☰ Item Navigation

Clustering text data with Gaussian mixtures

In a previous assignment, we explored K-means clustering for a high-dimensional Wikipedia dataset. We can also model this data with a mixture of Gaussians, though with increasing dimension we run into several important problems associated with using a full covariance matrix for each component.

In this section, we will use an EM implementation to fit a Gaussian mixture model with **diagonal** covariances to a subset of the Wikipedia dataset.

If you are using Turi Create

An IPython Notebook has been provided below to you for this assignment. This notebook contains the instructions, quiz questions and partially-completed code for you to use as well as some cells to test your code.

- Download the Wikipedia people dataset in SFrame format:

people_wiki.sframe

ZIP File

Download file

- Download the companion IPython notebook:

CLU04-NB02.ipynb

ZIP File

Download file

- Download a collection of helper functions:

em_utilities.py

ZIP File

Download file

- Save all the files in the same directory (where you are calling IPython notebook from) and unzip the data file.

Open the companion IPython notebook and follow the instructions in the notebook. The