Assignment

Ryan Dill
CS 432
April 25, 2018

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.

For this assignment I created assignment7_1.py to find a list of 100 blogs. Here is a snippet of that code.

```
import urllib.request
from time import sleep
x = 0

for x in range (0, 100):
    try:
        url = 'https://www.blogger.com/next-blog?
navBar=true&blogID=953024975153422094'
        response = urllib.request.urlopen(url)


        with open("blogs.txt", "a") as text_file:
            print(response.geturl(), file=text_file)
        sleep(2)

    except:
        x = x - 1
        pass
```

Next I used assignment7_2.py to clean and prep the blog list for later use. Then I used assignment7_3.py to create a list of words and blogs. Here is the list of words:

```
thoughts  ideas  bank   progress  board  easily selection  government experts
   across teachers  evening   form  interesting  matter difference whom   partner
   journey   production universities  approach  helps  tests  busy  november
   writing   americans allow  east  desire name  hospital  o  america   chance
   april  regular   spread original  stuff  happens   visited   rest  chapter
   practice  changed   join   effective child  track  rights taught feelings
   playing   hear  internal  v  exactly   community seat   purchase  hall
   associated club  built  responsible  wait   rules  sense  gave  average
   balance   places tell  presence  moment came  overall   season thanks realize
   state  f  example   staff  somehow   designed  attend invited   india  quality
   various   pre actually  birthday  master near   technical central   usa handle
```

either positive allowed fell smile k mention offering trip fine
closer decided groups comes recently created liked released continued
giving especially terms meeting providing names case online half weird
thus money hotel girls glad common usual south involved obviously deal
tuesday program trying view image d picked desk towards plans
listening guide members beauty month wide documents several boy mass
nature write anyone team particular cultural states international l window
responsibility pretty yourself others theme beyond floor opportunity
driving call proud joy shot admit action worry council author
applications location possible sat events executive rate minute maybe
institute directly rule ones she blog met largest teaching academic
friday gives provided tv information beginning push talked assignment
capacity door respect cold become basically colleges super links mr
guest j written road ahead try task middle reflect highly therefore
readers fight woman false north ll continue king live support
customer real sister fear complete however sounds hell foreign
christmas story cup advance bill touch final amp early established
comments couldn shop popular influence en born related feels june
risk outside learn dad major hours cause problem conflict planning
district cool choice february sunday stated heard eating carry missing
mind marketing herself cover ever expected load administration
visiting although competition spot beach seeing raised line york didn
path features experience parts x rd afraid california considered platform
stop british fact quite opportunities appear agree pm subject primary
father lived wonder river p white conditions successful cost force report
aren wondering self safe eight seriously font additional indeed physical
stand sharing posts conversation doesn tomorrow teacher offer letter
certain note dropped close th clearly choose position taking
requirements says men losing remains believe completed individual usually
series starting favorite alone ask death record distance ok seven scale
song eye enjoyed contact prepared travel building main further
weather july theory source mso test mean worth purpose naturally
coffee determine announced rather interview calls night skills tstyle police
remove single funds save assistant walk united feeling else wouldn exam
ages ways worried unless campus space mostly pressure ride training
gt fellow spoke works reasons short activity seems copy financial
courses means studying provides list session b institutions
ultimately student advantage onto developed soon except book studies
tech earth enjoy leaving creating requires org age potential face
water paid nearly minor sports model speed adding e dance sorry required
mother receive together remain director difficult phone margin examples
morning fill design military stress issue aspect pdf ago land confidence
considering okay loss deep video rates release somewhere stories
calling resources young advanced law discuss challenge thousands
continues race term toward graduate economic left legal whatever
candidates based available gets direction count teams surprised practices
share movement personally speech lives period might hope products
certainly leadership despite problems january rise setting weight
variety greater center answer experienced placed option whose simple
section business articles saying advice decisions amazing hair st late
question latin title held dream papers actions phase players values
forced happened began hour background resolution committee plan music
words parents yesterday watched makes para effect room entire focus
talking price wonderful run hit evidence truth below serve power faculty
positions willing notes buying accepted honest projects hansi table
infrastructure bigger current policy images random knew managed forward
growing small everyone colleagues ability moving john presentation
employees social clean anywhere anymore semester apart solutions limited
expect wear spend internet found hot actual summer prefer modern grade

