Recipes data extraction

Maciej Chrabąszcz
Aleksander Kozłowski

Goal

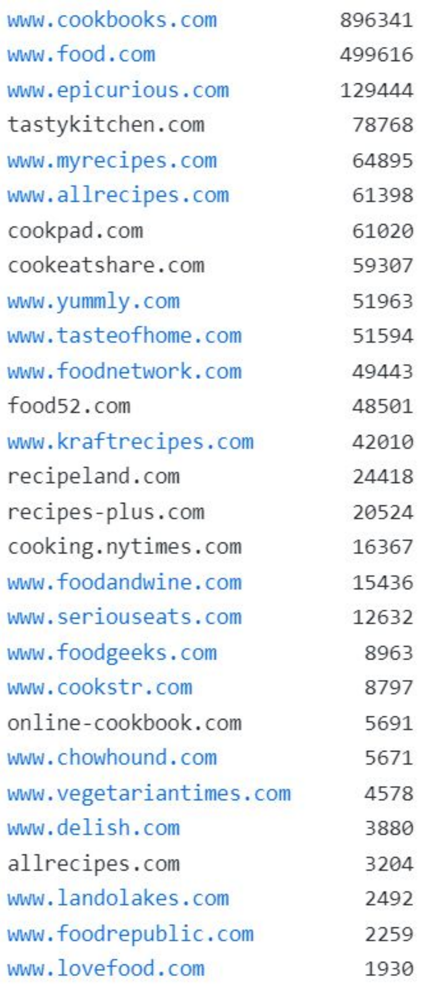
Enrich the RecipeNLG dataset

Dataset description

The RecipeNLG dataset contains 2231142 cooking recipes (>2 million). Including recipes ingredients and instruction scraped from the internet. It's main goal is to allow creation of models which generate recipes conditioned on products.

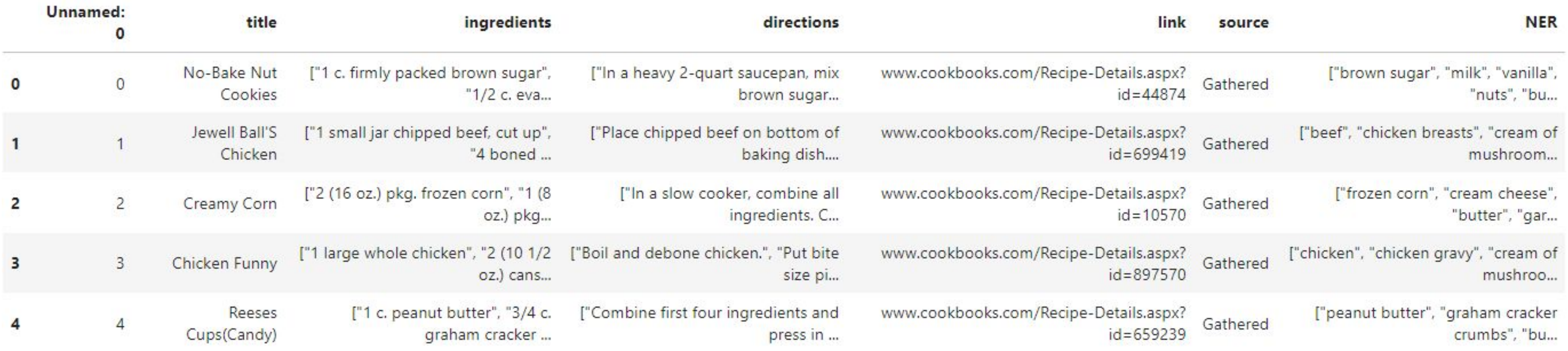
Bień, Michał, et al. "RecipeNLG: A cooking recipes dataset for semi-structured text generation."

