CS 535: Assignment #7

Due on 11:59pm March 31, 2018

Alexander C. Nwala 16:20

David Sinclair

March 25, 2018

Contents

1 Problem 1

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

1.1 Question 1

1. Create a blogterm 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 entrytitle (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.

1.2 Answer 1

To start with I used the following bash program 01_getblog.sh to gather 150 blogs. This program created a file called blogurl.txt. Examples of blogurl.txt are:

http://ridingaborrowedbike.blogspot.com/?expref=next-blog
http://theslowmusicmovement.blogspot.com/?expref=next-blog
http://www.gypsyrhapsody.com/?expref=next-blog
http://nathaliealves.blogspot.com/?expref=next-blog
http://dana9morgan.blogspot.com/?expref=next-blog

I then manually added the following two blogs per requirement 1 to the top of the file.

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

After adding the above blogs a ran 02_cleandata.py. This did two things. Removed the ?expref=next-blog from the end of every line in blogurl.txt and then added feeds/posts/default to the end of the line. I created two files to do this process called urlclean.txt and blogclean.txt. The blogclean.txt is listed below.

http://f-measure.blogspot.com/feeds/posts/default
http://ws-dl.blogspot.com/feeds/posts/default
http://ridingaborrowedbike.blogspot.com/feeds/posts/default
http://theslowmusicmovement.blogspot.com/feeds/posts/default
http://www.gypsyrhapsody.com/feeds/posts/default
http://nathaliealves.blogspot.com/feeds/posts/default

I then used the 03_blogmatrix.py to create the blogdata1.txt.

```
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)
        words = getwords(e.title)
        for word in words:
            wc.setdefault(word, 0)
            wc[word] += 1

    return (d.feed.title, wc)

apcount = {}
wordcounts = {}
feedlist = [line for line in open('blogclean.txt','r')]
for feedurl in feedlist:
    try:
        (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)
```

The following blogs did not parse correctly and were not included in any of the blogdata1

Failed to parse feed http://www.gypsyrhapsody.com/feeds/posts/default
Failed to parse feed http://globalgoon.blogspot.com/feeds/posts/default
Failed to parse feed http://momslilprincess.blogspot.com/feeds/posts/default
Failed to parse feed http://globalgoon.blogspot.com/feeds/posts/default
Failed to parse feed http://www.punkrockteaching.org/feeds/posts/default
Failed to parse feed http://www.gypsyrhapsody.com/feeds/posts/default
Failed to parse feed http://www.chrisanne-grise.com/feeds/posts/default
Failed to parse feed http://www.chrisanne-grise.com/feeds/posts/default

I initally looked at the title only which had 14 words on it. This was enough to do any real data analysis with. I saved the data as blogdata1_titleonly.txt with the title words only. Here is the list of words and the first three blogs:

Blog in it music a for you s and is new to of my i
Fran Brighton 0 0 0 1 0 0 0 0 01 0 0 0 0
Bleak Bliss 1 0 2 3 00 2 1 0 1 0 5 0 0
Green Eggs and Ham Mondays 8-10am 0 0 1 1 0 0 1 10 0 0 0 0 0

I then added the summary by changing the following line in the 03_blogmatrix.py program.

```
        # Extract a list of words
        words = getwords(e.title + ’ ’ + summary)
#        words = getwords(e.title)
        for word in words:
            wc.setdefault(word, 0)
            wc[word] += 1
```

I looked at the title and summary and had 1023 words. I believe this is enough to do some data analysis with. I saved the data as blogdata1.txt with the title words and summary. Here is the list of words:

