IST736 Text Mining
HW4 (10 points)

Multinomial Naïve Bayes

Format:

Individual homework

Task 1: Illustrate multinomial naïve Bayes algorithm. (4 points)

You will be using a science drawing tool to illustrate how multinomial naïve Bayes works.

Here is the instruction:

Step 1: Go to http://doodlebook.org/.
Step 2: Click "log in" on the top-right corner and use your Google account to log in.
Step 3: Enroll in the Spring 2018 Text Mining class. Then go to assignment "How does Multinomial Naïve Bayes work?"
Step 4: Start to draw by clicking the "Google Drawing" button. After finishing the drawing, publish it online and post the link to the specified location. If any questions on tool use, check out the video tutorial and FAQ on the tool website.

Task 2: Use Weka's multinomial naïve Bayes algorithm to build models to classify the customer reviews by (1) sentiment (positive or negative) and (2) authenticity (true or fake, lie detection). (6 points)

For each of the two classification tasks, combine MNB with different types of word vectors to build the models, and evaluate them using tenfold cross validation methods. Compare the following models' performance. Report confusion matrix and overall accuracy. Also report the precision, recall, and F-measure for each category.

Word vectors:

(1) Unigram Boolean
(2) Unigram term frequency
(3) Unigram normalized term frequency
(4) Unigram Tfidf
(5) The best of the above four representations + stemming
(6) The best of the above four representations + bigrams (no stemming)
(7) The best of the above four representations + bigrams + trigrams (no stemming)

When comparing these models, discuss which text representation is the most effective, and whether stemming improves or hurts the sentiment classification model and the lie detection model. Also, compare the difficulty level of sentiment classification vs. lie detection. Discuss whether you believe computers can detect fake reviews by the words.

Submission:

For Task 1, submit your screen name, so that we can attribute your drawing to you.

For Task 2, submit a Word document in research paper format, similar to previous homework.

Grading rubrics:

Task 1: Accuracy and understandability
Task 2: Same as HW1