Domain

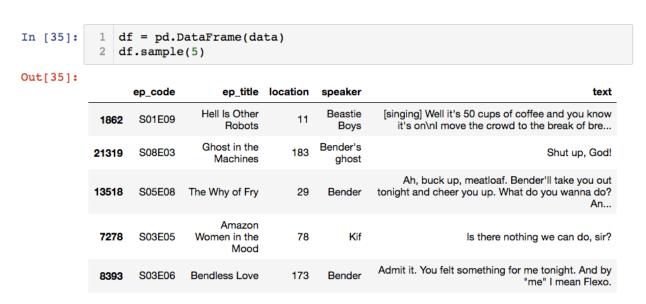
Fictional characters, especially cartoon characters, often have characteristic ways of speaking. In this project I will attempt to train an algorithm to predict the speaker of an unknown piece of dialogue by training it on tagged examples.

In the MVP version, I'll focus on the main cast (5-8 characters) of a single show, Futurama. I've chosen Futurama because I'm (very) familiar with it, and I know that its characters speak in characteristic ways (e.g. Farnsworth is rambling, bitter, and uses lots of science(y) words; Kif is servile and dots on his smizmar Amy; Bender is sociopathic and likes to talk about theft and alcohol).

Data

The minimal dataset will be everything that's been said by these characters over the course of 10 seasons and a hand-full of movies. Time permitting, I'll train the algorithm on character dialogue from other works to compare separability.

With the exception of a few episodes with oddly formatted transcripts, I managed to gather most of the this data on Wednesday by scraping <u>TheInfoSphere.org</u>. This comes to 24,034 lines of dialogue and is in the format of the below DataFrame.



Known unknowns

- How separable the dialogue is theoretically, given a perfect model.
- Whether this is enough dialogue to train a model.
- The number of words needed to make a piece of dialogue characteristic.