Blog mid moments die goes brother totally simple player created knew named honestly listen lp fact happens american company series matter frank popular hearing god holding welcome c should stand minutes folk resp ect days problem quality such moved listener space mother media rough rest eight completely copy phone originally decided open takes somehow piano march gives scene verse posted though seem subject miss impressive adding http walk perhaps sad stories men ending mom songs normal book mind news clearly current singles john e friend close similar due english happen clear overall sides positive whatever cover p yourself river sex far co side putting shot america forget skin low learned bought giving surprised indie guitars lose en hate anyway decade random service look held false v hope every performance melodies twin school exciting lyrics fighting research directly heaven hold o appreciate doubt youth goal concert loud pictures understand forward perfect alternative g winter break becoming variety sharing remember somewhere anymore fingers mike absolutely mix bigger taking history six west international process game began wild lo weather earlier keeps highlights brilliant pull wasn slightly magazine features change dog playing moving voice loss driving end minor snow particular become journey below short done impossible added rarely cause months releases nick try blog guitarist el nice sky dancing friends sorry certainly saw books particularly included passing north april plan christmas longer effort real lack smith bed based create bass brand part seems focused call terms musicians ain lost radiohead heart followed help mean left hook emotional whole tell actual word because paper lead tv together onto least trees middle dream otherwise begin knowing anna local write sun thinking offer r career amount font area group vocal via figure details prince wave questions thanks facebook sounding sit members self towards recording high else ben sets per wwwwithin bring fans hip forever talent awesome interesting kill add showing liked true sings hell songwriter control moment small cool community la monday experience official happy wonderful fight remix changing performed total single set closer stuck chance salt above lines choose singer passion allow stood throughout strange performances imagine purpose own trying al myself food bon pre dreams cry stone couldn likely soul store words culture eat l filled despite sitting picked ok wish da wall conversation lies trouble fully met bar easy answer parents corner share main couple east bit bright used need crowd ground plays james makes war doing none less table continue perfectly remains kids meaning machine spent famous hits fit personal extremely buy tunes classic why beats reading soundtrack room next lake names themselves shared please running thing realized november several demo split range thank animals soft brought backing artists latin paul baby attempt huge father recently writing turns team killing stop dark afraid staying fun moon across bad talking outro upcoming learn hadn alt catplenty latest apart title festival aren hours intro question character sat sing turn opportunityform shows went proud won run find various enjoyed might strong duo pain lives roll deal parts human believe hoping tend today released everyone thought many sweet birds morning general cut indeed studio july pop michael care albums missing style picture led excited glad learning came kept faith reach melody voices catchy york size sound response girl center nature seriously ahead usually especially smile took vocals public wants dear slow gone considered literally member tracks sense hit started effect camera era often daily air finding appeal nothing ways stuff knows half easily fresh age actually edge loved everything getting pink theme making date four star com places debut however car bob system guys start joy hand sea powerful weeks choice projects almost anthem nearly special large society already physical tomorrow attack given ability keeping among fellow hot check ep road including color warm u following titled week feature broken himself promise experiences solid themes turned concept kind instead material excellent incredible jazz link watch ready appear double top super ghost changes electric yeah magic wait eye highly closing king college interview between both entirely heads favorite outside each weekend point anyone note gets hair section under happening k power singing growing works result modern folks sort either sister against speak town feet fast catch weren step blue bowie example complete issue yes managed drummer helped living beginning leaves older featuring tape hands drum th front brown visit quite straight pieces fan haven biggest killer order reasons taken ourselves turning house ride creative leaving head st consider fair entire album better design records artist produced radio fall roots wonder de field white previous door extra yet collection young said tune clean women missed lived jack since type sent felt epic working known beyond focus online mostly incredibly likes state idea minute perform plastic difficult shouldn beach ideas worked full tom children season male seven cold issues third him whether george lovely once lights emotions film event business fear case coming mark ocean x hour possibly black others alone oh putclick became ears hard expect etc tried speaking simply listened rain name although drink number late keep deep written natural round sometimes involved feels pure money musical move string points read worth record past london ask bridge talented saying reality starts water seen street starting spend act listening any gonna able golden story notes b rolling needs thoughts acts looked taste original kid spot photo cd seeing female also woman problems times city hall evening probably leave mr pretty sounds n version midnight without memory fuck present national van nobody child para videos plus gold non behind constantly creating release reason shit available bill wrong decent blood unfortunately british certain stay face ones didn watching someone solo red class waiting its bands realize david talk told drop save results near vinyl eventually social truly asked rare body before hey caught august chorus beautiful pt spring return stars approach trip brian fantastic ten underground pick multiple needed boys drive three lucky gave using rather volume influence tour island dead project touch elements audience dj meet heat park free job riffs piece recorded major force greatest standing weird ended further currently birthday drums former heavy production wanna energy five shadow review fire breaking electronic soon track del quiet having country happened october typical rock success truth early relationship shape june attention stage count along different second punk cannot tears unique called beat play changed person content means lots family seat use beauty list immediately crazy influences looks funny future everywhere found line alive band says recordings inspired finish ago walking tonight length anything son dropped enough f meant y video bottom los empty channel rhythm hearts seemed familiar boy wouldn quickly green course genre noticed inside few sold during justin february guess itself limited guy exactly maybe recent follow month alex obvious favourite possible falling upon website label noise interested earth dan setting important hop january tiny appeared dance final wanted same mention california honest obviously waves brain friday summer serious poor level packed wide finally later jones metal death mentioned eyes mine image notice quick does light playlist played higher fine apparently must sure until acoustic looking losing art feeling amazing reference message download isn post heard enjoy hear include lot support box links okay whose doesn anti born sleep movie party brings guitar girls definitely youtube strings club wrote kiss page blues ryan opening land damn comes

