Assignment Nine

Jacob Berlin

CS432 – Spring 2016

1. 1. Choose a blog or a newsfeed (or something similar with an Atom or RSS feed). Every student should do a unique feed, so please "claim" the feed on the class email list (first come, first served). It should be on a topic or topics of which you are qualified to provide classification training data. Find something with at least 100 entries (or items if RSS).

Create between four and eight different categories for the entries in the feed:

examples:

work, class, family, news, deals

liberal, conservative, moderate, libertarian

sports, local, financial, national, international, entertainment

metal, electronic, ambient, folk, hip-hop, pop

Download and process the pages of the feed as per the week 12 class slides.

Be sure to upload the raw data (Atom or RSS) to your github account.

Work:

To start this assignment, I chose the blog 'http://gigabytedaily.blogspot.com/' due to my love for building and overclocking computers. This worked for me in that the feed had 150 entries (I stopped looking at that point) and was able to be parsed rather easily. To achieve this task, I modified the program from assignment eight (generatefeedvector.py, PCI) to show and save all of the data that we were looking for.

```python
# -*- coding: utf-8 -*-
import feedparser
import collections
import re
import feedfilter as f
entryFile = open('entryfile.txt', 'w')
outputFile = open('output.txt', 'w')
def getwords(html):
  text = re.compile(r'<[^>]+>').sub('', html)
  words = re.compile(r'[^A-Z^a-z]+').split(text)
  return [word.lower() for word in words if word]


def getwordcounts(url):
  # Parse the feed
  d=feedparser.parse(url)
  wc={}
  for e in d.entries:
    if 'summary' in e: summary=e.summary
    else: summary=e.description
    outputFile.write(e.title.replace(u"\u2019", "'").replace(u"\u2026", "...").replace(u"\u201c", "'").replace(u"\u201d", "'").replace(u"\u2122", "tm").replace(u"\u2013", "-"))
    outputFile.write("\n")
    entryFile.write(str(e))
    entryFile.write("\n")
    # Extract a list of words
    words=getwords(e.title+' '+summary)
    for word in words:
      wc.setdefault(word,0)
      wc[word]+=1

  if 'title' not in d.feed:
    print 'Invalid url', url
    return 'bogus data', wc
  return d.feed.title, wc
```

```python
def main():

    # XXX: break this up into smaller funtions, write tests for them

    apcount = collections.defaultdict(int)
    wordcounts = {}
    feedlist = open('feedlist.txt').readlines()
    for url in feedlist:
        title, wc = getwordcounts(url)
        wordcounts[title] = wc
        for word, count in wc.iteritems():
            if count > 1:
                apcount[word] += 1

    wordlist = []
    for w, bc in apcount.iteritems():
        frac = float(bc)/len(feedlist)
        if 0.1 < frac < 0.5: wordlist.append(w)

    out = file('blogdata.txt', 'w')
    out.write('Blog')
    for w in wordlist: out.write('\t' + w)
    out.write('\n')
    for blogname, counts in wordcounts.iteritems():
        blog = blogname.encode('UTF-8')
        out.write(blog)
        for w in wordlist:
            if w in counts: out.write('\t%d' % counts[w])
            else: out.write('\t0')
        out.write('\n')

if __name__ == '__main__':
    main()
    outputFile.close()
    entryFile.close()
```

