Assignment 9
CS532, Web Science, Spring 2017
Old Dominion University, Computer Science Dept

Hussam Hallak
CS Master's Student
Prof: Dr. Nelson

Question 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.
Create a table with 100 rows, like:

Listing 1: Command:

```
title                       classification
-----                       --------------
Ric Ocasek -                     80s
"Something To Grab
For" (forgotten song)

Weezer - "Pinkerton"           alternative
(LP Review)

Schon \& Hammer -                   80s
"No More Lies"
(forgotten song)
```

etc. This is your "ground truth" (or "gold standard") data.

Answer:

I claimed "Programming Tutorials blog" on blogspot:
http://tutorialsprogram.blogspot.com/
RSS Feed:
http://tutorialsprogram.blogspot.com/feeds/posts/default
The approach for grabbing 100 entries from the blog is divided into two steps:

1. Get the pages of the RSS feed. I modified the script, "getpages.py", I wrote for Assignment #8 to use it to get all pages in the feed. I renamed the script "getfeed.py". it takes feed url as a command line argument. It gets all feed pages and saves the output to a file names "pages.txt".

Listing 2: The content of getfeed.py

```
import sys
from bs4 import BeautifulSoup
import urllib2
import re
if len(sys.argv) < 2:
   print "Usage: python getfeed.py <feed_url>"
   print "e.g: python getfeed.py http://tutorialsprogram.blogspot.com/feeds/posts
       /default"
   exit()
fh_output = open('pages.txt', 'w')

def getNextPage(link):
   try:
      html = urllib2.urlopen(link).read()
      soup = BeautifulSoup(html, 'lxml')
      next_page = soup.find('link', rel="next")
      if(next_page != []):
         next_page = next_page.get('href')
         return next_page
   except:
      return False

def getAllPages(link):
   all_pages = []
   next_page = getNextPage(link)
   while(next_page != False):
      all_pages.append(next_page)
      next_page = getNextPage(next_page)
   return all_pages

try:
   html = urllib2.urlopen(sys.argv[1]).read()
   soup = BeautifulSoup(html, 'lxml')
   title = soup.title.string.encode('ascii')
   rss = soup.find('link', type='application/atom+xml')
   rss = rss.get('href')
   pages = getAllPages(rss)
   pages.insert(0,rss)
   for page in pages:
      fh_output.write(page + '\n')
except:
   print 'Error'
fh_output.close()
```

Listing 3: Running getfeed.py

```
root@ima-app:/var/www/Hussam/A9# python getfeed.py http://tutorialsprogram.
    blogspot.com/feeds/posts/default
root@ima-app:/var/www/Hussam/A9# ls
downloadrss.py getfeed.py pages.txt
root@ima-app:/var/www/Hussam/A9# cat pages.txt
http://tutorialsprogram.blogspot.com/feeds/posts/default
http://www.blogger.com/feeds/7802129033476755546/posts/default?start-index=26&max
    -results=25
https://www.blogger.com/feeds/7802129033476755546/posts/default?start-index=51&
```

```
    max-results=25
https://www.blogger.com/feeds/7802129033476755546/posts/default?start-index=76&
    max-results=25
https://www.blogger.com/feeds/7802129033476755546/posts/default?start-index=101&
    max-results=25
```

2. Get all entries in each feed page. I wrote a simple python script, “downloadrss.py”, to do that using feedparser. It takes the feed url as a command line argument. It gets the title and the url for all entries and saves them to a file named “rss.txt”

