Ryan Condotta

Final Report Assignment 9

Question 1:

The blog that I chose was http://theworldsfirstinternetbaby.blogspot.com/ which is a blog that displays music and a little bit of history about them. Since I listen to music all the time, I figured this would be a good blog for me to choose and provide quality training data. The different categories that I created for this that I thought would be relevant were Rock N Roll, Other, Alternative, and Metal. I thought these categories would be great for the entries that were on the blog. I used the RSS of:

http://theworldsfirstinternetbaby.blogspot.com/feeds/posts/default?max-results=100&alt=rss

I also created the python script of A9Q1.py. In this file, it parses the url using the feedparser to get entries and where I clarified that there were at least 100 entries and processes the page as we did in week 12 slides as shown in Figure 1.

```
# Returns title and dictionary of word counts for an RSS feed
def getwordcounts(url):
  # Parse the feed
  d=feedparser.parse(url)
  wc={}

  # Loop over all the entries
  count = 0
  print(len(d.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)
    for word in words:
      wc.setdefault(word,0)
      wc[word]+=1
  return d.feed.title,wc

def getwords(html):
  # Remove all the HTML tags
  txt=re.compile(r'<[^>]+>').sub('',html)
  # Split words by all non-alpha characters
  words=re.compile(r'[^A-Z^a-z]+').split(txt)

  # Convert to lowercase
  return [word.lower() for word in words if word!='']
```

Figure 1

Lastly, I output the html of the RSS link to a text file named RawRSS.txt and is created by the code as show in Figure 2.

```
filename2 = r"C:\Users\Ryan\Documents\WebScience\Assignment9\Question1\RawRSS.txt"
out=open(filename2,'w')
response = urllib.request.urlopen(feedlist[0])
soup = BeautifulSoup(response, 'html.parser')
out.write(str(soup.encode("utf-8")))
```

Figure 2

Question 2:

In order to classify the entries in from the RSS feed, I had to create a python script of A9Q2.py. As a starting point I borrowed a class library source code from the Programming Collective Intelligence in the way of docclass.py. That file has many functionalities but has two classes of classifier and a subclass of fisherclassifier. In A9Q2.py, it contains two functions of remove_html_tags and readfile. The first function, remove_html_tags, which removes all of the html from the RSS feed. Readfile takes in the link to get the feed from and the fisherclassifier. The function works by getting the feed from the entry to get the information such as title and entries that could be used in classifying information. The classifier will take a guess on what category it believes it belongs and for the first 50 entries, I choose the correct category that it belongs to in order to train the data. After the first 50, fisher classifier will classify the rest of the items. Figure 3 shows the readfile function. The title, predicted category, actual category, cprob(), and fisherprob is outputted to an excel file named Table and the table is shown in Figure 4.

```
def readfile(feed, fisherclassifier):
    # Get feed entries and loop over them
    title =[]
    guess =[]
    actual=[]
    f=feedparser.parse(feed)
    print(len(f))
    count = 0
    for entry in f['entries']:
        count = count + 1
        print(count)
        print ('-----')
        # Print the contents of the entry
        print ('Title: '+str(entry['title'].encode('utf-8')))
        print()
        fulltext = remove_html_tags(entry['title'])
        title.append(str(fulltext))
        fulltext = fulltext +" " +remove_html_tags(entry['summary'])
        # Print the best guess at the current category
        if count < 50:
            #print('Guess: '+str(fulltext.encode('utf-8')))
            value = str(fisherclassifier.classify(fulltext))
```

```
                print('Guess: '+ value)
                guess.append(value)
                # Ask the user to specify the correct category and train on that
                temp = input('Enter Category:')
                print("value: ",temp)
                actual.append(temp)
                fisherclassifier.train(fulltext,temp)
        else:
                value1 = str(fisherclassifier.classify(fulltext))
                print(value1)
                guess.append(value1)
        print()

#input("Help")
filename2 = r'C:\Users\Ryan\Documents\WebScience\Assignment9\Table2.xlsx'
workbook = xlsxwriter.Workbook(filename2)
worksheet = workbook.add_worksheet()

bold = workbook.add_format({'bold' : 1})

#data headers
worksheet.write('A1', 'Title', bold)
worksheet.write('B1', 'Predicted', bold)
worksheet.write('C1', 'Actual', bold)

row = 1
col = 0

#throw data into
for name in title:
        worksheet.write_string(row, col, name)
        row += 1

row1 = 1
col1 = 1
for id in guess:
        worksheet.write_string(row1, col1, id)
        row1 += 1

row1 = 1
col1 = 2
for id1 in actual:
        worksheet.write_string(row1, col1, id1)
        row1 += 1
workbook.close()

return c2
```

