University of Colorado at Colorado Springs

Home Work Assignment 2

Due 04-24-2015

---

1 Naïve Bayes Classifier

Write a program in a language of your choice that classifies datasets into two classes. The two classes here are *Charles Dickens* and *Thomas Hardy*.

1.1 Dataset

For Dickens, you can find his writings in text form at `http://www.gutenberg.org/ebooks/author/37`. For Hardy, you can find documents at `http://www.gutenberg.org/ebooks/author/23`. You may be able to find data in other sites as well.

1.2 Things to Do

1. Develop a Naïve Bayes classifier to classify an unseen document. You can use parts of the documents you download for training and parts for testing.
2. Provide results in terms of metrics you think are relevant to this problem.
3. Improve your basic algorithm in any way you want. For example, you may want to reduce the number of features to use, in various ways.
4. Compare the performance of your programs with any other programs or published results you can find.
5. Perform research into the problem of author identification.

1.3 What to Hand in

You will submit a 2+ page paper with a title and your name. Use the IEEE Author style you have been using for the semester project papers you have been writing for the class. In this paper, you will document salient aspects of your program, results you have obtained, improvements you have implemented and relevant research you have conducted.