After running this program I received both a list of the titles of each of the 100 entries (limited in URL - http://gigabytedaily.blogspot.com/feeds/posts/default?max-results=100) and the entire feed data. The data is stored in 'entryfile.txt'.

```
Which VR Headset is Right For You?
Here's a Guide to GIGABYTE's 3D OSD
It's Time for Some PC Spring Cleaning
Win a Trip to Taipei With GIGABYTE
SlipperySkip latest build: SkyVue
Here Are Your Winter OC Challenge Winners!
Microsoft Really Wants You to Upgrade to Windows 10
Check Out This Interview With GIGABYTE's Very Own Sofos1990
Winter is Coming: Keep Warm By Overclocking
Time to Upgrade: Microsoft Cuts Support For New CPUs on Windows 7/8.1
SlipperySkip new project (In progress)
GIGABYTE Z170X-Gaming G1 is Tom's Hardware 2015 Editor's Choice
TeamAU and GIGABYTE Break World Records!
The GIGABYTE Z170X-Gaming G1 Gets Reviewed by The Overclocker
Anandtech Recommends The GIGABYTE GA-Z170X-Gaming G1!
Vote for GIGABYTE at the European Hardware Awards 2015/2016
Desktop Power in a Console Frame: Check out TheOriginalQ's Awesome Mini-ITX Build
GIGABYTE Leads The Industry in Thunderbolt 3 Support
Check Out Our PC Building Guide And Don't Forget About Tomorrow's OC Livestream!
Learn How To Build a PC and Overclock with GIGABYTE
TweakTown Gives The GIGABYTE Z170X-UD5 TH a Best Features Award
Don't Miss Dinos22 on The Next OC Show!
The magic of building your own custom PC
GIGABYTE's Latest OC Contest Let's You Overclock on Your Favorite Chipset!
Gotta Have SLI On A Budget? Check Out This Review of The GIGABYTE Z170XP-SLI
8Pack Shatters New Records On A X99 Champion
TweakTown Reviews the GIGABYTE Z170X-SOC FORCE
The Innovative Tech Behind the Killer Ethernet Controller
Don't Miss Out on Your Chance to Go to Blizzcon!
GIGABYTE Tweak Launcher for Z170 chipset
More Recognition of GIGABYTE's Overclocking Leadership
Another Record Broken on a GIGABYTE SOC Champion
Add Some Color and Illumination to Your System Build With GIGABYTE
Let's Build a CustoMac With GIGABYTE Parts
Here's Why It's Time to Upgrade Your Gaming System
Having Trouble Installing Windows 7 By USB With Your GIGABYTE 100 Series Motherboard?
Here Are Your Winners of the GIGABYTE Z97 Last Hurrah Contest!
Win a Trip to BlizzCon From GIGABYTE and Blizzard
```

GIGABYTE OC Guru Dinos22 is Livestreaming Now!
Congrats to all the Winners of the GIGABYTE Z97 Last Hurrah Challenge!
GIGABYTE OC guru from Australia does a 1v1 interview with OCTV, an interesting watch!
The Awards Keep Coming in For GIGABYTE
Turn Your GIGABYTE BRIX-s Into a CustoMac!
TweakTown Gives The Z170X-Gaming G1 Top Marks!
GIGABYTE Z170X Gaming G1 Custom Build
GIGABYTE's 100 Series Motherboards Come With Intel's USB3.1 Controller
Win Two Z170X-SOC Force Motherboards With GIGABYTE and HWBOT
Here's Why GIGABYTE's 100 Series Motherboard is a KILLER
Want to Learn How to Overclock? Check Out These Videos From HWBOT
GIGABYTE BRIX S BXi5H-5200 Gets a Recommended Award From Legit Reviews
Check Out This Beast of a PC From NCIX Powered by GIGABYTE
IEM Season 10 Kicks Off in Shenzhen
We're #1! Consumers Vote GIGABYTE Motherboards To The Top on Ranker
Give GIGABYTE's Z97 a Last Hurrah on HWBOT
Meet LEENA: A Robot Powered by a GIGABYTE BRIX
GIGABYTE X99 Case mode Made By Kier
GIGABYTE Z97X-SOC Mobo Featured in Project Modern Orange
GIGABYTE X99 Champion Challenge Winners!
GIGABYTE X99 bios and software updates (Windows XP)
GIGABYTE New 100 Series Overview by Sin0822
New day, new XTU world record!
GIGABYTE X99 SOC Champion voted the best overclocking board by European Hardware Award
Gigabyte Overclocking Crew Set New XTU World Record
GIGABYTE X99-Gaming 5P Receives the Editor's Choice Award from TT
G.Skill announces DDR4 3666MHz kit, only capable of running on GIGABYTE X99 SOC Champi
GIGABYTE X99-SOC Champion... is still the Champion!
Katowice 2015 | Official Aftermovie
'All you need are the GIGABYTE boards & an Australian passport'
8pack and der8auer light up the record books with X99 SOC Champion and TitanX
Win a G1tm Gaming Motherboard with The Summit 3 by GIGABYTE Giveaway!
eTeknix reviews the GIGABYTE X99 UD4P, make sure you check it out!
New GIGABYTE OC Challenge Coming May 1st on HWBOT.org
Interesting things on the net today
TweakTown gives GIGABYTE BRIX BXi7H-5500 an EDITOR'S CHOICE AWARD
GIGABYTE X99 SOC Champion overclocking guide for uncore and memory overclocking

