MSDS 7333: Quantifying the World

Case Study 5: Text Processing and Machine Learning

Instructions

1. Choose your group
2. You can use any platform/package you would like
3. Built in Case Study: Module 10.8 outlines a possible case study for the unit. It is to compare the performance of Naïve Bayes, CART, and one other classification algorithm of your choosing in classifying messages as spam or ham. If you do this option, make sure you describe carefully (but not necessarily mathematically) how each of the classifiers work, what implementation of the algorithms you are using, and how you are comparing their performance.
4. You can alternatively come up with a question of interest regarding making a text-based classifier of Wikipedia articles à la the TMNT example. I’ve included the code “caseStudy.r” for getting a document term matrix from a Wikipedia article. f you do this option, make sure you describe the scientific problem you are trying to solve and the processing you do to get a document term matrix, what methods you are using for supervised (e.g. logistic lasso) or unsupervised (e.g. clustering) learning, and the implementation of the algorithms you are using (e.g. glmnet).
5. Write a report on your analysis, including an abstract, introduction/background, methods, results, conclusions/discussions, and references. I’m not planning on including a rubric for this case study due to its unstructured nature.