile In Mine
Web Science and Digital Libraries Research Group
A H T A P O T
ORGANMYTH
FOLK IS NOT HAPPY
What Am I Doing?
MarkFisher’s-MusicReview
Mark Wynn
the traveling neighborhood
MarkEOrtega’s Journalism Portfolio
Out And Down In The Colonies
SEM REGRAS
Angie Dynamo
One Stunning Single Egg
Tears of silence
|[Tu queres ver isto]|
funky little demons
Dust and Water Studios
Radio Rithard’s Folkways
*Sixeyes: by Alan Williamson
curiouscollage
Tony J. Cocuzza
Primitive Offerings
Abu Everyday
Everything Starts With an A...
Bleak Bliss
my music world
Cuz Music Rocks
turnitup!
persona mia
Desolation Row Records
Spotirama
LOST PLACES
The Campus Buzz on WSOU
Green Eggs and Ham Mondays 8-10am
60@60 Sounding Booth
Lost in the Shuffle
Diskomoderator Kondratieff
Helen McCookerybook
Friday Night Record Party
Check Out These Bands.
CardrossManiac2
But we still have the radio!
The Run-Out Groove
Did Not Chart
The Ideal Copy
My Name Is Blue Canary
Aiming to misbehave
no gift for the gab
www.doginasweater.com Live Show Reviev
DaveCromwell Writes
Some Call It Noise....
The Jeopardy of Contentment
Steel City Rust
Pithy Title Here
Eli Jace
Coyote Doc Music Co-op
The Music Binge
Cakes and Guns
amearful
Skiptrack music
KiDCHAIR
Captain Panda’s Local & Independent Music Shov
300 Vinyl Challenge 2017
the fast break of champions
Alex Denney
Playing Favorites
Compton Novels
My Life From A to Z
The Stearns Family
What A Wonderful World
Luke And The Real Blog
bittersweet
Nothing But Ordinary Glances At Extı
Diagnosis: No Radio
She May Be Naked
Rants from the Pants
ALL MY BROTHERS AND SISTERS
PSI LAB
An Earful O’ Wax: Guided by Voices/Robert Pollard Sor
Earthly Pleasures
Stereo Pills
the roofy leak
.
Broken Biscuit Records
Wyoming Beat
from a voice plantation
The Fleshy Fresh
Hip In Detroit
macthemost
MAGGOT CAVIAR
Too Poppy
The Stark Online
Floorshime Zipper Boots
New Amusements
Ian Hill Stuff
TheNorthernGirl Games
Sonology
Myopiamuse
Incarnate Green
For the Other Things
You Should Be In Sweden
guardtheguardians
Year 13 coursework
kaleidoscopekanvas-KK
Friday Night Dream
The Nosebleed Section
Music-Drop Magazine
Unexpectedly Bart (King!)
fractalpress.gr
SEVEN1878
Fools Rush In
A Music History by Wayne R. Flower
tumbleweed
Happy Accidents
She’s mad but she’s magic. There’s no li
nonsense a la mode
Morgan’s Blog
Lyrically Speaking
Kid F
unter diesem gesichtspunk
Avidd Wallows’ Blog
IoTube :)
Radiohead Bootlegs
sweeping the kitchen
MUSIC LYRCS
PALMIRA A PISTA TRES
earenjoy
You might feel the same
Who needs a TV?
DE ALTERNATIEVE MUZIEKMAN
dazey rosie
These Hungry Times
Stonehill Sketchbook
2+2=5?
Stephanie Veto Photography
GLI Press
mattgarman
adrianoblog
RingtoneLirik.com : Download Mp3 Ringtones for Cell Phone
The Themes of My Life
The Cheat Codes For Drugs
Riley Haas’ blog
Now I am become Death, the destroyer of worlds
hmmhannahmary
(Insert World Problem Here) Sucks.
Make Up, Music & Fashion
hello my name is justin.
isyeli’s
Yestermorrow
Words
Faland’hoje
This Is Not Going To Last
Unfini...
Andrea Jesse’s Favorite Songs