august sit order four healthy guys women inside football sets yes
enter g costs according wake party told necessary visit older mission
lead rich reality female sell books discussed natural measures
december useful journal slow became link green museum second degree
analysis nor won national association access informed correct remember
bit sent thursday items increasing previously assignments son public low ten
account red buildings basis programs effort generation show grow aware
benefits spent h round history tired wall happy american eat forget
develop improve pt attending pictures among encourage connection general
dr accessible environmental bottom require earlier changes least piece
de higher third tried version easier immediately myself becoming comment
raise machine afternoon prior ensure options die side absolutely florida
levels currently walked cards planned creative talk decide compared seen
office fun catch relationships facilities screen published determined fish sad
families understand communication secondary records former notice service
level field haven reading shopping techniques key growth includes complex
completely wife perfect efforts ended meet print saw review parent
capital strong trump vote personal hands connect whole added hand
teach instead meaning development act shared cases street wrong recent
technology wants learning poor country concept arrived looked march
check turned relationship surprise brother knows energy address
happening girl included u needed mobile interest https himself miss
served looks attention conference voice activities event seem company
selling engineering earn ms global jobs passion awareness interested
kids facebook themselves lunch wasn taken aside specific budget needs
proper update perhaps size putting happiness cute avoid group beat
fall often months speaking wanted organizations living facts security yeah
lots opened presented six videos five please shall within store food uk
town wrote article heart figure health somewhat buy standing pages
perspective computer c clear la huge less traditional society lack
http finding traffic art air per imagine twitter experiences press
mom africa fan sun miles anything developing west industry classes etc
regarding direct big increase opinion dark n senior along london previous
appreciate percent pay private september product head ve airport
weekend code participants rare bought body already news behind fast
helpful department someone boys using longer blue covered systems
otherwise education edu running forms park passed gain bad text normal
front october r picture open stuck game hoping comfortable recall
suggest kind message fourth application president applied speak
types appeared area oh sort everyday co bus spending straight freedom
learned watch word black paper countries manager organized discussion
focused house web loved type population justice sessions read hate
similar reports followed washington management itself solution ground
located moved play professional attempt break leader till quiet
situation answers city kept customers true return idea historical
culture accept understanding called seemed beautiful pass reason collection
doubt changing importance sitting due drink listen faces photos individuals
welcome following ascii range success attended movie research entered
twice unfortunately reach friend started paying signs professionals non
anyway large media happen yet attitude re software youtube plus brought
becomes david cut resource multiple fit above wedding numbers friends
exercise brain companies offered control apply serious materials isn
finally relevant children offers box method awesome guess professor
million played lake telling services site biggest updated bed
asking whether knowledge ran specifically pick stopped extremely
challenges filled present stay ready educational identify leave friendly
areas graduation indian benefit probably childhood result court special
bring content addition annual questions step arts walking photo
finished sometimes via reached click add sign extra goes topics care

consider job english mine helping cross prepare doctor page meant received against light mark provide quick ii looking man material science god monday trust car leaders devices total takes chose human coming bar value tips worse spring lines memory likely build fire methods points almost sector amount net thinking hold create under upon www war submitted system moments person achieve feedback role occasion guy explore dead returned shows context secretary though statement email network famous named nice feet station data political issues generally project thank lost search nothing sweet campaign member mentioned style display including noticed peace sleep process quickly daily attack demand smart schedule led local wish topic eventually digital cannot character truly honestly restaurant match felt supposed easy train color lower weeks minutes performance response standard excited appropriate saturday funny winter went couple holding top movies feature sound officer active watching far sport maintain move goals schools enough results date number none dreams win charge knowing google language turn waiting drive medical safety birth score curriculum trade asked eyes finish listed decision worked realized website wednesday tax leading critical allows follow basic organization goal ourselves impact official study career simply agency begin include library throughout exams definitely stage particularly showed registration fair unique games market nation dinner send helped gone lose recognize besides known environment keeping details fully

