Matthew Gomez

CS 432

Michael Nelson

Assignment 9

4/21/16

Question 1:

RSS Feed:

http://www2c.cdc.gov/podcasts/createrss.asp?t=r&c=183

URL Link:

http://www2c.cdc.gov/podcasts/rss.asp

Categories:

Zika
Research
Cancer
Ebola
HIV
Outbreak

For question 1, I picked the World News Room Blog from the Center of Disease Control website. I categorized the 100 blogs from the site listed above. I was able to use a feed parser code I found from the *Program Collective Intelligence* textbook and from the direction from my

fellow colleague.

```
def classifyEntries(settings):
    database = FeedDatabase(settings['database'])
    unclassifiedEntries = database.get_unpredicted_entries()
    #for i in unclassifiedEntries:
    #    print(i)
    #print(len(unclassifiedEntries))
    database.close_database()

    classifier = fisherclassifier(getwords)
    classifier.setdb(settings['database'])
    counter = 0
    size = len(unclassifiedEntries)
    results = []
    for entr in unclassifiedEntries:
        a = open('script50.txt','w+')
        for i in results:
            a.write('{0}|{1}\n'.format(i['guid'],i['category']))
        a.close()
        category = classifier.classify(entr['description'])
        #print('{0}|{1}'.format(entr['guid'],category))
        results.append({'guid':entr['guid'],'category':category})

        counter += 1
        sys.stderr.write('...Classified {0} of {1} entries\n'.format(counter,size))
```

Question 2:

```
def extractFeedInformation(settings,categorizedDat=None):
    # Get data from feeds

    data = feedparser.parse('../newsfeeds/CDCNews/CDCnews.xml')
    for key in categorizedDat.keys():
        print (key)
    #print("EXTRACT feed information")
    tmp = {}

    for entry in data.entries:
        #print(entry.title)
        #print (entry.summary)
    #    print(entry.link)
    #    guid = entry.link
        #print(categorizedDat[entry.guid])
        categorizedDat[entry.link]['title'] = entry.title #unicodedata.normalize('NFKD', entry.title).encode('ascii','ignore')
        categorizedDat[entry.link]['description'] = entry.summary#unicodedata.normalize('NFKD',entry.description).encode('ascii','ignore')
        categorizedDat[entry.link]['guid'] = entry.guid
        tmp[entry.link] = {'title':entry.title,'description':entry.summary,'guid':entry.link}

    #for entr in tmp.keys():
        #print(categorizedDat[entr])
        #print(str(entr))
        #pass
        #v = entr.replace(';','')
        #print(v)
    v1 = sorted(tmp.keys())
    v2 = sorted(categorizedDat.keys())
    #print(len(v1))
    #print(len(v2))
    print(len(data.entries))
    #for i in range(len(v1)):
    #    print('{0}\n{1}\n'.format(v1[i],v2[i]))
        #print(categorizedDat[v])
if __name__ == '__main__':
    settings = getSettings()

    trainingDataFile = 'script50.txt'

    allCategories = 'CDCfeed.txt'

    # Put all the feeds and their categories in the database
    allCats = readCategorizedData(allCategories)

    #trainingDat = readCategorizedData(trainingDataFile)


    extractFeedInformation(settings,categorizedDat=allCats)


    for key in trainingDat:
        trainingDat[key]['description'] = allCats[key]['description']
        extractFeedInformation(settings,categorizedDat=trainingDat)

    sys.stderr.write('Uploading Entries to database...\n')

    loadFeedInformationToDatabase(settings,allCats)

    sys.stderr.write('Training Classifier...\n')
    train_classifier(settings,trainingDat)
def loadFeedInformationToDatabase(settings,allCats):
    # Put feed information in database

    database = FeedDatabase('database.db')
    counter = 0
    size = len(allCats)
```

Title	Predicted Category	Actual Category	Fisher Probability
Half of those who need them not taking cholesterol-lowering medications	Zika	Research	0.286453712
New CDC Vital Signs: E-cigarette Ads and Youth	Cancer	Cancer	0.528647951
Transcript for CDC Telebriefing: Zika Summit Press Conference	Zika	Zika	0.567895642
Updated Guidelines for Healthcare Providers Caring for Infants or Children with Possible Zika Virus Infection	Research	Zika	0.145654135
Updated Guidelines for Healthcare Providers Caring	Cancer	Zika	0.015468465

for Infants or Children with Possible Zika Virus Infection - Media Statement			
The Centers for Disease Control and Prevention and the National Institute of Allergy and Infectious Diseases will discuss the latest on Zika virus..	Research	Zika	0.289846123
CDC Adds Fiji to Interim Travel Guidance Related to Zika Virus - Media Statement	Cancer	Zika	0.035794513
CDC adds 1 destination to interim travel guidance related to Zika virus - Media Statement	Outbreak	Zika	0.256465431
Drug overdose deaths hit record numbers in 2014 - Press Release	Zika	Outbreak	0.158985456
New CDC Laboratory Test for Zika Virus Authorized for Emergency Use by FDA - Media Statement	Zika	Zika	0.652646845
CDC adds Cuba to interim travel guidance related to Zika virus - Media Statement	Cancer	Zika	0.025846941
CDC Telebriefing: New Vital Signs Report When is a daily pill the right option to prevent HIV? - Media Advisory	Zika	HIV	0.425341255
CDC responds to broad challenges facing US cancer survivors - Press Release	Zika	Cancer	0.456876355
CDC issues interim travel guidance related to Zika virus for 14 Countries and Territories in Central and South America and the Caribbean - Media Statement	Zika	Zika	0.548632149
E-cigarette ads reach nearly 7 in 10 middle and high-school students - Press Release	Zika	Cancer	0.012678239
Updates on Zika response efforts - Media Advisory	Zika	Zika	0.589643252
CDC Museum to Host Places & Spaces: Mapping Science Exhibition - Media Statement	Outbreak	Research	0.405546641
Enhanced Entry Airport Screening and Routing for Ebola to End for Travelers from Guinea to the United States - Media Statement	Zika	Ebola	0.325467946
CDC adds 2 destinations to interim travel guidance related to Zika virus - Media Statement	Ebola	Zika	0.345647412

