# Web Crawler Project

**Authors:** Jay Patel and Ami Patel
**Class:** CS 4395
**Instructor:** Dr. Karen Mazidi
**Date:** October 20, 2019

# Topic:

The topic chosen for this project is diets and different types of diets available. The web crawler script starts crawling using the url: https://en.wikipedia.org/wiki/List_of_diets

# Usage:

The usage of the script is as follows:

python3 web_crawler.py

# Files Created:

Total number of files created include:

1. The files created by scraping all the webpages.
2. The files created after extracting the sentences from each file.
3. The knowledge base text file, which is actually a dictionary including the important terms as tokens, and values being the list of sentences including the important term.

# Most important terms:

The top 35 terms based on tf-idf values across all documents are:

```
calorie -> 0.012020925137863599
weight -> 0.009854763178618177
jpg -> 0.007859835667064661
loss -> 0.0063824905919529105
people -> 0.00561869781176984
kidney -> 0.0055745803904398835
protein -> 0.00543253347576528
also -> 0.005316947657131977
kb -> 0.005009043249380765
fasting -> 0.004928252229229462
much -> 0.004854604382598761
carbohydrate -> 0.004854604382598761
may -> 0.00484746120907816
food -> 0.004822277510597726
atkins -> 0.004732092593552159
eating -> 0.004523323802660152
atkin -> 0.004351490482981146
body -> 0.0042011330478677385
disease -> 0.004190215317252085
cancer -> 0.004161089470798939
blood -> 0.004045503652165635
type -> 0.004045503652165635
medicine -> 0.004045503652165635
need -> 0.0038092866520473495
vegetarian -> 0.0035235901531437985
based -> 0.0035235901531437985
week -> 0.0033403006542721122
meal -> 0.0033124318262034093
many -> 0.003237893654240247
drink -> 0.003237893654240247
woman -> 0.0032364029217325074
patient -> 0.0031426614879390634
lose -> 0.0030817468518432004
treatment -> 0.0030474293216378795
gluten -> 0.0030052312844658998
```

Out of these 35 terms, manually 10 terms were chosen based on the domain of dieting.

These 10 terms were:
**['calorie', 'weight', 'food', 'eating', ' loss', ' fasting', 'disease', 'body', 'protein', 'meal']**

This was mainly because apart from 'disease', all the 9 words are directly related to or imply dieting in some or the other way. However, the word 'disease' was included because some of the valuable sentences that were found related to how the type of diet changes based on diseases.