Who speak it?

This project aims to learn character model from the animated sitcom – The Simpsons. Then for an input sentence, we predict who speak it.

The dataset is form Kaggle.com - The Simpsons by the Data. It contains the text spoken during each episode (including details about which character said it and where).

Main work including:

* Analysis and preprocess the dataset. Eg calculate tf-idf.
* Model the scripts with respect to the speakers using two methods (Latent Dirichlet Allocation and Latent Semantic Analysis)
* Predict for a new input sentence, who are most likely speak it. Show the prediction result in terms of credible interval.
* Compare the behavior of the two methods.

Dataset: <https://www.kaggle.com/wcukierski/the-simpsons-by-the-data>