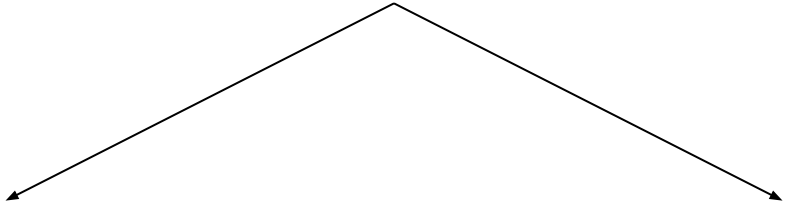
Recipes Data Extraction

Maciej Chrabąszcz
Aleksander Kozłowski

Our goals



Dietary tags classification

Nutritional values extraction

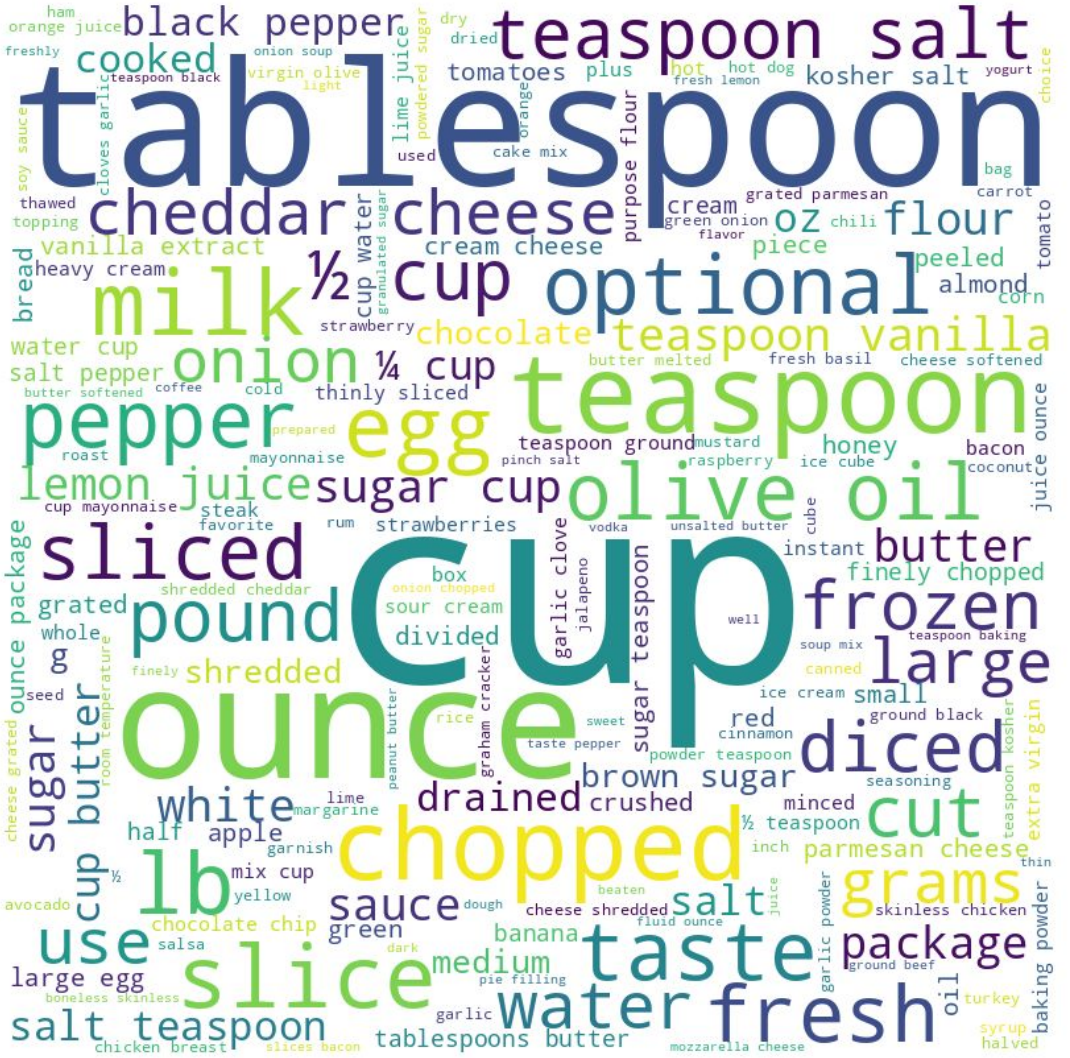
EDA

We looked into, tried, and tested several datasets:

- RecipeNLG
- TASTEset
- Food.com based found on kaggle

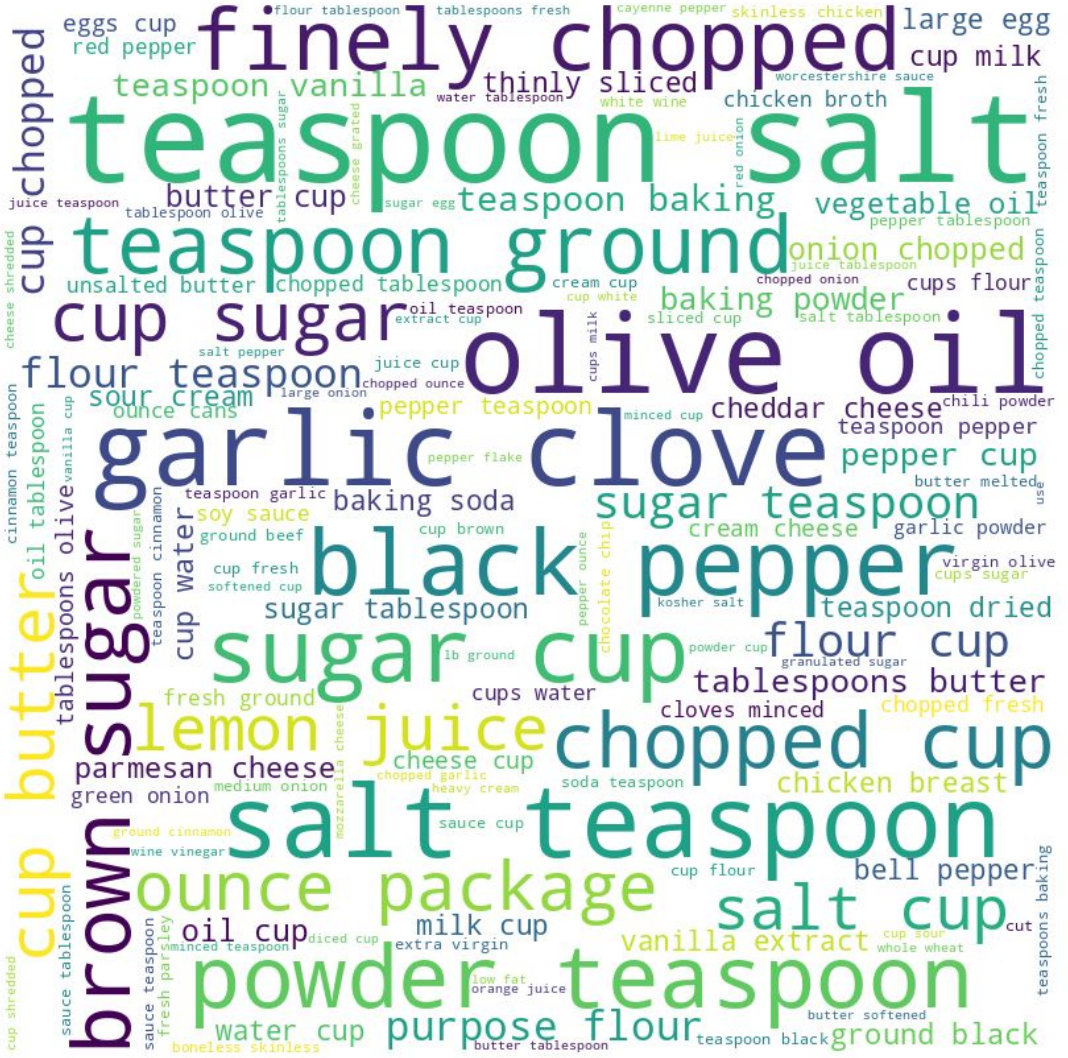
TASTEset

Word Cloud



Food.com

Word Cloud



RecipeNLG

Word Cloud

We conclude that RecipeNLG dataset is far more noisier than we thought compared to other datasets.