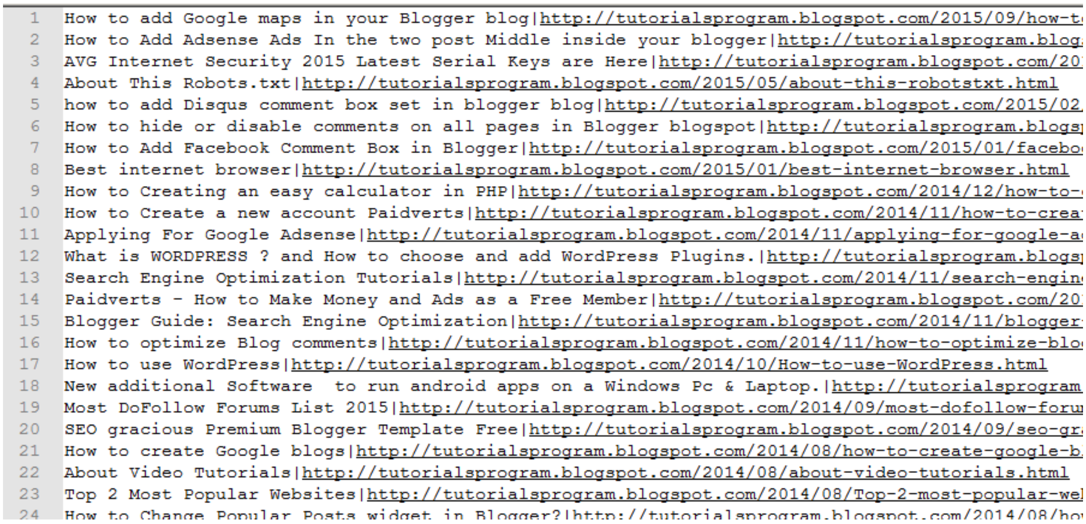
Listing 4: The content of downloadrss.py

```
#! /usr/bin/python
import sys
import feedparser
import socket
if len(sys.argv) < 2:
   print "Usage: python downloadrss.py <feed_url>"
   print "e.g: python downloadrss.py http://tutorialsprogram.blogspot.com/feeds/
       posts/default"
   exit()
timeout = 120
socket.setdefaulttimeout(timeout)
feed_url = sys.argv[1]
d = feedparser.parse(feed_url)
fh_output = open('rss.txt', 'a')
for s in d.entries:
   fh_output.write(unicode(s.title).encode("utf-8") + "|" + unicode(s.link).
       encode("utf-8") + "\n")
```

Listing 5: Running downloadrss.py

```
root@ima-app:/var/www/Hussam/A9# python downloadrss.py http://tutorialsprogram.
    blogspot.com/feeds/posts/default
root@ima-app:/var/www/Hussam/A9# python downloadrss.py http://www.blogger.com/
    feeds/7802129033476755546/posts/default?start-index=26&max-results=25
root@ima-app:/var/www/Hussam/A9# python downloadrss.py https://www.blogger.com/
    feeds/7802129033476755546/posts/default?start-index=51&max-results=25
root@ima-app:/var/www/Hussam/A9# python downloadrss.py https://www.blogger.com/
    feeds/7802129033476755546/posts/default?start-index=76&max-results=25
root@ima-app:/var/www/Hussam/A9# python downloadrss.py https://www.blogger.com/
    feeds/7802129033476755546/posts/default?start-index=101&max-results=25
root@ima-app:/var/www/Hussam/A9# ls
downloadrss.py getfeed.py pages.txt rss.txt
root@ima-app:/var/www/Hussam/A9# cat rss.txt | wc -l
120
root@ima-app:/var/www/Hussam/A9#
```