New 'Parents for Healthy Schools' Website - Media Advisory	Zika	Outbreak	0.435782512
CDC adds 2 destinations to interim travel guidance related to Zika virus - Media Statement	Research	Zika	0.156486456
Transcript for CDC Telebriefing: New Vital Signs Report - Why are millions of US women at risk of alcohol-exposed pregnancies? - Transcript	Outbreak	Outbreak	0.784568134
ATSDR and CDC Analysis Finds Possible Health Effects Associated with Formaldehyde in Select Laminate Flooring - Media Advisory	Outbreak	Research	0.663127985
CDC Telebriefing: Updates on CDC recommendations related to Zika virus - Media Advisory	Ebola	Zika	0.288476432
CDC Issues Updated Zika Recommendations: Timing of Pregnancy after Zika Exposure, Prevention of Sexual Transmission, Considerations for Reducing Unintended Pregnancy in Areas with Zika Transmission - Media Statement	Research	Zika	0.584667924
ATSDR and CDC Analysis Finds Possible Health Effects Associated with Formaldehyde in Select Laminate Flooring - Media Statement	Zika	Research	0.448964655
Daily Pill Prevents HIV – Reaching People Who Could Benefit From PrEP - Digital Press Kit	HIV	HIV	0.841654222
CDC Year in Review: What’s Next? - Press Release	Outbreak	Research	0.575641999
2015: What Kept Us Up At Night and What Will Keep Us Busy in 2016 - Digital Press Kit	Research	Research	0.785648146
CDC Releases Guideline for Prescribing Opioids for Chronic Pain - Digital Press Kit	Outbreak	Outbreak	0.486465467
CDC Releases Guideline for Prescribing Opioids for Chronic Pain - Press Release	Zika	Research	0.856421385
CDC Telebriefing: Updates on CDC’s Zika virus response efforts - Media Advisory	Outbreak	Zika	0.452789665

New CDC estimates underscore the need to increase awareness of a daily pill that can prevent HIV infection - Press Release	Outbreak	HIV	0.486469465
Zika Action Plan Summit - Media Advisory	HIV	Zika	0.568799985
Superbugs threaten hospital patients - Digital Press Kit	Ebola	Research	0.456874694
CDC Director travels to Puerto Rico to assess Zika response - Media Advisory	Zika	Zika	0.745863413
National Zika Summit Focused on Coordinated U.S. Response - Press Release	HIV	Zika	0.048968446
CDC adds 2 destinations to interim travel guidance related to Zika virus - Media Statement	Research	Zika	0.568998743
CDC To Play Key Role in National Multidrug-Resistant TB Plan - Media Statement	HIV	Research	0.627781546
CDC/ATSDR Revises Report of Possible Health Effects Associated with Formaldehyde in Select Laminate Flooring	Cancer	Research	0.889651532
Changes in the CDC/ATSDR Formaldehyde in Laminate Flooring Report - Media Statement	Zika	Research	0.756412354
Powerful new ads mark the 5th year of the successful “Tips From Former Smokers” campaign - Press Release	HIV	Cancer	0.017554134
PulseNet saves lives and money by reducing foodborne illness - Press Release	Ebola	Outbreak	0.456876233
PulseNet, a national network of public health laboratories, prevents an estimated 270,000 cases of food poisoning and saves half a billion dollars every year.	Research	Outbreak	0.464488656
Transcript for CDC Telebriefing: Daily Pill Prevents HIV - Transcript	Ebola	HIV	0.567751432
CDC issues Interim Guidelines for Preventing Sexual Transmission of Zika Virus and Updated Interim Guidelines for Health Care Providers Caring for Pregnant Women and Women of Reproductive Age with Possible Zika Virus Exposure - Media Statement	Ebola	Zika	0.533765517

First-of-its-Kind PSA Campaign Targets the 86 Million American Adults with Prediabetes - Press Release	Zika	Research	0.760714634
More than 120 Partners Join CDC to Fight Antibiotic Resistance - Press Release	Ebola	Research	0.577333549
Get Smart About Antibiotics Week 2015 - Digital Press Kit	Research	Research	0.756811236

For question 3, I ran the calculations necessary for the assignment.

correct	0
Nprediction	45
Ncategory	0
precision	0
recall	0
Fmeasure	0

Zika

correct	0
Nprediction	45
Ncategory	0
precision	0
recall	0
Fmeasure	0

Cancer

correct	0
Nprediction	45
Ncategory	0
precision	2.5
recall	0
Fmeasure	0

Ebola

correct	0
Nprediction	45
Ncategory	0

precision	0
recall	0
Fmeasure	0
HIV	
correct	0
Nprediction	45
Ncategory	3
precision	0
recall	0
Fmeasure	0
Outbreak	