UCLA Extension Machine Learning: Recipe Naming Using Text Analysis

March 17, 2020

Concept and Data Source

Concept: Given a recipe (consisting of a list of ingredients and a set of corresponding instruction set), produce an accurate title for the recipe.

Main data source:

- 250,000 recipes scraped from various websites (foodnetwork.com, epicurious.com, allrecipes.com). The program that created this set is MIT lisenced and was created by Ryan T. Lee (github: rtlee9), and the dataset used is ODC lisenced.
- Dataset is a json file with recipe key and corresponding title, ingredient list, and instruction set.

Basic concept: vectorizing text

- In order to do any form of efficient computational analysis, the text that you want to analyze must be turned into a vector. Any part of text, such as letters,words, sentences, paragraphs, or even entire manuscripts, can be converted into vectors.

- The typical starting point for text-based machine learning is to look at the words in a text, and how those words correspond to features that you are trying to predict.