Parsing USDA API


```
{'totalHits': 181156,
 'currentPage': 1,
 'totalPages': 3624,
 'pageList': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
 'foodSearchCriteria': {'query': 'water',
  'generalSearchInput': 'water',
  'pageNumber': 1,
  'numberOfResultsPerPage': 50,
  'pageSize': 50,
  'requireAllWords': False},
 'foods': [{'fdcId': 2047057,
   'description': 'WATER',
   'lowercaseDescription': 'water',
   'dataType': 'Branded',
   'gtinUpc': '041270037556',
   'publishedDate': '2021-10-28',
   'brandOwner': 'Iga, Inc.',
   'brandName': 'IGA',
   'ingredients': 'CARBONATED WATER, HIGH FRUCTOSE CORN SYRUP, CARAMEL COLOR, PHOSPHORIC ACID, NATURAL AND ARTIFICIAL FLAVORS, SODIUM BENZOATE (PRESERVATIVE), CAFFEINE.',
   'marketCountry': 'United States',
   'foodCategory': 'Soda',
   'modifiedDate': '2018-01-19',
   'dataSource': 'LI',
   'packageWeight': '67.6 fl oz/2 L',
   'servingSizeUnit': 'ml',
   'servingSize': 240.0,
   'tradeChannels': ['NO_TRADE_CHANNEL'],
   'allHighlightFields': '<b>Ingredients</b>: CARBONATED <em>WATER</em>, HIGH FRUCTOSE CORN SYRUP, CARAMEL COLOR, PHOSPHORIC ACID, NATURAL AND ARTIFICIAL FLAVORS, SODIUM BENZOATE (PRESERVATIVE),
CAFFEINE.',
   'score': 885.38617,
   'microbes': [],
```

]:

	product_name	carbohydrates_100g	energy-kcal_100g	fat_100g	proteins_100g
1862566	Green Onion	53.0	526.0	31.0	6.1

Ingredient	Description	Fat	Fat_unit	Protein	Protein_unit	Carbohydrates	Carbohydrates_	Energy	Energy_unit
apple	APPLE	0	G	0	G	11.7	G	46	KCAL
beef	BEEF	5.88	G	20	G	0	G	129	KCAL
green onion	Onions, green, r	0.19	G	1.83	G	7.34	G	32	KCAL
chicken breast	CHICKEN BREAS	8.93	G	14.3	G	3.57	G	143	KCAL
water	WATER	0	G	0	G	10.8	G	42	KCAL
bell pepper	Peppers, bell, gr	0.11	G	0.72	G	4.78	G		