The file “rss.txt” is too big to include in this report, but here is a screen shot. The title and the url of each entry are saved on the same line separated by the “pipe” character.



```
How to add Google maps in your Blogger blog|http://tutorialsprogram.blogspot.com/2015/09/how-t
How to Add Adsense Ads In the two post Middle inside your blogger|http://tutorialsprogram.blog
AVG Internet Security 2015 Latest Serial Keys are Here|http://tutorialsprogram.blogspot.com/20
About This Robots.txt|http://tutorialsprogram.blogspot.com/2015/05/about-this-robotstxt.html
how to add Disqus comment box set in blogger blog|http://tutorialsprogram.blogspot.com/2015/02
How to hide or disable comments on all pages in Blogger blogspot|http://tutorialsprogram.blogs
How to Add Facebook Comment Box in Blogger|http://tutorialsprogram.blogspot.com/2015/01/facebo
Best internet browser|http://tutorialsprogram.blogspot.com/2015/01/best-internet-browser.html
How to Creating an easy calculator in PHP|http://tutorialsprogram.blogspot.com/2014/12/how-to-
How to Create a new account Paidverts|http://tutorialsprogram.blogspot.com/2014/11/how-to-crea
Applying For Google Adsense|http://tutorialsprogram.blogspot.com/2014/11/applying-for-google-a
What is WORDPRESS ? and How to choose and add WordPress Plugins.|http://tutorialsprogram.blogs
Search Engine Optimization Tutorials|http://tutorialsprogram.blogspot.com/2014/11/search-engin
Paidverts - How to Make Money and Ads as a Free Member|http://tutorialsprogram.blogspot.com/20
Blogger Guide: Search Engine Optimization|http://tutorialsprogram.blogspot.com/2014/11/blogger
How to optimize Blog comments|http://tutorialsprogram.blogspot.com/2014/11/how-to-optimize-blo
How to use WordPress|http://tutorialsprogram.blogspot.com/2014/10/How-to-use-WordPress.html
New additional Software  to run android apps on a Windows Pc & Laptop.|http://tutorialsprogram
Most DoFollow Forums List 2015|http://tutorialsprogram.blogspot.com/2014/09/most-dofollow-foru
SEO gracious Premium Blogger Template Free|http://tutorialsprogram.blogspot.com/2014/09/seo-gr
How to create Google blogs|http://tutorialsprogram.blogspot.com/2014/08/how-to-create-google-b
About Video Tutorials|http://tutorialsprogram.blogspot.com/2014/08/about-video-tutorials.html
Top 2 Most Popular Websites|http://tutorialsprogram.blogspot.com/2014/08/Top-2-most-popular-we
How to Change Popular Posts widget in Blogger?|http://tutorialsprogram.blogspot.com/2014/08/ho
```

Categories: I regretted claiming this blog because it is not rich nor is it organized to say the least. I am a full-time programmer, so I thought the best topic, for which I am qualified to provide classification training data, would be programming. I searched for programming tutorials blogs on blogspot in Google and randomly selected one of the results. I chose to have the following categories:

1. Tutorials
2. Articles
3. Downloads
4. Other

I finally created a simple python script, "getrawdata.py" to get the raw data. The script accepts the file "pages.txt" from the command line as input. It goes through each url in the input file and downloads the raw data, and saves it to files named 1.RawRSS, 2.RawRSS, 3.RawRSS, ...etc.

Listing 6: The content of getrawdata.py

```python
import sys
import urllib2

if len(sys.argv) != 2:
   print "Usage: python getrawdata.py <feed_urls_input_file>"
   print "e.g: python getrawdata.py pages.txt"
   exit()

fh_input = open (sys.argv[1], 'r')
urls = []

for line in fh_input:
   urls.append(line.rstrip())

counter = 1
for url in urls:
   response = urllib2.urlopen(url)
   page = response.read()
   output_file = str(counter) + '.RawRSS'
   fh_output = open(output_file, 'w')
   fh_output.write(page)
   fh_output.close()
   counter = counter + 1

fh_input.close()
```

Listing 7: Running getrawdata.py

```
root@ima-app:/var/www/Hussam/A9# python getrawdata.py pages.txt
root@ima-app:/var/www/Hussam/A9# ls
1.RawRSS 2.RawRSS 3.RawRSS 4.RawRSS 5.RawRSS downloadrss.py getfeed.py getrawdata
    .py pages.txt rss.txt
root@ima-app:/var/www/Hussam/A9#
```

For the title-classification table, I will only use the first 100 entries per the question in the assignment. I saved the first 100 entries to a file and named it “100rss.txt”. I manually categorized all 100 entries.

```
root@ima-app:/var/www/Hussam/A9# head -n 100 rss.txt > 100rss.txt
root@ima-app:/var/www/Hussam/A9# ls
100rss.txt 1.RawRSS 2.RawRSS 3.RawRSS 4.RawRSS 5.RawRSS downloadrss.py
    getfeed.py getrawdata.py pages.txt rss.txt
root@ima-app:/var/www/Hussam/A9#
```

For the table, I created a python script, “maketable.py”, to generate LaTex code necessary to make the table. This is just to save typing time.

