Recipe Naming Using Text Analysis

Elliot Lehman

UCLA Extension

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

- 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.
- In order to do any form of efficient computational analysis, the text that you want to analyze must be turned into a vector.
- There exist many different ways to vectorize text. Methods to vectorize text based on letters, words, relationships between words, proximity between words, etc.

- The goal is to input a *recipe vector* and output a *title vector*.

  Note: We will treat these as different vector spaces so, in general, a *recipe vector* does not equal a *title vector*.

A toy example of text vectorization

Given the below dataset, what would our *recipe vector* and *title vector* look like?

Recipe		Title
bread	avocado	avocado toast
lettuce	avocado	salad

1. Define your basis vectors.
   - There are 3 unique words in the Recipe section, meaning that any recipe using ingredients in this data set can be represented by a vector of size 3.
   - The word “avocado” would be represented by the vector $(1, 0, 0)_r$, “bread” by $(0, 1, 0)_r$, and lettuce by $(0, 0, 1)_r$
2. Vectors of recipes.
   - Now that basis vectors are defined, we can add them to get recipes.
   - The phrase “avocado lettuce” is therefore represented by $(1, 0, 1)_r$.
   - The phrase “avocado bread lettuce” is the vector $(1, 1, 1)_r$.

We have vectors... now what?

EDA!!!

Now that we have transformed a word based data set into a set of vectors and numbers, we can start to really see what our data looks like.
- highest frequency words in recipes - highest frequency words in titles (good to know so we can normalize) - highest frequency nouns, verbs, adjectives, etc. - most related words - perhaps the most important.

eda continued

Ingredient word	Title word
cup	and
cup	with
teaspoon	with
chopped	and
tablespoons	and
tablespoons	with
cups	and
teaspoon	and
fresh	and
chopped	with

Future improvements

- Expand vocabulary, the program can only currently title recipes using words seen in other titles
-
-
-