Skip The Pizza

Analyzing the ingredients of 29k recipes

July 27, 2012 Uncategorized

I love cooking. In fact I'm trying to cook as often as possible. Another passion of mine – which happens to be as well my job – is software engineering. The other day i asked myself if and how I could combine these two. I came up with an idea about a little web application which is since then my side project on rainy weekends – aka SkipThePizza. I will write about the application later on, first i'll stick to what i did so far and the results i got.

This and the following blog post(s) will describe how i analyzed the ingredients of 29200 recipes and what i learned from that. If you like what I'm doing please share, follow or comment. Feedback is highly appreciated.

What are we looking for?

A main feature of the upcoming application is a recommendation engine for ingredients. The algorithm should be able to propose ingredients based on popularity and based on which ingredients we already chose. Technically it comes down to the following questions:

- **How many ingredients does a typical recipe consist of?**
- **Which are the most frequent ingredients?**
- **Which are the most frequent combinations of ingredients?**
- **Given a set of ingredients, which other ingredients go probably along with it?**

If you think this sounds familiar to you, maybe you already read articles about n-grams related to other domains. Traditionally, n-grams are used to describe a continuous sequence of letters or words inside a text. Once a set of n-grams is extracted from a text, a probabilistic model of the text can be computed. This model can then be used to determine the probability of the transition from one letter to another.

There are other excellent articles related to this topic, for example the ones of Dave from hooktheory. His work is about musical progressions but the underlying math is more or less a more complex version of what I was doing with recipes and ingredients.

In the case of ingredients and recipes, there is one big difference to the work stated above though: I'm not necessarily looking at a continuous sequence. I don't care whether tomatoes are mentioned first and then potatoes or the other way around. What I'm looking at is simply the fact that they both occur in the same recipe.

What kind of data do we process?

As always, you need a lot of data if you want to generate quantitative statements. Obviously there are a lot of cooking websites which have huge amounts of data available. Unfortunately, most of them are not really into sharing their data and I would not like to risk any trouble by scraping their websites. After some search I stumbled on Wikia Recipes. Perfect! It's a sufficient amount of data with over 40k content pages and all provided under the Creative Commons license.

You might ask if the quality of the data is any good and if the set of recipes is representative. Well, this question is not really easy to answer. Browsing through the recipes on wikia I got the impression that the majority of the recipes is well written. According to the stats of the wiki, there are 51 active users managing the data right now from which 25 are marked as wikia staff. Let's just say the quality of the data is good!

I need to confess that as a European I'm not aware of all ingredients in the data set. I have really no idea how a *vegetable spray* or *milk chips* could taste like. And for sure there are some other things I never heard about.

Preprocessing the Data

The recipes are stored as wiki markup language inside a single XML file. Besides recipes, the xml includes a page for each ingredient and for each category. This is how a typical recipe looked like:

```
* Makes 6 to 8 servings

== Ingredients ==
* 2 tbsp extra virgin [[olive oil]]
* 3 cloves [[garlic]], finely chopped
* 1½ pounds fresh ripe [[tomato]]es, seeded and
chopped (about 3 cups)
* 1 tbsp [[tomato paste]]
* 1 tsp dried [[oregano]]
* ⅛ tsp [[ground red pepper]]
* ½ cup pitted brine cured [[black olives]] coarsely
chopped
* 2 tbsp [[capers]]
```

** [[salt]] and [[pepper]]*
** 1 pkg. thin [[spaghetti]] (16 oz)*
** grated [[Parmesan cheese]]*

== Directions ==
Heat olive oil and garlic in large skillet over low heat until garlic begins to sizzle.
Add tomatoes, tomato paste, oregano and red pepper; simmer, uncovered, until sauce is thickened, about 15 minutes.
Add olives, capers and salt and pepper to taste.
Toss spaghetti with sauce.
Sprinkle with cheese before serving.

[[Category:Cathy's Recipes]]
[[Category:Garlic Recipes]]
[[Category:Tomato Recipes]]
[[Category:Black olive Recipes]]
[[Category:Caper Recipes]]
[[Category:Spaghetti Recipes]]
[[Category:Parmesan cheese Recipes]]
[[Category:Tomato paste Recipes]]

As you can see, a recipe has more or less a clear structure and with some regex magic it should be possible to identify the recipe pattern and extract the ingredients. As you also can see, some of the ingredients contain the information that they are a synonym to another ingredient. This is great as it reduces the set of real ingredients and thus leads to more accurate results later on.

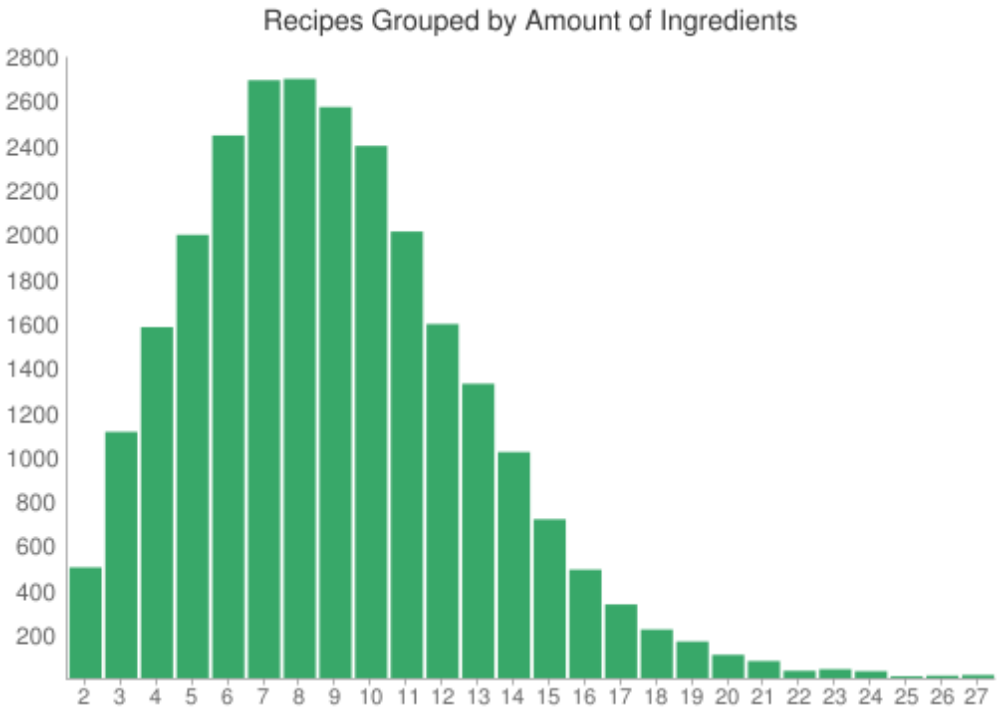
The next step was to write some scripts which transform the original XML file into a relational database which will be used for further processing. After processing all pages I ended up with a set of data which has

the following key values:

Total number of recipes:	29 920
Unique ingredients	4254
Ingredient marked as synonym:	2564

First result: How much ingredients does a typical recipe consist of?

It's difficult to say when you can call a combination of ingredients already a recipe. Does mixing two ingredients really count as a recipe? Is milk + chocolate or garlic + butter a recipe? Maybe yes, maybe not. The most complex recipe in the database consists of 42 ingredients – it's a Casserole and i really should give it a try once i have plenty of free time. As you can see in the following chart, the answer to how much ingredients a typical recipe consists of lies at around 8-10. This might be a little bit higher as expected but you need to take in account that really all ingredients are counted. This means also basic things like water and salt.



If you read that far you probably want to read the follow up on this post. The next post will cover the following questions:

- Which are the most frequent ingredients?
- Which are the most frequent combinations of ingredients?
- Given a set of ingredients, which other ingredients go probably along with it?

Follow me or subscribe to my mailing list and you won’t miss any upcoming article. In the mean time, feel free to comment!