Figure 3

Title	Predicted	Actual	Cprob	Fisher Prob
Wayne Shorter "JuJu" (1964)	None	Other	0	0.75
Pavement "Wowee Zowee" (1995)	Other	Metal	0	0.75
Pixies "Trompe le Monde" (1991)	Metal	Other	0	0.75
The Smiths "Meat is Murder" (1985)	Metal	Alternative	0.33	0.75
David Bowie "The Man Who Sold the World" (1970)	Alternative	Rock N Roll	0	0.75
Mott the Hoople "All the Young Dudes" (1972)/"Mott" (1973)	Rock N Roll	Rock N Roll	0	0.75
Cheap Trick "Cheap Trick" (1977)	Rock N Roll	Metal	0	0.75
T. Rex "The Slider" (1972)	Rock N Roll	Metal	0.75	0.75
T. Rex "Electric Warrior" (1971)	Metal	Metal	0.75	0.75
Bad Brains "Rock for Light" (1983)	Rock N Roll	Rock N Roll	0.75	0.75
Syd Barrett "The Madcap Laughs" (1970)	Rock N Roll	Alternative	0.75	0.75
Silver Apples "Silver Apples" (1968)	Rock N Roll	Alternative	0.75	0.75
Bob Dylan "Bringing It All Back Home" (1965)	Rock N Roll	Rock N Roll	0.9	0.75
Bob Dylan "The Times They Are A-Changin'" (1964)	Rock N Roll	Rock N Roll	0.9	0.75
Curtis Mayfield "Super Fly" (1972)	Metal	Other	0	0.75
Bruce Springsteen "Nebraska" (1982)	Rock N Roll	Alternative	0	0.75
John Lennon "Imagine" (1971)	Metal	Alternative	0	0.75
The United States of America "The United States of America" (1968)	Alternative	Metal	0	0.75
A Tribe Called Quest "The Low End Theory" (1991)	Rock N Roll	Rock N Roll	0.75	0.75
Ice Cube "Death Certificate" (1991)	Other	Other	0	0.75
Ice Cube "AmeriKKKa's Most Wanted" (1990)	Other	Other	0	0.75
Bob Dylan "The Freewheelin' Bob Dylan" (1963)	Rock N Roll	Rock N Roll	0.9	0.75
The Rolling Stones "Beggars Banquet" (1968)	Metal	Alternative	0	0.75
The Clash "London Calling" (1979)	Rock N Roll	Metal	0	0.75
X "Under the Big Black Sun" (1982)/"More Fun in the New World" (1983)	Rock N Roll	Metal	0	0.75
X "Los Angeles" (1980)/"Wild Gift" (1981)	Rock N Roll	Alternative	0	0.75
Wipers "Youth of America" (1981)	Metal	Metal	0	0.75
nts "Sorry Ma, Forgot to Take Out the Trash" (1981)/"Stink" (1982)"Hootenanny" (1983)/"	Metal	Metal	0	0.75
Joy Division "Unknown Pleasures" (1979)/"Closer" (1980)	Metal	Rock N Roll	0	0.75
Sonic Youth "Sister" (1987)	Metal	Alternative	0.75	0.75
Sonic Youth "EVOL" (1986)	Alternative	Alternative	0.75	0.75
Sonic Youth "Bad Moon Rising" (1985)	Alternative	Alternative	0.75	0.75
Buzzcocks "Singles Going Steady" (1979)	Rock N Roll	Rock N Roll	0	0.75
Pink Floyd "The Piper at the Gates of Dawn" (1967)	Rock N Roll	Rock N Roll	0	0.75
Todd Rundgren "Runt" (1970)	Rock N Roll	Alternative	0	0.75
The Mothers of Invention "Freak Out!" (1966)	Metal	Metal	0	0.75
Alice Cooper "Love It to Death"/"Killer" (1971)	Metal	Metal	0	0.75
Love "Love" (1966)/"Da Capo" (1967)/"Forever Changes" (1967)	Metal	Rock N Roll	0	0.75
Bob Dylan "Blood on the Tracks" (1975)	Rock N Roll	Rock N Roll	0.9	0.75
The Flying Burrito Brothers "The Gilded Palace of Sin" (1969)	Rock N Roll	Metal	0	0.75
Bob Dylan "Another Side of Bob Dylan" (1964)	Rock N Roll	Rock N Roll	0.9	0.75
Bob Dylan "John Wesley Harding" (1967)	Rock N Roll	Rock N Roll	0.9	0.75
Nick Drake "Pink Moon" (1972)	Rock N Roll	Alternative	0	0.75
The Germs "(GI)" (1979)	Rock N Roll	Metal	0	0.75
The Smiths "The Queen is Dead" (1986)	Alternative	Metal	0.5	0.75
Lydia Lunch "Queen of Siam" (1980)	Alternative	Other	0	0.75
Johnny Thunders "So Alone" (1978)	Rock N Roll	Rock N Roll	0	0.75
We're Gonna Die (Send Those Curse Words to Hell!)	Rock N Roll	Rock N Roll	0	0.75
Patti Smith "Horses" (1975)	Rock N Roll	Rock N Roll	0.33	0.75

