Assignment 10

CS595

Introduction to Web Science

Old Dominion University

Computer Science

Due: 11:59 pm Dec 12

Lulwah Alkwai

Support your answer:include all relevant discussion, assumptions, examples, etc.

Question 1

Choose a blog or a newsfeed (or something similar with an Atom or RSS feed). 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.

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.

Answer One-

The blog I choose for this assignment is "Rebecca Woolf" blog called "Girl's Gone Child!", she started to blog since 2005 and have almost 1910 blog posts. A link to the blog is: `http://www.girlsgonechild.net`

The categories I choose for the feeds entries are: Family, style, music, food

From the page source I found the xml link of the page; which is:

`http://www.girlsgonechild.net/feeds/posts/default?alt=rss`

To get the xml of 100 blog entries I used curl to the following url:`http://www.blogger.com/feeds/18751784/posts/default?max-results=100`

And I saved the content of the href link in a file called girlsgonechild.xml.

To view all the blogs titles I created a python code called c.py which reads the curl result and lists all titles.

Question 2

Manually classify the first 50 entries, and then classify (using the fisher classifier) the remaining 50 entries. Report the cprob() values for the 50 titles as well. From the title or entry itself, specify the 1-, 2-, or 3-gram that you used for the string to classify. Do not repeat strings; you will have 50 unique strings. For example, in these titles the string used is marked with *s:

Rachel Goswell - "Waves are Universal" (LP Review)
The *Naked and Famous* - "Passive Me, Aggressive You" (LP Review)
Negativland* - "Live at Lewis's, Norfolk VA, November 21, 1992" (concert)
Negativland - "*U2*" (LP Review)

Note how "Negativland" is not repeated as a classification string.

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

Answer Two-

To solve this question and make my life easier I only manually classify the first 10 entries and user fisher classifier the remaining 10 entries.
The same steps is preformed if I was to work on any other number of data.
Note: The strategy to solve this question was based on Correns thoughts, so thank you Corren :)

I got the following counts for the different categories I have:

Family=2
Style=4
Music=2
Food=2

The feature, category count database file is called fc in the girls.db.

I used the existing codes feedfilter.py and docclass.py and only minor changes are made. And the database I created was girls.db.

The following table is the result of the fisher classifier.

Title	Feature	Predicted	Actual	CP
red, green, blue and white	red	music	family	0.0
Thank you + Much love	thank	music	family	0.0
Last Minute Thanksgiving Foodstuffs	thanksgiving	food	food	0.0
12 Days of Giving: Starlight and Santa Clara Valley Medical Center	giving	family	family	0.0
"...the most revolutionary thing a father can do is take care of his children."	revolutionary	family	family	0,0
"Its, I suppose... more adventurous."	adventurous	style	family	0.0
"Be your own best friend"	friend	music	family	0.0
the gift that keeps on sprouting	gift	food	food	0.0
187/100	187/100	food	music	0.0
talking to strangers	strangers	family	family	1.0

Question 3

Assess the performance of your classifier in each of your categories by computing precision and recall. Note that the definitions are slightly different in the context of classification; see:

`http://en.wikipedia.org/wiki/Precision_and_recall#Definition_.28classification_context.29`

Answer Three-

Based on the result data the following is the precision and recall:

Result	Family	Style	Music	food
precision	3/3	0/1	0/3	2/3
Recall	3/7	0/0	0/1	2/2

Question 4

==========
The questions below is for 5 points extra credit
==========

Redo the questions above, but with the extensions on slide 26 and pp. 136–138.

Answer Four-