The list of blogs is listed below.

Oh Yes Jónsi!!
hello my name is justin.
My Name Is Blue Canary
Indie, Rock And Other Great Music
Music-Drop Magazine
Radiohead Bootlegs
CLOUDBUSTING
Skiptrack music
Primitive Offerings
Green Eggs and Ham Mondays 8-10am
Unexpectedly Bart (King!)
simone goes
TheNorthernGirl Games
kaleidoscopekanvas-KK
Stephanie Veto Photography
Fran Brighton
SEVEN1878
A to Zappa - Song of the day
ORGANMYTH
isyeli’s
Kid F
She May Be Naked
The Stearns Family
Stereo Pills
Helen McCookerybook
Faland’hoje
www.doginasweater.com Live Show Review Archive
You might feel the same

The Run-Out Groove
Wyoming Beat
the traveling neighborhood
The Themes of My Life
BEEHIVE CANDY
Ringtone Lirik
Morgan’s Blog
Paulina Gamero. Media Studies A2
Jasmine Hodge
headphonehead
earenjoy
não ponho música
Three Pups, One City
Web Science and Digital Libraries Research Group
Who needs a TV?
sweeping the kitchen
IoTube :)
Incarnate Green
New Amusements
.
PSI LAB
She’s mad but she’s magic. There’s no lie in her fire.
F-Measure
KiDCHAIR
SEM REGRAS
fractalpress.gr
What Am I Doing?
The Campus Buzz on WSOU
Chemical Robert!
Lost in the Shuffle
mattgarman
Rants from the Pants
INDIEohren.!
Stories From the City, Stories From the Sea
FACSO S2 Blog sessions
Hip In Detroit
GLI Press
ELLIA TOWNSEND A2
One Stunning Single Egg
Pewter & Puddles
Pithy Title Here
Soundcheck
Angie Dynamo
Nothing But Ordinary Glances At Extraordinary Things
The Cheat Codes For Drugs
Spinitron Charts
Out of my Mind
Bleak Bliss
a duchess nonethelesss
Spotirama
The Fleshy Fresh
Dust and Water Studios
Words
Yestermorrow
The Nosebleed Section
Too Poppy

MAGGOT CAVIAR
Radio Rithard's Folkways
Stonehill Sketchbook
Encore
Friday Night Dream
macthemost
Happy Accidents
Captain Panda's Local & Independent Music Showcase
persona mia
MPC
The Slow Music Movement
Riley Haas' blog
The World's First Internet Baby
A Music History by Wayne R. Flower
hmmhannahmary
Skywriting
the fast break of champions
i'm in too truthful a mood
Did Not Chart
Luke And The Real Blog

2 Problem 2

2.1 Question 2

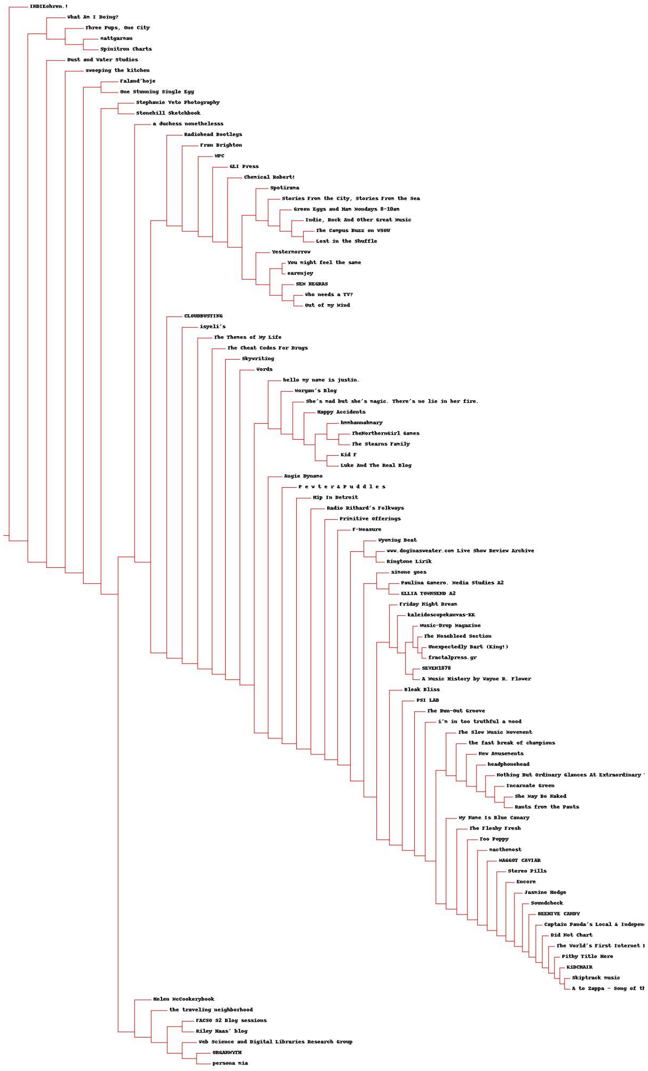
2. Create an ASCII and JPEG dendrogram that clusters (i.e., HAC) the most similar blogs (see slides 13 & 14). Include the JPEG in your report and upload the ascii file to github (it will be too unwieldy for inclusion in the report).

2.2 Answer 2

Because of some font issues with some of the blog titles. They needed by be removed. The following blogs were removed and a new file was created called blogdata1rm.txt file.

Oh Yes Jónsi!!
não ponho música
IoTube :)
.

To create the JPEG dendrogram from slide 13 & 14. I used programs 04_ascii.py and 05_demjpg.py. The ascii file printed on the screen and I copied it into file ascii.txt. The JPEG dendrogram can be seen below.



3 Problem 3

3.1 Question 3

3. Cluster the blogs using KMeans, 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?

3.2 Answer 3

For this I created 3 programs called 06_kmod_5.py, 06_kmod_10.py, 06_kmod_20.py. Each program listed the iterations required then printed out the Centriods. Each one is listed below.

For 5 KMeans it took 9 Iterations

1st Centriod

$'hellomynameisjustin.','CLOUDBUSTING','TheNorthernGirlGames','StephanieVetoPhotography',"isyeli's",'KidF','S$
,

2nd Centriod

$'PrimitiveOfferings','www.doginasweater.comLiveShowReviewArchive','WyomingBeat','RingtoneLirik','TheCampusBu$

3rd Centriod

$'Music-DropMagazine','GreenEggsandHamMondays8-10am','UnexpectedlyBart(King!)','kaleidoscopekanvas-KK','S$

4th Centriod

$'MyNameIsBlueCanary','Indie,RockAndOtherGreatMusic','Skiptrackmusic','simonegoes','AtoZappaSongoftheday','Ste$

5th Centriod

$'RadioheadBootlegs','FranBrighton','ORGANMYTH',"Faland'hoje",'Youmightfeelthesame','earenjoy','WebScienceand$

For 10 KMeans it took 7 Iterations

1st Centriod

$'Skiptrackmusic','simonegoes','AtoZappa-Songoftheday','StereoPills','TheRunOutGroove','BEEHIVECANDY','Jasm$