Chips, Buses and Compatibility!
Fastest dual core in the world at 7862.83MHz
Windows 10 Personal and Secure Solution
The Overclocker magazine issue 33 is out, GIGABYTE X99 SOC Champion makes a mark!
[HWBOT] GIGABYTE X99-SOC Champion Dominates X99 2D Multi-threaded Benchmarks with 1 World Record and 5 Global First Place
GIGABYTE BRIX i7 5500U unboxing, RAM/SSD & Windows 10 installation guide
It's Time to Overclock with GIGABYTE's BIG XTU Challenge
VRWORLD: INTEL PUSHES PAST 10NM TO A POST-SILICON ERA!
Case Mod: Terminator Velox Project
GIGABYTE Z97X Gaming G1 Black selected as 'Motherboard of the Year' by PC PowerPlay Magazine
GIGABYTE Brix BXBT-2807 Compact PC Unboxing and Linux Installation
Steponz from USA strikes again, takes out all major single GPU scores!
Steponz first overclocker to reach 10,000+ points with 1x GPU in Fire Strike Extreme!
How to insulate a motherboard for subzero cooling, guide inside!
OCAU overclocking competition with a twist, how low can you score in 3DMARK!
GIGABYTE X99 SOC Champion breaks XTU world record and highest DDR4 memory frequency
Top 10 Best-Selling Games of 2014
CES 2015 Coverage
Steam Daily Deal: 3DMARK 80% OFF
GIGABYTE X99 SOC Champion Debuts at CES 2015 with One More WR Under its Belt
DDR4 Memory Clock Reaches New High, GIGABYTE X99-SOC Champion rocks!
HWBOT Country Cup is over, Australia wins again!
Hardware Asylum podcast on evolution of overclocking motherboards
X99-SOC Champion - Let it Rain Records

After looking at all of the articles, I decided that it would be best to make my categories:

1. HowTo – How to guides, articles on building and fixing.

2. Competitions – entries on previous, occurring, or future competitions and outcomes.

3. Sales – Entries attempting to sell a product.

4. Information – General information.

2. Manually classify the first 50 entries, and then classify (using the fisher classifier) the remaining 50 entries.

Create a table with the title, predicted category, actual category, and cprob() and fisherprob() for the actual category.

To begin this problem, I modified PCI's 'feedfilter.py' file to let the user input the first 50 categories and then let the program learn from it and make guesses for the next 50. The output from compilation is shown in 'computing.txt'. With help from Ryan Condotta's variation of PCI's 'docclass.py', I was able to run this smoothly and efficiently. It also helped with the compilation of the fisher and cprob scores. I ended up with the table below:

Title	Predicted	Actual
Which VR Headset is Right For You?	x	Informational
Here's a Guide to GIGABYTE's 3D OSD	x	HowTo
It's Time for Some PC Spring Cleaning	x	Informational
Win a Trip to Taipei With GIGABYTE	x	Sales
SlipperySkip latest build: SkyVue	x	Informational
Here Are Your Winter OC Challenge Winners!	x	Competitions
Microsoft Really Wants You to Upgrade to Windows 10	x	HowTo
Check Out This Interview With GIGABYTE's Very Own Sofos1990	x	Informational
Winter is Coming: Keep Warm By Overclocking	x	Sales
Time to Upgrade: Microsoft Cuts Support For New CPUs on Windows 7/8.1	x	Sales
SlipperySkip new project (In progress)	x	Informational
GIGABYTE Z170X-Gaming G1 is Tom's Hardware 2015 Editor's Choice	x	Sales
TeamAU and GIGABYTE Break World Records!	x	Competitions
The GIGABYTE Z170X-Gaming G1 Gets Reviewed by The Overclocker	x	Informational
Anandtech Recommends The GIGABYTE GA-Z170X-Gaming G1!	x	Sales
Vote for GIGABYTE at the European Hardware Awards 2015/2016	x	Sales
Desktop Power in a Console Frame: Check out TheOriginalQ's Awesome Mini-ITX Build	x	Informational
GIGABYTE Leads The Industry in Thunderbolt 3 Support	x	Informational
Check Out Our PC Building Guide And Don't Forget About Tomorrow's OC Livestream!	x	Competitions
Learn How To Build a PC and Overclock with GIGABYTE	x	HowTo
TweakTown Gives The GIGABYTE Z170X-UD5 TH a Best Features Award	x	Competitions
Don't Miss Dinos22 on The Next OC Show!	x	Informational
The magic of building your own custom PC	x	Sales
GIGABYTE's Latest OC Contest Let's You Overclock on Your Favorite Chipset!	x	Informational
Gotta Have SLI On A Budget? Check Out This Review of The GIGABYTE Z170XP-SLI	x	Sales
8Pack Shatters New Records On A X99 Champion	x	Competitions
TweakTown Reviews the GIGABYTE Z170X-SOC FORCE	x	Informational
The Innovative Tech Behind the Killer Ethernet Controller	x	Informational
Don't Miss Out on Your Chance to Go to Blizzcon!	x	Sales
GIGABYTE Tweak Launcher for Z170 chipset	x	HowTo
More Recognition of GIGABYTE's Overclocking Leadership	x	Competitions
Another Record Broken on a GIGABYTE SOC Champion	x	Competitions

Title		
Another Record Broken on a GIGABYTE SOC Champion	x	Competitions
Add Some Color and Illumination to Your System Build With GIGABYTE	x	HowTo
Let's Build a CustoMac With GIGABYTE Parts	x	Sales
Here's Why It's Time to Upgrade Your Gaming System	x	Informational
Having Trouble Installing Windows 7 By USB With Your GIGABYTE 100 Series Motherboard?	x	HowTo
Here Are Your Winners of the GIGABYTE Z97 Last Hurrah Contest!	x	Competitions
Win a Trip to BlizzCon From GIGABYTE and Blizzard	x	Competitions
Want to Learn More About the Killer Wireless AC-1535?	x	Informational
GIGABYTE OC Guru Dinos22 is Livestreaming Now!	x	Informational
Congrats to all the Winners of the GIGABYTE Z97 Last Hurrah Challenge!	x	Competitions
GIGABYTE OC guru from Australia does a 1v1 interview with OCTV, an interesting watch!	x	HowTo
The Awards Keep Coming in For GIGABYTE	x	Competitions
Turn Your GIGABYTE BRIX-s Into a CustoMac!	x	HowTo
TweakTown Gives The Z170X-Gaming G1 Top Marks!	x	Informational
GIGABYTE Z170X Gaming G1 Custom Build	x	Sales
GIGABYTE's 100 Series Motherboards Come With Intel's USB3.1 Controller	x	Informational
Win Two Z170X-SOC Force Motherboards With GIGABYTE and HWBOT	x	Competitions
Here's Why GIGABYTE's 100 Series Motherboard is a KILLER	x	Informational
Want to Learn How to Overclock? Check Out These Videos From HWBOT	x	HowTo
GIGABYTE BRIX S BXi5H-5200 Gets a Recommended Award From Legit Reviews	Informational	Informational
Check Out This Beast of a PC From NCIX Powered by GIGABYTE	Informational	Sales
IEM Season 10 Kicks Off in Shenzhen	Informational	Competitions
We're #1! Consumers Vote GIGABYTE Motherboards To The Top on Ranker	Informational	Informational
Give GIGABYTE's Z97 a Last Hurrah on HWBOT	Competitions	Informational
Meet LEENA: A Robot Powered by a GIGABYTE BRIX	Informational	Informational
GIGABYTE X99 Case mode Made By Kier	Informational	Informational
GIGABYTE Z97X-SOC Mobo Featured in Project Modern Orange	Sales	Sales
GIGABYTE X99 Champion Challenge Winners!	Informational	Competitions
GIGABYTE X99 bios and software updates (Windows XP)	Informational	HowTo
GIGABYTE New 100 Series Overview by Sin0822	Informational	Informational
New day, new XTU world record!	Competitions	Competitions
GIGABYTE X99 SOC Champion voted the best overclocking board by European Hardware Awards	Sales	Competitions
Gigabyte Overclocking Crew Set New XTU World Record	Competitions	Competitions