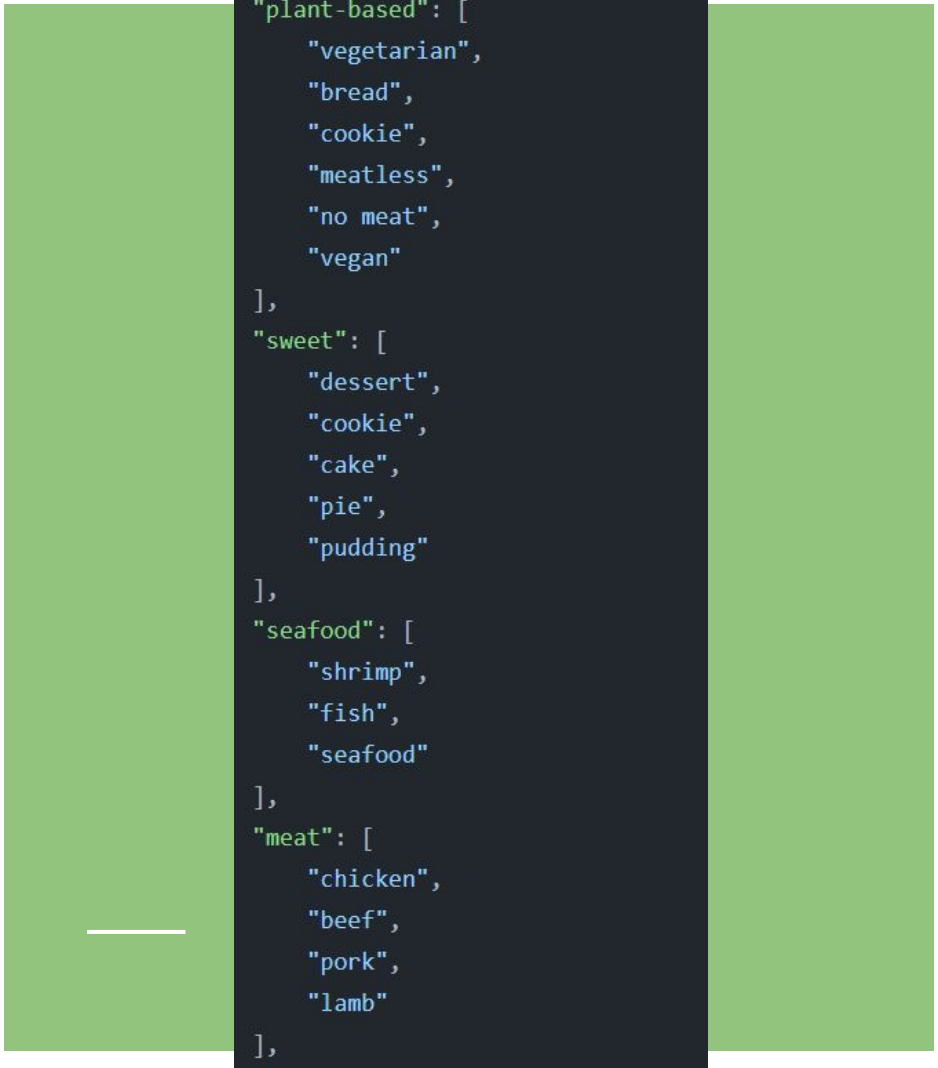
	product_name	carbohydrates_100g	energy-kcal_100g	fat_100g	proteins_100g
62985	apple	11.000000	47.000000	0.0	0.000000
217264	apple	13.768116	50.724638	0.0	0.289855
1887166	apple	12.000000	54.300000	0.1	0.800000
2596078	apple	14.200000	61.000000	0.0	0.200000

Recipes Data Extraction

Problems with parsing USDA API

- Lack of search engine based on popularity (doable e.g. Google)
- Noisy dataset (RecipeNLG) yields 80k+ distinct ingredients
- Rate limits – 1000 requests per hour

Creating classes from tags



```
"plant-based": [
    "vegetarian",
    "bread",
    "cookie",
    "meatless",
    "no meat",
    "vegan"
],
"sweet": [
    "dessert",
    "cookie",
    "cake",
    "pie",
    "pudding"
],
"seafood": [
    "shrimp",
    "fish",
    "seafood"
],
"meat": [
    "chicken",
    "beef",
    "pork",
    "lamb"
],
```

Dietary tags classification on ingredients list

One of our ideas was to use ***pretrained language models*** on ingredients list text to classify dietary tags.

Example input:

“

4 cups water
1 cup uncooked old fashion grits
1 teaspoon salt
4 ounces shredded cheddar cheese
1 -2 clove garlic, minced
1 tablespoon olive oil

”

Classification using only products

Because of failure trying to use language models we decided to create **TF-IDF** representation using only products names.

Example input
“
water
grits
salt
cheddar cheese
garlic
olive oil
“

Results

low-carb

	precision	recall	f1-score	support
False	0.94	0.74	0.82	74017
True	0.49	0.84	0.61	22169
accuracy			0.76	96186
macro avg	0.71	0.79	0.72	96186
weighted avg	0.83	0.76	0.78	96186

seafood

	precision	recall	f1-score	support
False	1.00	0.99	0.99	88352
True	0.90	0.97	0.93	7834
accuracy			0.99	96186
macro avg	0.95	0.98	0.96	96186
weighted avg	0.99	0.99	0.99	96186

nuts

	precision	recall	f1-score	support
False	0.99	0.83	0.90	90088
True	0.25	0.84	0.39	6098
accuracy			0.83	96186
macro avg	0.62	0.84	0.65	96186
weighted avg	0.94	0.83	0.87	96186

Future work ideas

- Use entity linking for better API results extraction
- Create rules which lower common NER errors

Thank you for your attention