CS/INFO 4300 Language and Information, Spring 2015

Assignment 3 - "Tragedy Tomorrow, Comedy Tonight"

Due: Friday, April 10, 5:00pm

This assignment can be completed in groups of 2 to 3. Indicate your groups at least one week before the deadline using CMS. Every group member is expected to know and understand their group submission completely.

In this assignment we will explore the classification problems we have been discussing in class and get more familiar with scikit-learn. We are going to look at text features to in the context of classifying movies into genres based on the text of the dialogs in the script. The data is based on the Cornell Movie Dialogs corpus Many of the questions rely on reading the scikit-learn documentation.

The assignment focuses on vectorizers, support vector machine classification, multi-label evaluation, cross-validation and grid search.

The assignment is structured as an IPython Notebook that you will have to complete and submit via CMS by the due date.

Documentation and tutorials for working with IPython Notebooks are available on the IPython Notebook website.

The bundled ZIP file is available on the course website, and contains:

- This description,
- The IPython Notebook with the assignment,
- An HTML version of the IPython notebook, for reading on other platforms,
- A JSON file with the movie script data.

The zip file is password protected; the password will be made available in class (and it can be obtained by emailing Vlad using your Cornell email). It is the same as for the previous assignments.

Required libraries. These will be useful throughout the course, so it's worth getting accustomed to them:

- numpy
- scikit-learn

You should not use any additional libraries except for the bonus question. If you use additional libraries for the bonus question please be sure to cite them and explain why you used them.

It is OK to take inspiration from the in-class demos (available on the course website); if you adapt any code from there you need to acknowledge it in your comments. Make sure to understand the code: more often than not, the in-class demo code is not directly applicable in the assignment.