**Analysis of Recipe’s website Data – a recommendation system**

Health is one of the most important worries of the human being. And it’s clear that it’s totally related to the quality of food you eat and the type of life you live. As the time seems to be “shorter” every day, people tend to eat more pre-processed food or fast food, in order to not waste time at this moment of the day. But with this concern of having a healthy life, people are getting more worried about cooking their own food, in a healthier way, using their own favorite and available ingredients.

Recipe’s websites are the most free and easy source for finding a way to cook new recipes. There are thousands of recipes, from different kind of types, using different kind of ingredients. But if the user is an inexperienced cooker, would he be able to choose the recipe that match with his desire of cooking a health meal?

My project proposal is to acquire data from food recipes websites, analyzing all the data available, to propose a recommendation website/program, offering different kinds of recommendation, based on some entries, like the available/desired ingredients, based on the nutrition facts (if the user has any restriction of some food, based on health needs), based on price, or maybe using all the entries together, to propose the healthier recipe, using the ingredients the had been chosen by the user.

One other aim is to provide decision makers with a better insight on the data, specially thinking about marketing actions can, understanding the best spot to sell advertisement, based on the analysis of the recipe’s popularity (more positive rating and more positive reviews) and even based on user’s review (in this case, using sentimental analysis of the reviews, to understand the user thinking).

**The Data**

The analyzed data were scraped from the “AllRecipes.com” website using the language R. The scraped data is organized as follow:

1. Data from the recipe: name of the recipe, preparation time, ingredients and quantities of this ingredients, how to prepare the recipe, recipe rating stars, number of reviews, categories that the recipe fit, nutrition facts);
2. Data from the review: comments, date of the review, rating of the recipe from the user.

All these data constitute in 5 files, for 200 MB. As the time to scrape the data was tight, I only scraped 10% of the recipes (5054 recipes) and the first 100 reviews of each recipe. These files were stored on excel sheets.

These are the tables scraped and their attributes:

**category\_db**

1. id\_recipe
2. category
3. Order

**ingredients\_db**

1. id\_recipe
2. ingredientid
3. ingredients
4. qtygrams
5. id\_ingredient

**nutrition\_db**

1. type
2. id\_recipe
3. description
4. qty
5. unity
6. percentage

**recipes\_db**

1. id\_recipe
2. recipe\_name
3. prep\_time
4. stars
5. qty\_reviews
6. directions

**reviews\_db**

1. id\_review
2. id\_recipe
3. review\_stars
4. dt\_review
5. coment

**The Project**

At this point of the project, the analysis were mainly exploratory, and relies on the strong assumption that the use of innovative data analysis technics on the available data can yield informative results. In this regard, three main questions of interest can be explored with this data:

1. Thinking about health, a deeper analysis of the recipe’s content is important, based on the ingredients for itself (text analysis of single words or expressions, like “chicken”, a healthy source of protein, and “fried chicken”, a unhealthy source of protein) and the nutrition facts of the recipe (quantity of fat, calories, carbohydrates, cholesterol, fiber, sodium and protein). The naturally nutrient rich (NNR) score, which is based on mean percentage daily values (DVs) for 14 nutrients in 2000 kcal food, can be used to assign nutrient density values to foods within and across some food groups. Theses analysis would help to identify the healthier choice for the user, based on he’s needs. Chosen the best diet that fit with the user is very important, because many diseases have strong relation with a poor or unbalanced diet. For example, analyzing the information from the Centers for Disease Control and Prevention, 6 of 10 leading causes of Death have some relation with a unhealthy diet

Leading Causes of Death from USA (2013)

|  |  |  |  |
| --- | --- | --- | --- |
| **Position** | **Leading Causes of Death** | **number of # Deaths** | **Influenced by diet** |
| 1 | Heart disease | 611,105 | yes |
| 2 | Cancer | 584,881 | yes |
| 3 | Chronic lower respiratory diseases | 149,205 | possibly |
| 4 | Accidents (unintentional injuries) | 130,557 | No |
| 5 | Stroke (cerebrovascular diseases) | 128,978 | yes |
| 6 | Alzheimer's disease | 84,767 | yes |
| 7 | Diabetes | 75,578 | yes |
| 8 | Influenza and Pneumonia | 56,979 | no |
| 9 | Nephritis, nephrotic syndrome, and nephrosis | 47,112 | yes |
| 10 | Intentional self-harm (suicide) | 41,149 | no |

Source: http://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

1. Many users want to begin to cook, but are too afraid of finding the best recipe based on the ingredients available and personal preferences (food type or some specification as low fat, low sodium, more protein, etc). This recipe recommendation system will be designed for new users or even for experienced users who wants to discover new kinds of recipes, and for those who wants to choose the best option for the needs;
2. Advertisement can bring more revenue and profitability to the website. Based on the sentimental analysis of the users reviews, and the combination of the popularity of the recipes and most used ingredients, can help to f chosen the best “spot” to show some advertisement.

**Key challenges**

The key challenges we encounter in this project are of different kinds:

1. Scrape more recipes from different websites (and finding different websites structures and different kind of information);
2. A good definition of “healthy recipes”, based on a profound study of nutrient information, diseases
3. Difficulties for recommender systems in nutrition as well as the challenges to overcome in implementing and designing these kind of systems