2nd Centriod

$'CLOUDBUSTING','TheNorthernGirlGames','StephanieVetoPhotography','ORGANMYTH',"isyeli's",'KidF','SheM$

3rd Centriod

$'Indie,RockAndOtherGreatMusic','RadioheadBootlegs','GreenEggsandHamMondays8-10am','FranBrighton','sweeping$

4th Centriod

$'hellomynameisjustin.','SEVEN1878','WhoneedsaTV?','SEMREGRAS','OutofmyMind','Yestermorrow'$

5th Centriod

$'MyNameIsBlueCanary','PrimitiveOfferings','WyomingBeat','ThreePups,OneCity','NewAmusements','Pewter\&Pud$

6tht Centriod

$$'www.doginasweater.comLiveShowReviewArchive','RingtoneLirik'$$

7th Centriod

$$'WhatAmIDoing?','mattgarman','DustandWaterStudios'$$

8th Centriod

$'Music-DropMagazine','UnexpectedlyBart(King!)','kaleidoscopekanvas-KK','Youmightfeelthesame','earenjoy','fract$

9th Centriod

$$"Faland'hoje",'TheThemesofMyLife','PaulinaGamero.MediaStudiesA2','ELLIATOWNSENDA2'$$

10th Centriod

$$'AMusicHistorybyWayneR.Flower'$$

For 20 KMeans it took 4 Iterations

1st Centriod

$'Youmightfeelthesame','earenjoy','WhoneedsaTV?','SEMREGRAS','ChemicalRobert!','INDIEohren.!','OutofmyMind$

2nd Centriod

$$'RadioheadBootlegs','FranBrighton','OneStunningSingleEgg'$$

3rd Centriod

4th Centriod

$$'SpinitronCharts','MAGGOTCAVIAR','StonehillSketchbook','personamia'$$

5th Centriod

$'PrimitiveOfferings','www.doginasweater.comLiveShowReviewArchive','WyomingBeat',"RadioRithard'sFolkways"$

6th Centriod

$'AtoZappaSongoftheday','PSILAB','BleakBliss','TheFleshyFresh','Encore','macthemost',"CaptainPanda'sLocal\&Inde$

7th Centriod

$$'TheCheatCodesForDrugs','Skywriting'$$

8th Centriod

$$'F-Measure','GLIPress','Spotirama'$$

9th Centriod

$'Music{-}DropMagazine','UnexpectedlyBart(King!)','kaleidoscopekanvas{-}KK','SEVEN1878','fractalpress.gr','TheNose$

10th Centriod

$$'RingtoneLirik'$$

11th Centriod

$'hellomynameisjustin.','TheNorthernGirlGames','StephanieVetoPhotography','KidF','SheMayBeNaked','TheStearnsF$

12th Centriod

$$'JasmineHodge'$$

13th Centriod

$'MyNameIsBlueCanary','simonegoes','StereoPills','BEEHIVECANDY','ThreePups,OneCity','Soundcheck','TooPopp$

14th Centriod

$$'ORGANMYTH','WebScienceandDigitalLibrariesResearchGroup'$$

15th Centriod

$$'WhatAmIDoing?','mattgarman'$$

16th Centriod

$'Indie,RockAndOtherGreatMusic','GreenEggsandHamMondays8{-}10am','sweepingthekitchen','StoriesFromtheCity,Stor$

17th Centriod

$'Skiptrackmusic','TheRun{-}OutGroove','KiDCHAIR','PithyTitleHere','DustandWaterStudios','TheSlowMusicMoveme$

18th Centriod

$'CLOUDBUSTING',' PaulinaGamero.MediaStudiesA2',' TheCampusBuzzonWSOU',' LostintheShuffle',' Pewter\&Puddl$

19th Centriod

$$"isyeli's","Faland'hoje",' NothingButOrdinaryGlancesAtExtraordinaryThings',' Yestermorrow'$$

