



A Network Tour of Cooking



Overview

Introduction to the Dataset

Cleaning

Data Exploration and Graph Creation

Making of Recipes

Final Considerations

Presentation of the Dataset

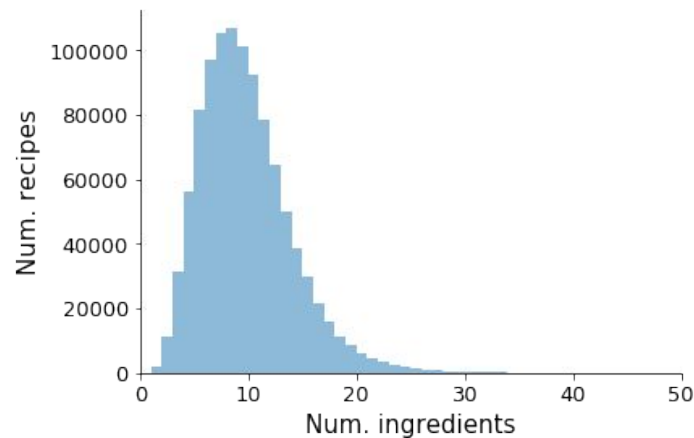
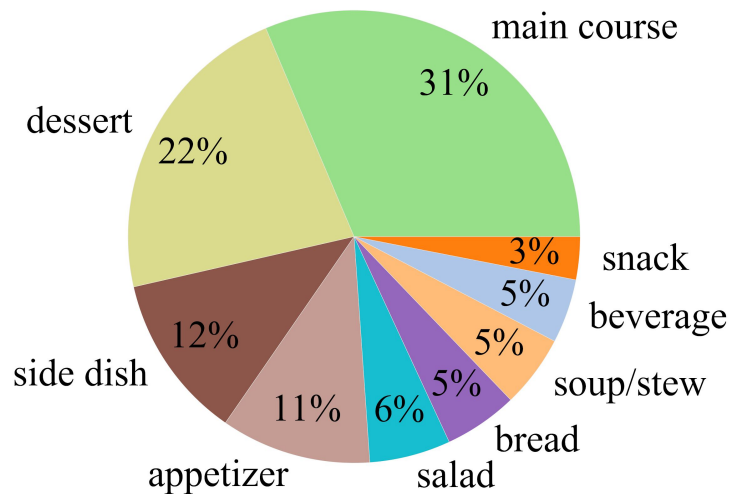
Recipe1M+ is a large-scale Dataset

1M structured (sometimes) cooking recipes

13M images (we don't use)

Made by scraping many different websites

Some statistics



Cleaning of the dataset

Examples of Recipe: Beer Simmered Pork Chops

```
[{'text': '4 pork chops'},  
 {'text': '1-1/2 cans mushroom soup'},  
 {'text': '3 oz (84 grm). can of mushrooms'},  
 {'text': '1/2 cup (125 ml) beer'},  
 {'text': '2 Tbsp (30 ml). of pkg. dry onion soup mix'}]
```

Cleaned

```
['pork chop', 'mushroom soup', 'mushroom', 'beer', 'onion soup']
```

Cleaning of the dataset

Hearty Minestrone Soup

```
[{'text': '1 12 lbs ground sweet Italian sausage (remove from casing if using links)'},  
{ 'text': '1 onion, chopped'},  
{ 'text': '2 tablespoons olive oil'},  
{ 'text': '2 -3 tablespoons fresh basil'},  
{ 'text': '3 fresh garlic cloves, minced'},  
{ 'text': '2 cups carrots, peeled and thick sliced'},  
{ 'text': '2 small zucchini, halved then thick sliced'},  
{ 'text': '4 -8 cups chicken broth'},  
{ 'text': '24 ounces diced tomatoes (drained)'},  
{ 'text': '3 cups shredded cabbage'},  
{ 'text': '12 teaspoon salt'},  
{ 'text': '12 teaspoon pepper'},  
{ 'text': '1 (16 ounce) can great northern beans, undrained'}]
```

Cleaned

```
[ 'ground sausage',  
  'onion',  
  'olive oil',  
  'basil',  
  'carrot',  
  'zucchini',  
  'chicken broth',  
  'tomato',  
  'cabbage',  
  'salt',  
  'pepper',  
  'bean']
```

Cleaning of the dataset

Fatburger Banana Shake

```
[{'text': '2 cups premium vanilla ice cream'},  
 {'text': '3 tablespoons half-and-half (you can use milk)'},  
 {'text': '2 teaspoons instant banana pudding mix, powder'}]
```

Cleaned

```
['vanilla ice cream', 'half', 'banana powder']
```

Salty-Sweet Popcorn

```
[{'text': '6 cups fresh popped popcorn'},  
 {'text': '1 teaspoon salt'},  
 {'text': '2 tablespoons unsalted butter, melted'},  
 {'text': '2 teaspoons confectioners' sugar'}]
```

Cleaned

```
['salt', 'butter', 'confectioner sugar']
```

Cleaning of the dataset

Methodology

Sampling

NLP processing (POS tagging, Stop Words, Digits, “-less” , “-ing”)

Inconsistencies

Different languages

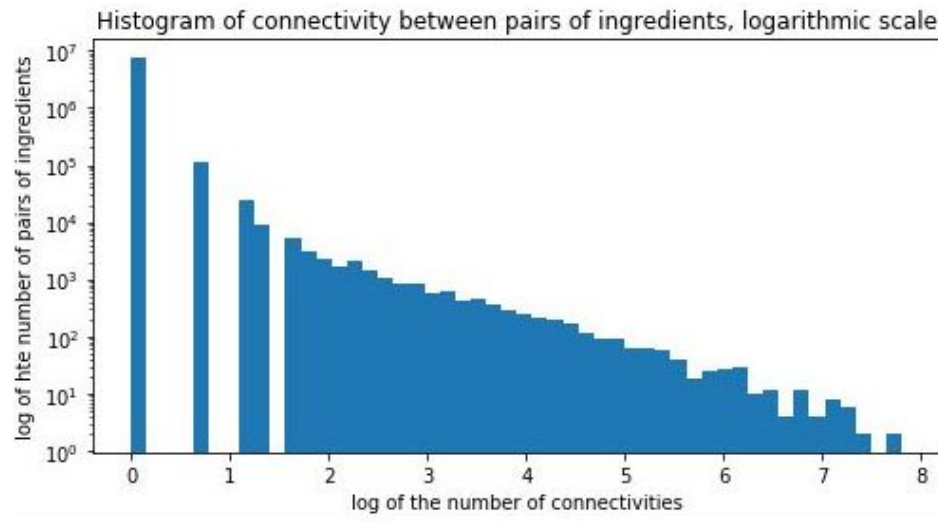
Typos

Brands
EPFL

Graph exploration

Hubs, most represented ingredients,
best pairs of ingredients,

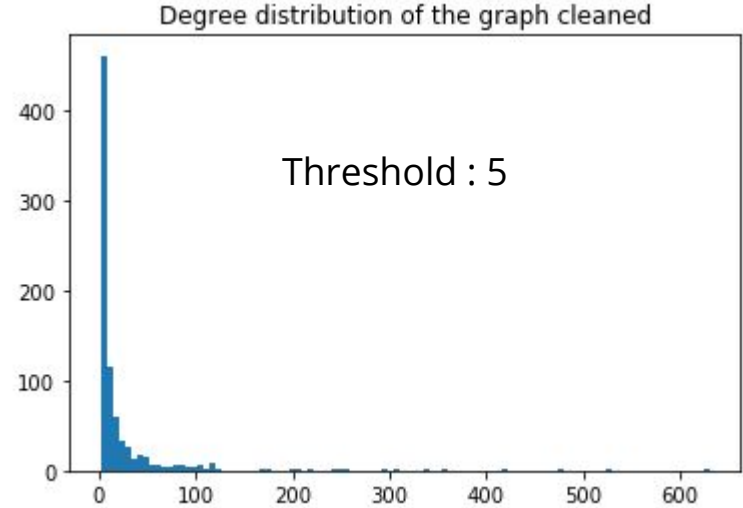
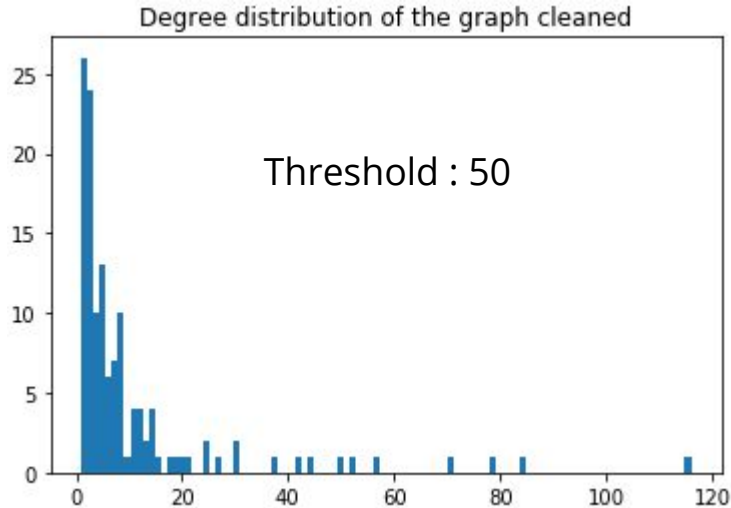
Rank	Ingredient 1	Ingredient 2	number of associations
1	pepper	salt	2448
2	onion	salt	1579
3	salt	sugar	1531
4	onion	pepper	1529
5	salt	flour	1383
5	salt	egg	1300
6	egg	sugar	1269
7	butter	salt	1240
8	flour	sugar	1216
9	egg	flour	1090