Listing 8: The content of maketable.py

```
import sys

if len(sys.argv) != 2:
   print "Usage: python maketable.py <input_file>"
   print "e.g: python maketable.py 100rss.txt"
   exit()

fh_input = open (sys.argv[1], 'r')
titles = []

for line in fh_input:
   line = line.split('|')
   title = line[0].strip()
   titles.append(title)

fh_output = open('tex.table','w')
fh_output.write('\\begin{longtable}{ |p{12cm}|p{2cm}| }'+'\n')
fh_output.write('\hline'+'\n')
fh_output.write('Title'+'\n')
fh_output.write('&'+'\n')
fh_output.write('Classification'+' \\\\ '+'\n')

for title in titles:
   fh_output.write('\hline'+'\n')
   fh_output.write(title)
   fh_output.write('\n&\n')
   fh_output.write('class'+' \\\\ '+'\n')
fh_output.write('\end{longtable}')
fh_input.close()
fh_output.close()
```

Listing 9: Running maketable.py

```
root@ima-app:/var/www/Hussam/A9# python maketable.py 100rss.txt
root@ima-app:/var/www/Hussam/A9# ls
100rss.txt 1.RawRSS 2.RawRSS 3.RawRSS 4.RawRSS 5.RawRSS downloadrss.py getfeed.py
     getrawdata.py maketable.py pages.txt rss.txt tex.table
root@ima-app:/var/www/Hussam/A9#
```

Title	Classification
How to add Google maps in your Blogger blog	Tutorials
How to Add Adsense Ads In the two post Middle inside your blogger	Tutorials
AVG Internet Security 2015 Latest Serial Keys are Here	Downloads
About This Robots.txt	Articles
how to add Disqus comment box set in blogger blog	Tutorials
How to hide or disable comments on all pages in Blogger blogspot	Tutorials
How to Add Facebook Comment Box in Blogger	Tutorials
Best internet browser	Articles
How to Creating an easy calculator in PHP	Tutorials
How to Create a new account Paidverts	Articles
Applying For Google Adsense	Articles
What is WORDPRESS ? and How to choose and add WordPress Plugins.	Articles
Search Engine Optimization Tutorials	Tutorials
Paidverts - How to Make Money and Ads as a Free Member	Articles
Blogger Guide: Search Engine Optimization	Articles
How to optimize Blog comments	Tutorials
How to use WordPress	Tutorials
New additional Software to run android apps on a Windows Pc & Laptop.	Downloads
Most DoFollow Forums List 2015	Articles
SEO gracious Premium Blogger Template Free	Downloads
How to create Google blogs	Tutorials
About Video Tutorials	Other
Top 2 Most Popular Websites	Articles
How to Change Popular Posts widget in Blogger?	Tutorials
Mobogenie Apps Download PC and Android Phone	Downloads
Microworkers Forum Posting Jobs top secret explication	Articles
Start Earning From Micro-workers Jobs	Articles
Face-book Like box For Blogger	Tutorials
Examples With Basic HTML Tags	Others
Search Engine Optimization means that a Web site visitor through SEO-much increase	Articles
How to Earn Money Online in Site Building	Articles
Job as a Search Engine Optimization	Others

How To Remove Powered By Blogger Attribution From Blogger	Tutorials
About Author	Others
Free Blogger Templates For Blogspot	Downloads
How to SEO Optimize Archive Links in Blogger	Tutorials
How to Change The Background Color in Blogger?	Tutorials
How to add your blog Livefyre Comment System	Tutorials
Some of the Nokia mobile required code	Others
Hill Climb Racing for PC free Download Game (Windows 7	Downloads
How to add Twitter Flying bird Widget to Blog?	Tutorials
A Blog Widget	Others
About Internet	Articles
MOBILE PHONE APPLICATIONS	Articles
Introduction	Others
Introduction of HTML	Tutorials
About Programming Tutorials	Others
Background Properties	Tutorials
Background Properties__CSS Background Color	Tutorials
Background Properties_ CSS Text/Font Color	Tutorials
Types - External Style Sheet	Tutorials
Types - User Defined	Tutorials
Internal Styles - Identifiers	Tutorials
Internal Styles - Universal	Tutorials
CSS Style Types >> Topic-01	Tutorials
Rule with CSS	Articles
HTML Anchor Tag	Tutorials
HTML >> acronym/abbreviation	Articles
HTML Phrase Tags	Tutorials
Html >> Canvas	Tutorials
Html >> Lists with HTML Bullets - (PART_53)	Tutorials
Html >> Subscript/Superscript - (PART_52)	Tutorials
Html >> Image tags - HTML form with submit button as image - (PART_51)	Tutorials
Html >> Image tags- MAP/Image Mapping - (PART_50)	Tutorials
HTML >> DTD - Document Type Definition - (PART_49)	Tutorials
Htm >> HTML Tooltip - (PART_48)	Tutorials
Html >> HTML Fieldset and Legend Tags - (PART_47)	Tutorials
Html >> Frame tags ¿¿Frames and Frameset Tag - (PART_46)	Tutorials
Html >> Meta Tag - (PART_45)	Tutorials
Html >> Media ¿¿Embedding Video to HTML File - (PART_44)	Tutorials
Html >> Media ¿¿Embedding Audio/Sound in background - (PART_43)	Tutorials
Html >> CSS an Intro - (PART_42)	Tutorials
Html >> Address Bar Icon in HTML page - (PART_41)	Tutorials
Html >> IFrame tag - (PART_40)	Tutorials
Html >> Auto refresh / Reload web page - (PART_39)	Tutorials
Html >> Web Page Auto Redirection - (PART_38)	Tutorials
Special/ASCII Characters in HTML - (PART_37)	Tutorials
Forms >> Html Label - (PART_36)	Tutorials

Forms >> Html Password Field - (PART_35)	Tutorials
Forms >> Combo Box / Dropdown - (PART_34)	Tutorials
Forms >> HTML TextArea Tag - (PART_33)	Tutorials
Forms >> HTML Check Box Tag - (PART_32)	Tutorials
Forms >> HTML Radio Button Tag - (PART_31)	Tutorials
Forms >> HTML Button Tag - (PART_30)	Tutorials
Forms >> HTML Text Field Code - (PART_29)	Tutorials
Forms¿¿ HTML Form Basic Types - (PART_28)	Tutorials
Using Tables >> Table Row Span - (PART_27)	Tutorials
Using Tables >> Table Col Span - (PART_26)	Tutorials
Using Tables >> Inner Tables - (PART_25)	Tutorials
Using Tables Table Alignment Code- (PART_24)	Tutorials
Using Tables >> HTML Table Size - code - (PART_23)	Tutorials
Using Tables >> Table Background Color and Image - (PART_22)	Tutorials
Using Tables >> HTML Table Border - (PART_21)	Tutorials
Using Tables >> Creating HTML Table - (PART_20)	Tutorials
Special Effects >> Special Effects - Blinking Text Tag- (PART_19)	Tutorials
Special Effects >> Special Effects - Marquee Tag- (PART_18)	Tutorials
Special Effects >> Special Effects - Marquee Tag -(PART_17)	Tutorials
LINKS >> Mail Link Code -(PART_16)	Tutorials
LINKS >> LINK Handling Code -(PART_15)	Tutorials
LINKS >> HTML LINKS Tag -(PART_14)	Tutorials

Included Files:

1.RawRSS
2.RawRSS
3.RawRSS
4.RawRSS
5.RawRSS
100rss.txt
downloadrss.py
getfeed.py
getrawdata.py
maketable.py
pages.txt
rss.png
rss.txt
tex.table

Question 2 & 3:

2. Train the Fisher classifier on the first 50 entries (the ”training set”), then use the classifier to guess the classification of the next 50 entries (the ”test set”).