20th Centriod

$$'AMusicHistorybyWayneR.Flower'$$

4 Problem 4

4.1 Question 4

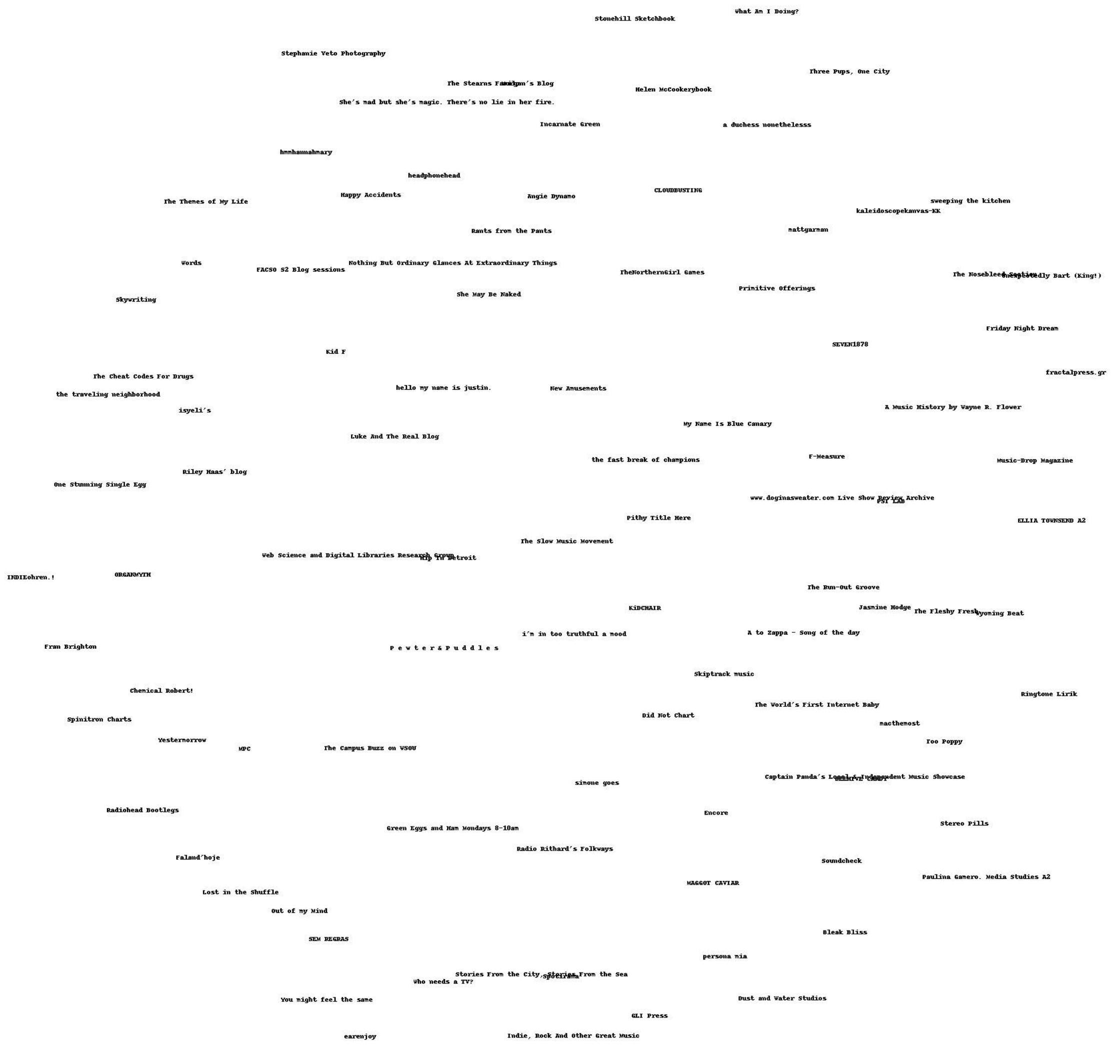
4. Use MDS to create a JPEG of the blogs similar to slide 29 of the week 11 lecture. How many iterations were required?