Dinosaur Jr. "You're Living All Over Me" (1987)	Rock N Roll	Rock N Roll	0	0.5
Bikini Kill "Pussy Whipped" (1993)	Rock N Roll	Rock N Roll	0	0.5
Jungle Brothers "Straight Out the Jungle" (1988)	Metal	Metal	0	0.75
3rd Bass "The Cactus Album" (1989)	Metal	Metal	0	0.75
Billie Holiday "Lady in Satin" (1958)	Rock N Roll	Alternative	0	0.5
Isaac Hayes "Hot Buttered Soul" (1969)	Metal	Other	0	0.5
Black Sheep "A Wolf in Sheep's Clothing" (1991)	Other	Metal	0	0.5
Boogie Down Productions "By All Means Necessary" (1988)	Rock N Roll	Alternative	0.5	0.75
Boogie Down Productions "Criminal Minded" (1987)	Alternative	Alternative	0.5	0.75
Nas "Illmatic" (1994)	Rock N Roll	Other	0	0.75
We Were the Best (Enjoy Poverty, Dumbass Hipsters)	Metal	Metal	0	0.75
Pere Ubu "Terminal Tower" (1985)	Alternative	Alternative	0	0.75
Pavement "Crooked Rain, Crooked Rain" (1994)	Metal	Metal	0	0.75
Joni Mitchell "Blue" (1971)	Metal	Metal	0	0.75
The Strokes "Is This It" (2001)	Metal	Alternative	0	0.75
Nirvana "MTV Unplugged in New York" (1994)	Metal	Alternative	0	0.75
Elliott Smith "Elliott Smith" (1995)	Rock N Roll	Rock N Roll	0	0.75
Can "Future Days" (1973)	Rock N Roll	Rock N Roll	0.33	0.5
Can "Ege Bamyasi" (1972)	Other	Rock N Roll	0.33	0.5
Can "Tago Mago" (1971)	Metal	Rock N Roll	0.33	0.5
Pretenders "Pretenders II" (1981)	Metal	Metal	0.5	0.75
Radiohead "The Bends" (1995)	Metal	Metal	0	0.75
Elliott Smith "Either/Or" (1997)	Rock N Roll	Rock N Roll	0.33	0.75
Bad Brains "I Against I" (1986)	Alternative	Alternative	0	0.75
Neutral Milk Hotel "In the Aeroplane over the Sea" (1998)	Metal	Alternative	0	0.75
Throbbing Gristle "20 Jazz Funk Greats" (1979)	Metal	Other	0	0.75
Nine Inch Nails "Pretty Hate Machine" (1989)	Rock N Roll	Metal	0	0.75
Bratmobile "Pottymouth" (1993)	Rock N Roll	Metal	0	0.75
Hole "Pretty on the Inside" (1991)	Rock N Roll	Metal	0	0.75
Wipers "Is This Real?" (1980)	Other	Alternative	0.5	0.75
The Breeders "Pod" (1990)	Metal	Rock N Roll	0	0.75
Miles Davis "Birth of the Cool" (1957)	Metal	Rock N Roll	0	0.75
The Who "The Who Sell Out" (1967)	Rock N Roll	Rock N Roll	0.75	0.75
The Who "A Quick One (Happy Jack)" (1966)	Rock N Roll	Rock N Roll	0.75	0.75
The Who "The Who Sings My Generation" (1965)	Rock N Roll	Rock N Roll	0.75	0.75
Led Zeppelin "Houses of the Holy" (1973)	Rock N Roll	Rock N Roll	0	0.5
Led Zeppelin "Led Zeppelin IV" (1971)	Metal	Rock N Roll	0	0.5
Alexander "Skip" Spence "Oar" (1969)	Metal	Metal	0	0.5
Neil Young "After the Gold Rush" (1970)	Rock N Roll	Rock N Roll	0.75	0.9
Neil Young "Everybody Knows This is Nowhere" (1969)	Rock N Roll	Rock N Roll	0.75	0.9
Neil Young "On the Beach" (1974)	Rock N Roll	Rock N Roll	0.75	0.9
Doctors of Madness "Late Night Movies, All Night Brainstorms" (1976)	Rock N Roll	Rock N Roll	0	0.75
Tav Falco's Panther Burns "Behind the Magnolia Curtain" (1981))	Metal	Metal	0	0.75
The Clash "Give 'Em Enough Rope" (1978)	Metal	Metal	0	0.75
The World's First Internet Baby Interviews former DEAD BOYS bassist JEFF MAGNUM!	Metal	Metal	0	0.75
Descendents "Milo Goes to College" (1982)	Metal	Rock N Roll	0	0.75
Green Day "Dookie" (1994)	Rock N Roll	Rock N Roll	0	0.75
Iggy and the Stooges "Raw Power" (1973)	Rock N Roll	Rock N Roll	0	0.75
Tom Petty & the Heartbreakers "Tom Petty & the Heartbreakers" (1976)	Metal	Metal	0.5	0.75
Wipers "Over the Edge" (1983)	Metal	Alternative	0.5	0.75
The Jesus and Mary Chain "Psychocandy" (1985)	Alternative	Alternative	0	0.75

