Importing Libraries

In [10]:
```
import pandas as pd
import numpy as np
import requests
from urllib.request import urlopen
import time
from bs4 import BeautifulSoup
from nltk.corpus import stopwords
from nltk.corpus import wordnet
from nltk.stem.wordnet import WordNetLemmatizer
from nltk.corpus import wordnet as wn
import nltk
#nltk.download('brown')
```

Task 1 : Web Scraping

The website I have used for the project is 'www.allrecipes.com'. To scrape all the recepies from the website, I first iterated over the first 11 pages of the website, and pulled the recipies in every page.

In [2]:
```
recipe_list = []

for i in range(1,11):
    url = 'https://www.allrecipes.com/?page='+str(i)

    response = urlopen(url)
    my_html = response.read()
    response.close()

    soup = BeautifulSoup(my_html, 'html.parser')
    attribute_list = soup.find_all('a')

    for link in attribute_list:
        sublink = str(link.get('href'))
        if '/recipe/' in sublink:
            recipe_list.append(sublink)

recipe_list = list(set(recipe_list))
len(recipe_list)
```

Out[2]:
```
100
```

recipe_list now contains sublinks for every single recipe in my list. There are a total of 100 recipies in my list.

Here is an example of a sublink for a recipe.

In [3]:
```
recipe_list[10]
```

Out[3]:
```
'https://www.allrecipes.com/recipe/229150/cheesy-amish-breakfast-casserole/'
```

Next, I will visit the sublink for every recipe, and extract the name of the recipe and all its ingredients. I will then make a dataframe for the url of the recipe, the name and the ingredient (with one separate row for every ingredient).

In [4]:
```
df = pd.DataFrame(columns=['url','name','ingredient'])

for link in recipe_list:
    url = link

    response = urlopen(url)
    my_html = response.read()
    response.close()

    soup = BeautifulSoup(my_html, 'html.parser')

    title = soup.find('h1').text

    attribute_list = soup.findAll('li', attrs={'class':"ingredients-item"})

    for line in attribute_list:
        new_row = {'url':url, 'name':title, 'ingredient':' '.join(line.text.split())}
        df = df.append(new_row, ignore_index=True)

df.head()
```

Out[4]:

	url	name	ingredient
0	https://www.allrecipes.com/recipe/236876/hot-d...	Hot Dogs with Coney Sauce	1 pound lean ground beef
1	https://www.allrecipes.com/recipe/236876/hot-d...	Hot Dogs with Coney Sauce	1 (12 ounce) bottle chili sauce
2	https://www.allrecipes.com/recipe/236876/hot-d...	Hot Dogs with Coney Sauce	¼ cup water
3	https://www.allrecipes.com/recipe/236876/hot-d...	Hot Dogs with Coney Sauce	1 (1.25 ounce) package chili seasoning mix
4	https://www.allrecipes.com/recipe/236876/hot-d...	Hot Dogs with Coney Sauce	1 tablespoon yellow mustard

Here is the shape of the raw dataframe.

In [5]:
```
df.shape
```

Out[5]:
```
(708, 3)
```

In [7]:
```
#Saving the file in the folder
df.to_csv(r'rawData.csv', index=False)
```

Task 2 : Data Cleaning

In [8]:
```
# Saving the original list of ingredients for comparison purposes later

ogingredients = df.ingredient
```

As a part of the Data Cleaning process, I will perform the following operations:

1. Convert all the documents in a lowercase format
2. Get rid of everything that is not an alphabet (numbers, special characters, etc.)
3. Lemmatize the words to their noun forms
4. Build a modified list of english stopwords which also include the terms for cooking related measurements and actions
5. Get rid of all the words that are not nouns or adjectives