Create a table with 50 rows, like

Listing 10: Command:

```
title                    actual            predicted
-----                    ------            ---------
Donnie Iris -              80s                 80s
"Ah! Leah!"
(Forgotten Song)

Black Sabbath -            metal               metal
"Vol. 4" (LP Review)

Catherine Wheel -          alternative         metal
"Ferment" (LP Review)
```

Assess the performance of your classifier in each of your categories by computing precision, recall, and F-measure. Use the "macro-averaged" label based method, as per:

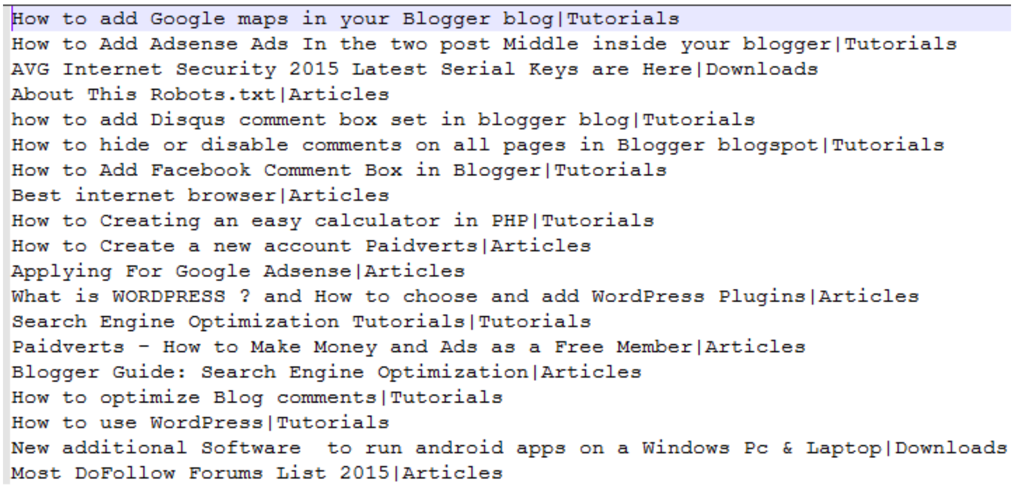
http://stats.stackexchange.com/questions/21551/how-to-compute-precision-recall-for-multiclass-multilabel-classification

For example, if you have 5 categories (e.g., 80s, metal, alternative, electronic, cover), you will compute precision, recall, and F-measure for each category, and then compute the average across the 5 categories.

3. Repeat question #2, but use the first 90 entries to train your classifier and the last 10 entries for testing.

Answer:

I put the titles and the categories for the 100 entries in a text file named "actual.txt". Every entry and its classification are separated by the "pipe" character. and placed on one line in the file "actual.txt". The file is included in the folder "Q2-3" and here is a screen shot.



```
How to add Google maps in your Blogger blog|Tutorials
How to Add Adsense Ads In the two post Middle inside your blogger|Tutorials
AVG Internet Security 2015 Latest Serial Keys are Here|Downloads
About This Robots.txt|Articles
how to add Disqus comment box set in blogger blog|Tutorials
How to hide or disable comments on all pages in Blogger blogspot|Tutorials
How to Add Facebook Comment Box in Blogger|Tutorials
Best internet browser|Articles
How to Creating an easy calculator in PHP|Tutorials
How to Create a new account Paidverts|Articles
Applying For Google Adsense|Articles
What is WORDPRESS ? and How to choose and add WordPress Plugins|Articles
Search Engine Optimization Tutorials|Tutorials
Paidverts - How to Make Money and Ads as a Free Member|Articles
Blogger Guide: Search Engine Optimization|Articles
How to optimize Blog comments|Tutorials
How to use WordPress|Tutorials
New additional Software  to run android apps on a Windows Pc & Laptop|Downloads
Most DoFollow Forums List 2015|Articles
```

To perform the prediction, I wrote a small python program "predict.py" that takes the file "actual.txt" as input from the command line, and the program trains for the number of entries specified in the second command line argument, in this case 50 and then another run for 90, and predicts the classification for the number of entries specified in the third command line argument, 50 and then 10 entries in our case. The output is saved to a file in each run. These files are named "predictedtable50_50.txt" and "predictedtable90_10". The program imports the file "docclass.py" which is taken from "PCI" book and uses its functions "train" and "classify" to perform the training and the prediction.