4.2 Answer 4

The following data was produced using the information from slide 29 of the week 11 lecture. I used the following program 07_mds.py to create these numbers.

```
4547.043521489129 3634.920533462348 3515.8839399478115 3458.2812091642527 3413.5608774776892 3374.276948606173
3348.5127490457294 3330.3940050771434 3317.3067505599533 3305.1969527040324 3295.4500908367454 3285.6891855480026
3277.2275986559007 3270.1385901969525 3263.082799426065 3255.8755514427567 3249.6470923772954 3244.011941280313
3238.7892469552226 3234.086579881611 3229.0613466355403 3224.193711293276 3218.728335385039 3212.8500053530183
3207.689594320295 3203.150266194056 3198.9533959022415 3194.604889994371 3190.2897368511835 3186.1387931376194
3182.1237121996005 3178.307526328884 3174.367903053554 3170.3350424791097 3166.752798856369 3163.6780409874955
3160.747964784802 3157.7328575197034 3154.77644914824 3152.340211326684 3150.327400154053 3148.4216402984066
3146.772738354147 3145.4361232921538 3143.734960076024 3142.2316167059453 3140.7912084980576 3139.3835955178583
3137.692572248288 3135.812385381646 3133.7239227427735 3131.648109416603 3129.5289450971286 3127.449995254269
3125.071438129394 3122.7193354907677 3120.511591126518 3118.388668796023 3115.7190542096773 3113.3641392680834
3111.59270032171 3109.946294358839 3108.4909193755507 3107.1045891536796 3105.8281331270755 3104.3869254043566
3102.7047333666487 3101.097304378567 3099.543313580798 3098.084514759206 3096.672042743501 3095.224982182776
3093.7313904382568 3092.027581045574 3090.2104044018242 3088.149298803071 3085.775778734686 3083.383085340877
3081.280947880271 3079.2365497560872 3077.298889320404 3075.8079645978023 3074.3080221743007 3072.8519787126897
3071.385544451662 3069.9776196329103 3068.520092978661 3067.149274019729 3066.0729659332546 3065.082969393555
3064.145666242849 3063.247777505517 3062.4154731034237 3061.4793494740084 3060.47169346936 3059.3756354829725
3058.2960591681913 3057.2799939873075 3056.3485490106896 3055.4951120820288 3054.558508211861 3053.5233297680397
3052.492524459361 3051.4628918740345 3050.5424254042914 3049.7517906129497 3049.011887797494 3048.4267260927068
3047.870466981071 3047.3233654694677 3046.720716078761 3046.08862047069 3045.3160050260144 3044.4723715462014
3043.6142328919223 3042.751093485595 3041.9547017515383 3041.11679117308 3040.203187631837 3039.3155044497316
3038.4337518292828 3037.3790657049644 3036.231612133021 3035.034051890081 3033.92414370466 3032.8730533644166
3031.8514384291143 3031.0115886619897 3030.1571915968875 3029.315418085879 3028.5644946694356 3027.8135124742435
3027.0743907044857 3026.3341503127654 3025.670914861522 3025.0419271220862 3024.319727365977 3023.55506678629
3022.896278116901 3022.24080916216 3021.625202638946 3021.0867707435423 3020.4753045684206 3019.804570141059
3019.190171500445 3018.5863459090597 3018.07085127555 3017.5614498521454 3017.012432153466 3016.4620731400714
3016.023992299319 3015.624103590204 3015.2562011442615 3014.830524961372 3014.4008117479275 3013.9826495271404
3013.4962474781896 3012.9286083410416 3012.269670369528 3011.5501389901888 3010.8170161469952 3010.2410459543544
3009.663304101802 3009.1072224531044 3008.5248957689 3007.9160790735205 3007.266301980161 3006.700046740064
3006.1348407605383 3005.5784810027235 3004.98192385602 3004.4065914605812 3003.833067295151 3003.338651271685
3002.945001280539 3002.6718369817313 3002.440720478205 3002.2488640560036 3002.129446978688 3002.0337810732885
3001.9484257631143 3001.8549163701864 3001.79396223132 3001.765834504294 3001.7570522955025 3001.7243353713498
3001.660285212706 3001.57275741207 3001.441502960767 3001.275228739525 3001.086431043717 3000.875074772635
3000.643429796154 3000.396713353474 3000.1492043924527 2999.9120164241035 2999.7110462006226 2999.5104573539993
2999.313485769039 2999.1182000752583 2998.9180995860715 2998.7166064961407 2998.5037312919344 2998.3141994136927
2998.1518423336875 2997.9983505784876 2997.8364408655525 2997.687367310575 2997.55462735149 2997.4192341791154
2997.2936262233748 2997.1901671639525 2997.108448852177 2997.0327283256684 2996.964839448901 2996.9327857105077
2996.914389943748 2996.9115785106906 2996.9144407560484
```

