Ingredient Identification and Extraction

Siva Kailash S 2016103593
Ben Stewart S 2016103513

Code : https://github.com/sbenstewart/nlp

Abstract

A system that automatically classifies the recipe (example French, Italian, etc..) and Extracts the Ingredients used in it .

Project is in two halves

Identification part

- This is identify the recipe and categorize them according to their cuisine (bayesian classification with empirical evaluation, linear regression ,Bokeh model ..)
- Recipe is the input and cuisine is output

Ingredient extraction part - this has 2 parts

- First part is identifying the main part of the ingredient (pepper,flour etc..) (Peter Norvig's method)
- Second part is to extract additional info about the ingredient used . This is done by assigning weights to the words near to the ingredient identified (sort of a n - gram problem) and identifying the adjunct (noun describer) that is used in the sentence (NLU).

Dataset :

Initially we tried scraping data from Hebbar's Kitchen (from recipes), which didn't work because the preprocessing part was a huge barricade for us. We can't classify the data into categories and the information the site provided wasn't that perfect for our project, so we planned to work with much better dataset.
Apparently we got our desired dataset which was submitted as a result for a competition named what's-cooking. It had sufficient data (ingredients) with pre classified labels according to their cuisine.
Link :
https://github.com/sbenstewart/nlp/tree/master/whats-cooking

Modules :

Pre-processing : This involves collecting data from the dataset, loading the data, removing duplicate tuples, white spaces and plotting the length of each ingredient (for a particular cuisine).

Feature Extraction : This involves figuring out how we can actually train the data. It contains finding the count of unique cuisines and their recipes and the unique ingredients for that cuisine and how many ingredients used and other details that is feeded into the kernel.

Adding External label Glossaries : After finding the labels, we need to feed those into our kernel as a basis for classification.

Identifying Part

Popular Items in each Cuisine

This lists the popular ingredients for each cuisine.For instance we can see that the top1 ingredient for each cuisine is a salty ingredient. This salty ingredient allows us to group the cuisines already: (look in code line no 10)

salt is the standard for most cuisines soy sauce is number one for chinese, japanese and korean cuisines fish sauce is number one for thai and vietnamese cuisines Another things that is easily seen from this table is that many ingredients have more than one name:

garlic cloves, garlic olive oil, extra-virgin olive oil ... Judging from this table, it can be interesting to see which ingredients among the top 10 ingredients are highly specific for a certain cuisine. A way to do this is to simply count the number of times an ingredient appears in a given cuisine and divide by the total number of recipes.

Subprocessing

This includes

- Concatenation - Get all the ingredients as one column
- Checking presence of ingredient in recipe
- Group based on presence of ingredient
- And Normalization

**Logistic regression - **The regression model that is being used in the training of dataset to get the desired result. Line 64 in code.

We got an avg precision value of 85% for a given cuisine.

The rank for each ingredient is found and updated via the Bokeh model and Bokeh plotting .

The Extraction Part Extracting the required ingredient from a sentence.

Inconsistencies

- special symbols that are not relevant (trademark, copyright, ...)
- brand names, for instance KRAFT, Pillsbury, Hidden Valley
- sentences of english words instead of ingredients i can't believe it' not butter!
- spelling errors burgundi wine sauc should actually be spelled burgundy wine sauce
- quantities of ingredients that shouldn't be there: 2 1/2 to 3 lb. chicken

Implementation of Model

Peter Norvig's approach on spell correcting approach to this problem.

Our error model will be that we can only delete words from our ingredient. So what we need is to generate a list of possible ingredients based on our original ingredient by subtracting words. Also, we will assume that word order doesn't matter, so we can represent an ingredient by a set.

Modelling ingredients using sets

Ingredients contain a fixed number of words. We will therefore model them as frozensets of words. This will allow us to manipulate them more easily in the remaining document.
(line 87)

Inference

- some simplifications are not the ones expected i can't believe it's not butter!® all purpose sticks, reduced sodium reduced fat cream of mushroom soup
- some get simplified too much: skinless and boneless chicken breast fillet becomes chicken, reduced sodium italian style stewed tomatoes becomes tomatoes

Building a slightly more elaborate model

The only thing we have to change is the way our candidates function works. Instead of generating all possibilities it should return only the ones that exist in our vocabulary of recipes with the least possible number of modifications. Here the modifications can be thought of leaving out a given number of words, by increasing the number of words left out. (line 96)

Conclusion

This result is interesting: the refined labels given here seem much better than the originals. However, there are still problems with the new ingredients. In particular, special symbols pollute the data.