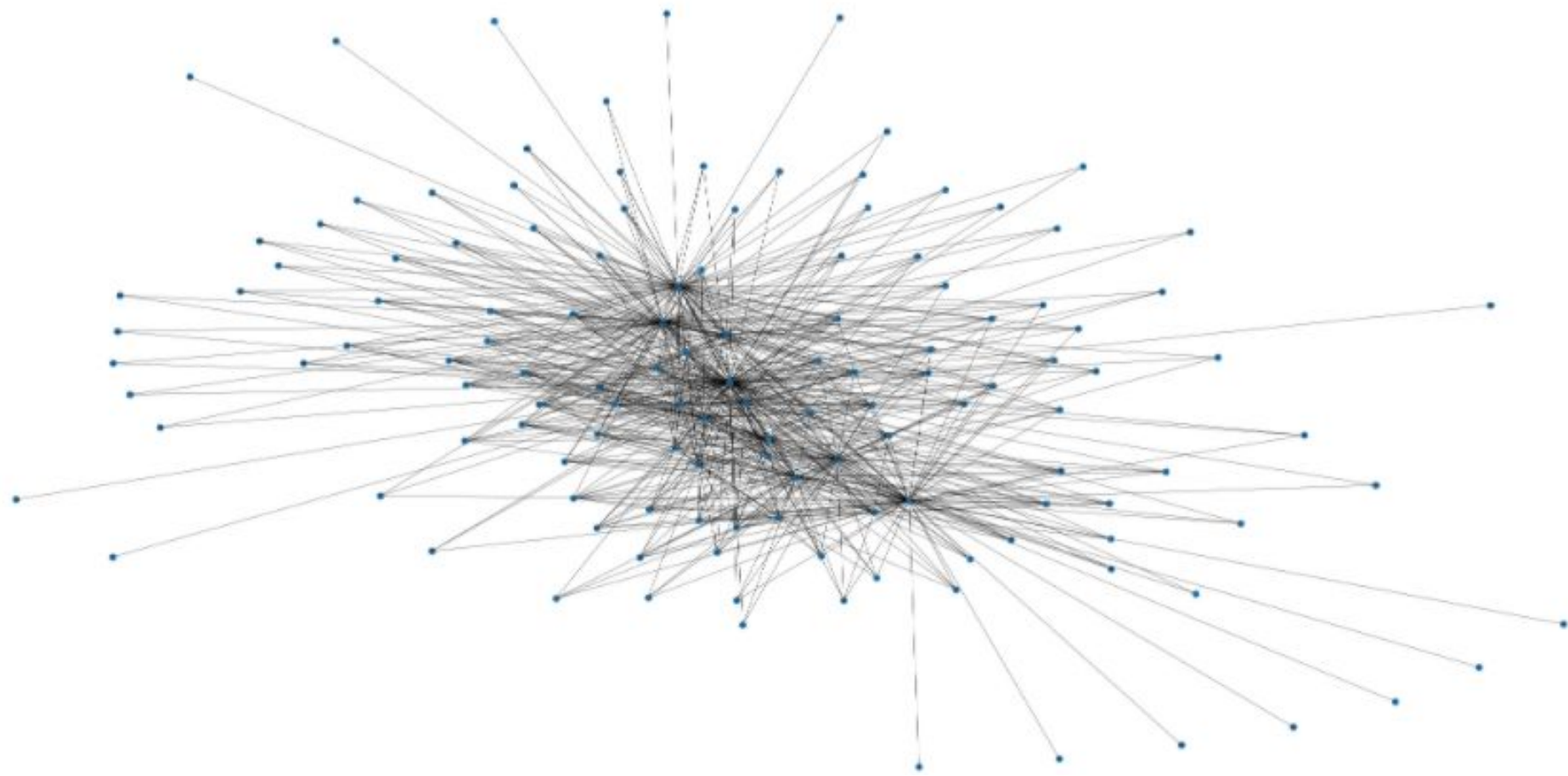


Graph exploration

Creation of the graph :

- thresholding
- pruning : delete 0-degree ingredients





Graph exploration

Playing with threshold

Threshold	50	5
Mean connectivity	0,52	6,42
Nb of nodes after pruning (from 2754)	131	818
Mean connectivity after pruning	10,92	21,60
Mean connectivity (removing 12 hubs)	1,50	11,08
Loss of connectivity (%)	86,30	48,68

Large threshold :



Less noise, less outliers + the networks of ingredients are more coherent, the recipes created will be more realistic (following section)



Overdependence on the 12 hubs : we miss common and good-sense associations, which is not good for the following section : creation of recipes

Creation of recipes using the graph

Algorithm :

- A first ingredient is selected randomly, and put in the list of ingredients
- All it's neighbours (Connected ingredients) are added to the list of neighbour ingredient
- While the required number of ingredients is not met:
 - One of the ingredients in the neighbour list is added to the list of ingredients
 - The neighbours of this last ingredients are added in the list of neighbours they are neither in the recipe or the neighbours list.

Creation of recipes using the graph

Variations of the algorithm :

- Possible to select the first ingredient
- Possible to select more than one ingredient. Dijkstra algorithm is used to find a path between the ingredients, and then the recipe is completed with neighbours ingredients

Examples of created recipes

- 'bamboo', 'baby bella mushroom', 'apple slice', 'avocado'
- 'chorizo', 'bone pork chop', 'cheddar cheese garnish', 'cornichon garnish', 'beef base', 'bosc pear', 'cheese herb', 'baby carrot'
- 'baby spinach', 'avocado safeway', 'almond extract', 'baby carrot', 'baby corn'

Evaluation of our recipes

2 Approaches to evaluate the originality of the created recipes :

- Measure of the minimum euclidean distance between the created recipe and recipes existing in the dataset
- Count the number of ingredients in common
 - Issue : Most of the ingredients are removed by the cleaning, making these methods irrelevant.

Conclusion

- Cleaning of the dataset difficult because of it's lack of consistency
- Bias in the results due to the scrapping : simple words as salt or pepper are easier to scrap than others.
- Structure to clean better the dataset ready to use, just need to add stop words while using the dataset.

Thank you