Websites used to create dataset

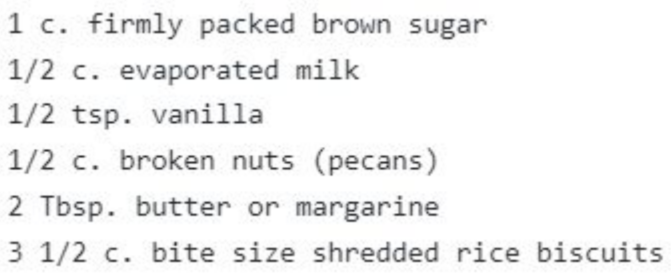


```
www.cookbooks.com          896341
www.food.com               499616
www.epicurious.com         129444
tastykitchen.com            78768
www.myrecipes.com           64895
www.allrecipes.com          61398
cookpad.com                 61020
cookeatshare.com            59307
www.yummly.com              51963
www.tasteofhome.com         51594
www.foodnetwork.com         49443
food52.com                  48501
www.kraftrecipes.com        42010
recipeland.com              24418
recipes-plus.com            20524
cooking.nytimes.com         16367
www.foodandwine.com         15436
www.seriouseats.com         12632
www.foodgeeks.com            8963
www.cookstr.com              8797
online-cookbook.com          5691
www.chowhound.com            5671
www.vegetariantimes.com      4578
www.delish.com               3880
allrecipes.com               3204
www.landolakes.com           2492
www.foodrepublic.com         2259
www.lovefood.com             1930
```

Example sample



	Unnamed: 0	title	ingredients	directions	link	source	NER
0	0	No-Bake Nut Cookies	["1 c. firmly packed brown sugar", "1/2 c. eva...	["In a heavy 2-quart saucepan, mix brown sugar...	www.cookbooks.com/Recipe-Details.aspx?id=44874	Gathered	["brown sugar", "milk", "vanilla", "nuts", "bu...
1	1	Jewell Ball'S Chicken	["1 small jar chipped beef, cut up", "4 boned ...	["Place chipped beef on bottom of baking dish....	www.cookbooks.com/Recipe-Details.aspx?id=699419	Gathered	["beef", "chicken breasts", "cream of mushroom...
2	2	Creamy Corn	["2 (16 oz.) pkg. frozen corn", "1 (8 oz.) pkg...	["In a slow cooker, combine all ingredients. C...	www.cookbooks.com/Recipe-Details.aspx?id=10570	Gathered	["frozen corn", "cream cheese", "butter", "gar...
3	3	Chicken Funny	["1 large whole chicken", "2 (10 1/2 oz.) cans...	["Boil and debone chicken.", "Put bite size pi...	www.cookbooks.com/Recipe-Details.aspx?id=897570	Gathered	["chicken", "chicken gravy", "cream of mushroo...
4	4	Reeses Cups(Candy)	["1 c. peanut butter", "3/4 c. graham cracker ...	["Combine first four ingredients and press in ...	www.cookbooks.com/Recipe-Details.aspx?id=659239	Gathered	["peanut butter", "graham cracker crumbs", "bu...



```
1 c. firmly packed brown sugar
1/2 c. evaporated milk
1/2 tsp. vanilla
1/2 c. broken nuts (pecans)
2 Tbsp. butter or margarine
3 1/2 c. bite size shredded rice biscuits
```

Planned workflow

- Add nutritional information
- Add dietary tags
- Product names and quantities (NER)
- Prepare strong foundations for 2nd project

Nutritional information

In order to enrich the data set with nutritional information we plan to use data provided by U.S. Department of Agriculture (https://fdc.nal.usda.gov). Their API contains information about micro and macro-nutrients as well as food category.

Dietary tags

In order to enrich the dataset we want to add dietary tags for example “vegan”, “dairy free” etc.

To achieve that we will train models on the dataset found on kaggle which contains recipe tags (including dietary).

Britto, L., Pacífico, L., Oliveira, E., & Ludermir, T. (2020, October). A cooking recipe multi-label classification approach for food restriction identification.

