

Project 3

This is a Team Project! For this project, please work with the entire class as one collaborative group! Your project should be submitted (as an IPython Notebook via GitHub) by end of day on Sunday, October 25th. The group should present their code and findings in our meetup on Thursday October 29th.

The ability to be an effective member of a virtual team is highly valued in the data science job market.

○

Using any of the three classifiers described in chapter 6 of *Natural Language Processing with Python*, and any features you can think of, build the best name gender classifier you can. Begin by splitting the Names Corpus into three subsets: 500 words for the test set, 500 words for the dev-test set, and the remaining 6900 words for the training set. Then, starting with the example name gender classifier, make incremental improvements. Use the dev-test set to check your progress. Once you are satisfied with your classifier, check its final performance on the test set. How does the performance on the test set compare to the performance on the dev-test set? Is this what you'd expect?

Source: Natural Language Processing with Python, exercise 6.10.2.