And here is the list of blogs:

GAME CHANGE HERS
IPS BUSINESS SCHOOL
Wear Your Heart on Your Shoulder, Be Almostwiser
My Writing
my artwork - aqua vitae
Driver's Ed for Yeshiva Students
The Scribbles of a Voyager
Sunday Brunch .
Engineers, Graduates & Post Graduates Seeking 1st Job
My Junior Year
JayMykels World
LAW&@UCU
FIELD OF VIEW
Shoalanda Speaks
Karis: (noun) greek; grace...
Alex Broderick
Inglesagil Looks @ EDITORIALS/VIDEO/MUSICA/DANZA/ART/TEXTILES/ EXPOVIAL / FERROVIAL 2012
Ferrum College Football Report
Post Secondary Education Blog
Artricia's World: A Little bit of everything!
Pier Points
Internet Marketing and Publishing
GTC Alumni Association
ISU Ed. Leadershop
Love and Rockets
BUIKETALKS
Freedom of Choice
M J C - Empowering the Girl Child
Online Homeland security
Department of Communication Studies
Cal Poly Dolly

nanopolitan
Kirie-lea Cussen
Noobie's Adventures In K-12 Computer Science Teaching
Signals Processed
Jacob G's Blog
Nathaniel Rey D. Giron
Young Journalists & Writers
F-Measure
Eddy County Ag News
Sarah the Explorer
Engineers Without Borders Bangladesh
Belog Makcik Anne
Kevin In Beijing
The HBCU Blog
THE CRITIC OUTLET
MEGAEVALUACIÓN
Florida Coal Cracker Chronicles
Free Nampeyo
Think Create Discover
Govt. Jobs 4 U
The 2015 blogger
Thinking Is Dangerous
..the moving finger writes...
CLASS BIAS AND RANDOM THINGS LAW REVIEW
Alex the Entrepreneur
Textual Studies, 1500-1800
LZ SISTERS
Cowboy's Gal's Blog
AdilaLieow's
Between the lines
University of Waterloo BCS Student Blog
M.A.JABBAR
Learning + Design
Breaking the Code
B.E/B.Tech Admission through Management Quota/NRI Quota
CSE Finance and Operations Blog
Wiwid's
Life is a Journey
Katherine Taylor 413-230-7477
SurfnPoetry
madsmaddad ramblings
Thought-o-louge : Thoughts-in-words
Times of wIKi
SBCC Hannah
wild throne
Borland Educational News and Views
Talk Time With Tori
Rotten Hub
NIRMAN for Education and the Arts, Varanasi, India
Web Science and Digital Libraries Research Group
National History Day - SF County
Blogger attempt
Otto's Random Thoughts
jenny.riley (@ginnywryly)
Facso 2017
Saniaga Kamnara
CLF's Journey from KL to Tokyo
A Sixties Girl
FAMOUS PAGE

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 one was difficult for me. I found some code that was very useful at https://stackoverflow.com/questions/12520537/python-how-to-make-a-dendrogram-from-a-dataframe. I used assignment7_4.py and assignment7_5.py to create a dendrogram and uploaded it to github.

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?

For this problem I created three separate programs, assignment7_6a – 7_6c. Each handled k=5, 10, and 20 respectively. For k=5 I got 11 iterations, for k = 10 I had 9 iterations, and for k = 20 I got 6 iterations.

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

I uploaded the answer to this as MDS.py.