Title		
'All you need are the GIGABYTE boards & an Australian passport'	Informational	Sales
8pack and der8auer light up the record books with X99 SOC Champion and TitanX	Competitions	Competitions
Win a G1tm Gaming Motherboard with The Summit 3 by GIGABYTE Giveaway!	Sales	Competitions
eTeknix reviews the GIGABYTE X99 UD4P, make sure you check it out!	Sales	Sales
New GIGABYTE OC Challenge Coming May 1st on HWBOT.org	Informational	Sales
Interesting things on the net today	Sales	Sales
TweakTown gives GIGABYTE BRIX BXi7H-5500 an EDITOR'S CHOICE AWARD	Informational	Informational
GIGABYTE X99 SOC Champion overclocking guide for uncore and memory overclocking	Informational	Informational
Chips, Buses and Compatibility!	Informational	Informational
Fastest dual core in the world at 7862.83MHz	Sales	Sales
Windows 10 Personal and Secure Solution	Informational	HowTo
The Overclocker magazine issue 33 is out, GIGABYTE X99 SOC Champion makes a mark!	Informational	Informational
[HWBOT] GIGABYTE X99-SOC Champion Dominates X99 2D Multi-threaded Benchmarks with 1 World Record and 5 Global First Place	Sales	Competitions
GIGABYTE BRIX i7 5500U unboxing, RAM/SSD & Windows 10 installation guide	HowTo	HowTo
It's Time to Overclock with GIGABYTE's BIG XTU Challenge	Sales	Competitions
VRWORLD: INTEL PUSHES PAST 10NM TO A POST-SILICON ERA!	Sales	Informational
Case Mod: Terminator Velox Project	Informational	Informational
GIGABYTE Z97X Gaming G1 Black selected as 'Motherboard of the Year' by PC PowerPlay Magazine	Sales	Sales
GIGABYTE Brix BXBT-2807 Compact PC Unboxing and Linux Installation	Sales	Sales
Steponz from USA strikes again, takes out all major single GPU scores!	Competitions	Competitions
Steponz first overclocker to reach 10,000+ points with 1x GPU in Fire Strike Extreme!	Competitions	Competitions
How to insulate a motherboard for subzero cooling, guide inside!	Sales	HowTo
OCAU overclocking competition with a twist, how low can you score in 3DMARK!	Informational	Informational
GIGABYTE X99 SOC Champion breaks XTU world record and highest DDR4 memory frequency	Sales	Sales
Top 10 Best-Selling Games of 2014	Informational	Sales
CES 2015 Coverage	Informational	Informational
Steam Daily Deal: 3DMARK 80% OFF	Informational	Sales
GIGABYTE X99 SOC Champion Debuts at CES 2015 with One More WR Under its Belt	Informational	Informational
DDR4 Memory Clock Reaches New High, GIGABYTE X99-SOC Champion rocks!	Informational	Competitions
HWBOT Country Cup is over, Australia wins again!	Informational	Competitions
Hardware Asylum podcast on evolution of overclocking motherboards	Informational	Informational
X99-SOC Champion - Let it Rain Records	Informational	Informational
X99-SOC Champion - Let it Rain Records	Informational	Informational

3. Assess the performance of your classifier in each of your categories by computing precision, recall, and F-measure.

To start this problem, I had to look at the predictions made by the program in problem two and compute the amount of TP (true positive), FP (false positive), and FN (false negative) values for each category.

	True Pos	False Pos	False Neg
Competitions	5	9	0
Informational	16	12	0
Sales	8	5	0
HowTo	1	1	1

From this, using the formula from the week 12 slides, I was able to calculate the precision, recall, and F-measure values:

	Precision	Recall	F-Measure
Competitions	0.55.	0.55	0.55
Informational	1.33.	1.33.	1.33.
Sales	1.6.	1.6.	1.6.
HowTo	1	.5.	0.66

The first three categories were the same due to the fact that I did not have any False Negatives, different on the 'HowTo' category. This shows that Informational and Sales were very accurate in their assumptions, while Competitions and HowTo were less accurate.