Figure 4

In the Excel file, it also has the fisher probability for the other categories as well.

Question 3:

The final part was to assess the performance of the classifier for Precision, Recall, and F-Measure for each of the categories which were Other, Metal, Rock N Roll, and Alternative. I added for more tables to the Table.xls file. Figure 5 shows the table of number of True positives, false negatives, and false positives. Rock N Roll was a category that had a lot of elements while other category had the least amount of elements. All these values were crucial in calculating Precision, Recall, and F-Measure.

	TP	FP	FN
Metal	17	18	2
Other	3	4	0
Alternative	7	5	1
Rock N Roll	28	18	3

Figure 5

In Figure 6, it displays all the tables for each of the categories for Precision, Recall, and F-Measure. As you can see from Figure 6, none of the categories were precise and were almost all below 60%, however the recall was generally higher than 85% which always put the F-Measure in between for each of the categories. The classifier seemed to be most proficient with the Rock N Roll than it was for any of the other categories while Other was the smallest because it had the least amount of training data.

Metal	**Precision**	**Recall**	**F-Measure**
	0.4857143	0.894737	0.62962963
Other	**Precision**	**Recall**	**F-Measure**
	0.4285714	1	0.6
Alternative	**Precision**	**Recall**	**F-Measure**
	0.5833333	0.875	0.7
Rock N Roll	**Precision**	**Recall**	**F-Measure**
	0.6086957	0.903226	0.727272727

Figure 6