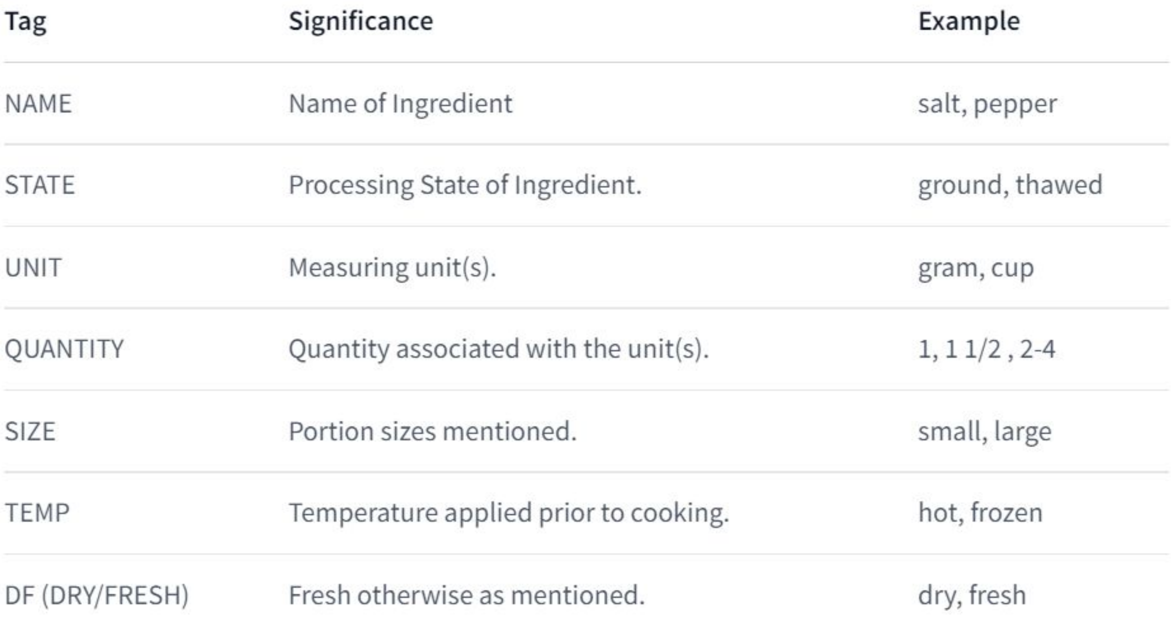
NER

To extract information about quantity and names of products used in recipes we will use NER model on recipes ingredients.
Pretrained model was introduced "A Named Entity Based Approach to Model Recipes".

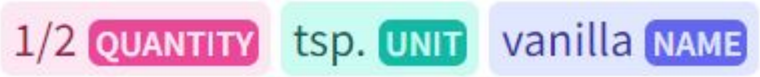
If this model won't be sufficient we can train our own NER model on "Tasteset".

- Diwan, N., Batra, D., & Bagler, G. (2020, April). A named entity based approach to model recipes.
- Wróblewska, A., Kaliska, A., Pawłowski, M., Wiśniewski, D., Sosnowski, W., & Ławrynowicz, A. (2022). TASTEset--Recipe Dataset and Food Entities Recognition Benchmark.

Pretrained NER model



Tag	Significance	Example
NAME	Name of Ingredient	salt, pepper
STATE	Processing State of Ingredient.	ground, thawed
UNIT	Measuring unit(s).	gram, cup
QUANTITY	Quantity associated with the unit(s).	1, 1 1/2 , 2-4
SIZE	Portion sizes mentioned.	small, large
TEMP	Temperature applied prior to cooking.	hot, frozen
DF (DRY/FRESH)	Fresh otherwise as mentioned.	dry, fresh



1/2 QUANTITY tsp. UNIT vanilla NAME

Future work ideas

- Recipe recommender based on tags e.g. ‘vegan’
- Language model
- Automatic nutritional information about given recipe

Time for your questions!