Listing 11: The content of predict.py

```
import sys
import docclass
import re
import math

if len(sys.argv) < 4:
   print "Usage: python predict.py <grounnd_truth_file> <train_num> <predict_num>
       "
   print "e.g: python predict.py actual.txt 90 10"
   exit()
fh_input = open(sys.argv[1],'r')
file_name = 'predictedtable' + sys.argv[2] + '_' + sys.argv[3] + '.txt'
fh_output = open(file_name,'w')

titles = []
cats = []
predictions = []
train_num = int (sys.argv[2])
predicted_num = int (sys.argv[3])
max_data = train_num + predicted_num
count = 0

for line in fh_input:
   item = line.strip()
   item = item.split('|')
   cat = item[1]
   title = item[0]
   cats.append(cat)
   titles.append(title)

cl = docclass.fisherclassifier(docclass.getwords)
cl.setdb('hhallak.db')

while count < train_num:
   cl.train(title, cats[count])
   count = count + 1

while count < max_data:
   prediction = cl.classify(titles[count])
   predictions.append(prediction)
   count = count + 1

fh_output.write("Title" + '|' + "Actual" + '|' + "Predicted" + '\n')

for i in range(0,predicted_num):
   fh_output.write(titles[i+train_num] + '|' + cats[i+train_num] + '|' +
       predictions[i] + '\n')
fh_input.close()
fh_output.close()
```

Listing 12: Running predict.py: Training 50 Predicting 50

```
root@ima-app:/var/www/Hussam/A9# python predict.py actual.txt 50 50
```

The file “predictedtable50_50.txt” is then formatted to make the following table:

Title	Actual	Predicted
Types - External Style Sheet	Tutorials	Tutorials
Types - User Defined	Tutorials	Tutorials
Internal Styles - Identifiers	Tutorials	Tutorials
Internal Styles - Universal	Tutorials	Tutorials
CSS Style Types >> Topic-01	Tutorials	Tutorials
Rule with CSS	Articles	Tutorials
HTML Anchor Tag	Tutorials	Tutorials
HTML >> acronym/abbreviation	Articles	Tutorials
HTML Phrase Tags	Tutorials	Tutorials
Html >> Canvas	Tutorials	Tutorials
Html >> Lists with HTML Bullets - (PART_53)	Tutorials	Tutorials
Html >> Subscript/Superscript - (PART_52)	Tutorials	Tutorials
Html >> Image tags - HTML form with submit button as image - (PART_51)	Tutorials	Tutorials
Html >> Image tags- MAP/Image Mapping - (PART_50)	Tutorials	Tutorials
HTML >> DTD - Document Type Definition - (PART_49)	Tutorials	Tutorials
Htm >> HTML Tooltip - (PART_48)	Tutorials	Tutorials
Html >> HTML Fieldset and Legend Tags - (PART_47)	Tutorials	Tutorials
Html >> Frame tags ¿¿Frames and Frameset Tag - (PART_46)	Tutorials	Tutorials
Html >> Meta Tag - (PART_45)	Tutorials	Tutorials
Html >> Media ¿¿Embedding Video to HTML File - (PART_44)	Tutorials	Tutorials
Html >> Media ¿¿Embedding Audio/Sound in background - (PART_43)	Tutorials	Tutorials
Html >> CSS an Intro - (PART_42)	Tutorials	Tutorials
Html >> Address Bar Icon in HTML page - (PART_41)	Tutorials	Tutorials
Html >> IFrame tag - (PART_40)	Tutorials	Tutorials
Html >> Auto refresh / Reload web page - (PART_39)	Tutorials	Tutorials
Html >> Web Page Auto Redirection - (PART_38)	Tutorials	Tutorials
Special/ASCII Characters in HTML - (PART_37)	Tutorials	Tutorials
Forms >> Html Label - (PART_36)	Tutorials	Tutorials
Forms >> Html Password Field - (PART_35)	Tutorials	Tutorials
Forms >> Combo Box / Dropdown - (PART_34)	Tutorials	Tutorials
Forms >> HTML TextArea Tag - (PART_33)	Tutorials	Tutorials
Forms >> HTML Check Box Tag - (PART_32)	Tutorials	Tutorials
Forms >> HTML Radio Button Tag - (PART_31)	Tutorials	Tutorials
Forms >> HTML Button Tag - (PART_30)	Tutorials	Tutorials
Forms >> HTML Text Field Code - (PART_29)	Tutorials	Tutorials
Forms¿¿ HTML Form Basic Types - (PART_28)	Tutorials	Tutorials
Using Tables >> Table Row Span - (PART_27)	Tutorials	Tutorials
Using Tables >> Table Col Span - (PART_26)	Tutorials	Tutorials
Using Tables >> Inner Tables - (PART_25)	Tutorials	Tutorials
Using Tables Table Alignment Code- (PART_24)	Tutorials	Tutorials
Using Tables >> HTML Table Size - code - (PART_23)	Tutorials	Tutorials

