☰ Item Navigation

Modeling text topics with Latent Dirichlet Allocation

In many cases, it is good to think of data as belonging to more than one cluster or category. For example, if we have a model for text data that includes both "Politics" and "World News" categories, then an article about a recent meeting of the United Nations should have membership in both categories rather than being forced into just one.

With this in mind, we will use Turi Create tools to fit an LDA model to a corpus of Wikipedia articles and examine the results to analyze the impact of a mixed membership approach.

In this assignment you will

- apply standard preprocessing techniques on Wikipedia text data
- use Turi Create to fit a Latent Dirichlet allocation (LDA) model
- explore and interpret the results, including topic keywords and topic assignments for a document

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 pretrained models:

topic_models

ZIP File

Download file

- Download the companion IPython notebook: