You are provided with the following data: a set a training set and test set from lines from characters from scripts of the television soap opera EastEnders scripts (from 2008). These are two comma-separated csv files of the format:

line, character, gender

Using the training data, we are asking you to use the skills learnt throughout the course (in the statistical methods, syntax and semantics lectures) to train a classification function which can go from the lines to the gender of the character first (but not using the character label), and then a program which (but not using the gender label) can predict the speaker of the lines. You can also combine the two classification tasks into one.

Start by doing the preprocessing steps you did in Lab 2 and training a simple multi-label classifier such as SVM on the lines of characters. Then,  use other methods learnt in the class to improve on your classifier by providing different features for the classifier from them. For example, using grammatical features can include POS tags and parse trees, number of special constituents, e.g. NP, VP, PP, dependency relations, e.g. nsubj, mod, det, etc. You can characterize features for  speech styles, e.g. by using the presence of pauses  and, of course, the typical vocabulary or dialect of speakers can also be taken into account.   The semantic content of the utterances can be  used as features, for example by employing  a Named Entity Recognizer or Word2Vec vectors (more to come in the Semantics class) as features;  the latter will be discussed in the proceeding lectures on semantics.  Finally, modelling the problem as a sequence classification task using a CRF or HMM tagger can also be explored to get possible improvements on the performance of your classifier.

The deadline is**the last day of term, Friday December 13th 10am**. The project must be done individually, though you can talk to each other for support and sharing of libraries. For submission, we expect a**short report of what you did (max 4 pages of 11pt font**) which should include a summary of your results, together with the **code you write (in Python or iPython)** including an explanation of how the program works. **We will run plagiarism detection, so if anyone has copied anyone else, the consequences are serious.**

**These are the conditions on which you have access to the scripts:**

* The data is supplied under the signed agreement signed between the BBC and QMUL
* The data is supplied for the purpose of training data for the agreed NLP Masters module at QMUL in November/December 2019, January 2020.  If the data is to be used for any other purpose this has to be agreed in writing.
* If there is a publication arising from use of this data the BBC should be provided with a chance to review the publication with respect to BBC sensitivities around the data
* The data has to be stored securely, only passed on by Julian Hough to specified students. Each student must agree to the following terms:
  + The data must not be passed to anyone else or placed online for others to access (e.g. **not on a website , Github, etc**)
  + The results of any processing of this data may not be placed online for public access
  + The data must be deleted after the end of the project.

**Marking Scheme:**  
Only use the training file for training, and report the results on the test file only. Report raw accuracy, precision, recall,  and F-Score of  **gender** and **character identification** on the test data. You should use cross-validation on the training data to develop your model (or split it into main training data and heldout data) and then use the test data once you have settled on the best model.  The details of the marking is as follows:

1. 10% for preprocessing.
2. 30% for feature engineering.
3. 30% for a sound evaluation and performance of the classifier (including baselines if possible).
4. 30% for the quality of the report and README as to how the code works.