Using Tables >> Table Background Color and Image - (PART_22)	Tutorials	Tutorials
Using Tables >> HTML Table Border - (PART_21)	Tutorials	Tutorials
Using Tables >> Creating HTML Table - (PART_20)	Tutorials	Tutorials
Special Effects >> Special Effects - Blinking Text Tag-(PART_19)	Tutorials	Tutorials
Special Effects >> Special Effects - Marquee Tag-(PART_18)	Tutorials	Tutorials
Special Effects >> Special Effects - Marquee Tag - (PART_17)	Tutorials	Tutorials
LINKS >> Mail Link Code -(PART_16)	Tutorials	Tutorials
LINKS >> LINK Handling Code -(PART_15)	Tutorials	Tutorials
LINKS >> HTML LINKS Tag -(PART_14)	Tutorials	Tutorials

Now it's time to compute precision, recall, and F-measure for all four categories for both splits, 50/50 and 90/10. I manually calculated precision, recall, and F-measure from the values of True Positive, False Positive, False Negative for each category using the following formulas:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The unfortunate choice of the blog feed brought involved math into the calculations of the results. It is the frustrating problem of $\frac{0}{0}$. I am going with what math scientists call it, "Indeterminate".

Category	TP	FP	FN	Precision	Recall	F-Measure
Tutorials	48	2	0	0.96	1	0.98
Articles	0	0	2	Indeterminate	0	Indeterminate
Downloads	0	0	0	Indeterminate	Indeterminate	Indeterminate
Other	0	0	0	Indeterminate	Indeterminate	Indeterminate

Listing 13: Running predict.py: Training 90 Predicting 10

```
root@ima-app:/var/www/Hussam/A9# python predict.py actual.txt 90 10
```

The file "predictedtable90_10.txt" is then formatted to make the following table:

Title	Actual	Predicted
Using Tables >> HTML Table Size - code - (PART_23)	Tutorials	Tutorials
Using Tables >> Table Background Color and Image - (PART_22)	Tutorials	Tutorials
Using Tables >> HTML Table Border - (PART_21)	Tutorials	Tutorials
Using Tables >> Creating HTML Table - (PART_20)	Tutorials	Tutorials
Special Effects >> Special Effects - Blinking Text Tag-(PART_19)	Tutorials	Tutorials

Special Effects >> Special Effects - Marquee Tag-(PART_18)	Tutorials	Tutorials
Special Effects >> Special Effects - Marquee Tag -(PART_17)	Tutorials	Tutorials
LINKS >> Mail Link Code -(PART_16)	Tutorials	Tutorials
LINKS >> LINK Handling Code -(PART_15)	Tutorials	Tutorials
LINKS >> HTML LINKS Tag -(PART_14)	Tutorials	Tutorials

Category	TP	FP	FN	Precision	Recall	F-Measure
Tutorials	10	0	0	1	1	1
Articles	0	0	0	Indeterminate	Indeterminate	Indeterminate
Downloads	0	0	0	Indeterminate	Indeterminate	Indeterminate
Other	0	0	0	Indeterminate	Indeterminate	Indeterminate

Included Files:

actual.txt
predict.py
predictedtable50_50.txt
predictedtable90_10.txt
docclass.py
docclass.pyc
hhallak.db