SNA Project: Vegetables and Health Benefits Network

In this project, I chose to analyze the the correlation between various vegetables and health benefits. http://www.nutrition-and-you.com provides a list of vegetables and descriptions of their health benefits. In this project, we use the information and network analysis to analyze the correlation between vegetables and corresponding health benefits. The basic workflow is as follows:

1. Parse the informatin from the website.
2. Use text mining to extract the key benefit claims.
3. Perform network analysis on the vegetables and their corresponding benefits.

This report is composed using RMarkdown. All the R/RMD codes and dataset for this SNA project can be found on Github: https://github.com/lifan0127/SNA_CourseProject.

```
library(tm)
library(knitr)
library(RWeka)
library(ggplot2)
library(gridExtra)
library(dplyr)  # version 0.3 required

# Load "vegetables" data frame from vegetable.RData
load("data/vegetables.RData")
```

Parse the informatin from the website

The data was parsed from the website using the RCurl and XML packages. The script can be found in the Github reposity associated with this project. The basic steps includes:

1. Parse vegetable names, images and links from http://www.nutrition-and-you.com/vegetable-nutrition.html.
2. Following the links, parse the health benefits for each vegetable.
3. Manual check to confirm consistency.

A sample of the data (first 5 vegetables) is shown below:

```
a <- data_frame(image=paste0("![", vegetables$Name, "](\"image/", vegetables$Name, ".gif\")"))
print("![\"abc\"](image/Artichoke.jpg)")
```

[1] “  ”

As an example, below is an excerpt of the textual description of health benefit related to Asparagus.

Data preprocessing