CS595 Intro to Web Science, Assignment #10

Valentina Neblitt-Jones

December 12, 2013

List of Tables

Listings

List of Figures

Question 1

Choose a blog or newsfeed (or something with an Atom or RSS feed). It should be on a topic or topics or 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 to Question 1

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 to Question 2

Question 3

Answer to 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`

Question 4 - Extra Credit (5 points)

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

Answer to Question 4

Not attempted.

A getFeedList.py

B generatefeedvector.py

C reduceTerms.py

D clusters.py

E Output from getKMeans.py

F Output from getMDS.py