Which then graphed the below information:


What Am I Doing?
Stonehill Sketchbook
Stephanie Veto Photography
Three Pups, One City
The Stearns Fa
gan's Blog
Helen McCookerybook
She's mad but she's magic. There's no lie in her fire.
Incarnate Green
a duchess nonethelesss
hmmhannahmary
headphonehead
CLOUDBUSTING
Happy Accidents
Angie Dynamo
sweeping the kitchen
The Themes of My Life
kaleidoscopekanvas-KK
Rants from the Pants
mattgarman
Words
Nothing But Ordinary Glances At Extraordinary Things
FACSO S2 Blog sessions
TheNorthernGirl Games
The Nosebleed Section
Unexpectedly Bart (King!)
Primitive Offerings
She May Be Naked
Skywriting
Friday Night Dream
SEVEN1878
Kid F
fractalpress.gr
The Cheat Codes For Drugs
hello my name is justin.
New Amusements
the traveling neighborhood
A Music History by Wayne R. Flower
isyeli's
My Name Is Blue Canary
Luke And The Real Blog
F-Measure
the fast break of champions
Music-Drop Magazine
Riley Haas' blog
One Stunning Single Egg
www.doginasweater.com Live Show Review Archive
PSY LAB
Pithy Title Here
ELLIA TOWNSEND A2
The Slow Music Movement
Web Science and Digital Libraries Research Group
Hip In Detroit
INDIEohren.!
ORGANMYTH
The Run-Out Groove
KiDCHAIR
Jasmine Hodge
The Fleshy Fresh
Wyoming Beat
i'm in too truthful a mood
A to Zappa - Song of the day
Fran Brighton
P e w t e r & P u d d l e s
Skiptrack music
Chemical Robert!
Ringtone Lirik
The World's First Internet Baby
Did Not Chart
Spinitron Charts
macthemost
Yestermorrow
Too Poppy
MPC
The Campus Buzz on WSOU
Captain Panda's Local & Independent Music Showcase
simone goes
Radiohead Bootlegs
Encore
Stereo Pills
Green Eggs and Ham Mondays 8-10am
Radio Rithard's Folkways
Faland'hoje
Soundcheck
Paulina Gamero. Media Studies A2
MAGGOT CAVIAR
Lost in the Shuffle
Out of my Mind
Bleak Bliss
SEM REGRAS
persona mia
Stories From the City, Stories From the Sea
Spotirama
Who needs a TV?
You might feel the same
Dust and Water Studios
GLI Press
earenjoy
Indie, Rock And Other Great Music

5 Problem 5

===
========The questions below is for 3 points extra credit=========== ===========================

5.1 Question 5

5. Re-run question 2, but this time with proper TFIDF calculations instead of the hack discussed on slide 7 (p. 32). Use the same 1000 words, but this time replace their frequency count with TFIDF scores as computed in assignment #3. Document the code, techniques, methods, etc. used to generate these TFIDF values. Upload the new data file to github.

Compare and contrast the resulting dendrogram with the dendrogram from question #2.

Note: ideally you would not reuse the same 1000 terms and instead come up with TFIDF scores for all the terms and then choose the top 1000 from that list, but I'm trying to limit the amount of work necessary.

5.2 Answer 5

6 Problem 6

===
========The questions below is for 5 points extra credit=========== ===========================

6.1 Question 6

6. Re-run questions 14, but this time instead of using the 98 ”random” blogs, use 98 blogs that should be ”similar” to:

http://fmeasure.blogspot.com/
http://wsdl.blogspot.com/

Choose approximately equal numbers for both blog sets (it doesn’t have to be a perfect 4949 split, but it should be close). Explain in detail your strategy for locating these blogs.

Compare and contrast the results from the 98 ”random” blogs and the 98 ”targeted” blogs.

6.2 Answer 6