In [11]:
```
#Converting all words to lowercase
df['ingredient'] = df.ingredient.apply(lambda x: " ".join(x.lower() for x in x.split()))

#Removes Punctuation from all sentences
df.ingredient = df.ingredient.str.replace('[^a-zA-Z_\s]',' ')

#Create a lemmatizer object
lemmatizer = WordNetLemmatizer()

#Lemmatizes all the words
df.ingredient = df.ingredient.apply(lambda x: " ".join(lemmatizer.lemmatize(x, wordnet.NOUN) for x in x.split
()))

#Removing stop words
#Get all the english stop words except for the word 'all'
stop = stopwords.words('english')
stop.remove('all')

#Define a list of all the possible words that are used for measurement
measurements = ['piece', 'cup','tablespoon','length', 'bunch', 'pound', 'ounce', 'dash', 'teaspoon','quart',
'inch','degree','optional','half','package', 'chunk', 'pinch','f','c','envelope','small','bulk','large']

#Define a set of all the actions (instructions)
actions = ['cut','chopped','diced','cubed','taste','packed','slice','use','peeled','pitted','mashed']

#Growing the stop words list by adding in measurements and actions
stop.extend(measurements+actions)

#Removing all the words in the new stop word list
df.ingredient =  df.ingredient.apply(lambda x: " ".join(x for x in x.split() if x not in stop))

#Attempting the use of the wordnet library for removing everything except nouns
#df_temp['method1'] = df_temp.clean_ingredient.apply(lambda x: " ".join([n for n,t in [(w, wn.synsets(w)[0].p
os()) for w in x.split()] if t in ['n']]))

#Removing words except nouns and adjectives
df.ingredient = df.ingredient.apply(lambda x: " ".join([n for n,t in [j[0] for j in [nltk.pos_tag([i]) for i
in x.split()]] if t in ['NN','NNS','JJ']]))

df.head()
```

Out[11]:

	url	name	ingredient
0	https://www.allrecipes.com/recipe/236876/hot-d...	Hot Dogs with Coney Sauce	lean ground beef
1	https://www.allrecipes.com/recipe/236876/hot-d...	Hot Dogs with Coney Sauce	bottle chili sauce
2	https://www.allrecipes.com/recipe/236876/hot-d...	Hot Dogs with Coney Sauce	water
3	https://www.allrecipes.com/recipe/236876/hot-d...	Hot Dogs with Coney Sauce	chili mix
4	https://www.allrecipes.com/recipe/236876/hot-d...	Hot Dogs with Coney Sauce	yellow mustard

Let us now compare the original ingredient list with the new cleaned ingredient list.

In [18]:
```
set(zip(ogingredients,df.ingredient))
```

Out[18]:
```
{('active dry yeast', 'active dry yeast'),
 ('almond', 'almond'),
 ('asian chile pepper sauce', 'asian chile pepper sauce'),
 ('asian dark sesame oil', 'asian dark sesame oil'),
 ('avocado', 'avocado'),
 ('bacon', 'bacon'),
 ('bag frozen mexican style corn', 'bag frozen mexican style corn'),
 ('bag mexican cheese blend', 'bag mexican cheese blend'),
 ('baking powder', 'baking powder'),
 ('baking soda', 'baking soda'),
 ('banana', 'banana'),
 ('basil', 'basil'),
 ('basil leaf', 'basil leaf'),
 ('bay leaf', 'bay leaf'),
 ('beef broth', 'beef broth'),
 ('beef chuck roast', 'beef chuck roast'),
 ('beef flank steak thick diagonal', 'beef flank steak thick diagonal'),
 ('beef hot dog', 'beef hot dog'),
 ('beef stock', 'beef stock'),
 ('beer', 'beer'),
 ('black bean', 'black bean'),
 ('black olive', 'black olive'),
 ('black pepper', 'black pepper'),
 ('boneless pork chop', 'boneless pork chop'),
 ('boneless pork shoulder roast', 'boneless pork shoulder roast'),
 ('boneless skinless chicken breast', 'boneless skinless chicken breast'),
 ('boneless skinless chicken breast thin strip',
  'boneless skinless chicken breast thin strip'),
 ('boneless skinless chicken thigh cube',
  'boneless skinless chicken thigh cube'),
 ('boston bibb butter lettuce leaf', 'boston bibb butter lettuce leaf'),
 ('bottle chili sauce', 'bottle chili sauce'),
 ('bottle mexican guajillo red chile cooking sauce herdez',
  'bottle mexican guajillo red chile cooking sauce herdez'),
 ('bread crumb', 'bread crumb'),
 ('bread flour', 'bread flour'),
 ('bread machine yeast', 'bread machine yeast'),
 ('brown sugar', 'brown sugar'),
 ('butter', 'butter'),
 ('buttermilk brushing', 'buttermilk brushing'),
 ('cajun', 'cajun'),
 ('caper', 'caper'),
 ('carrot', 'carrot'),
```

```
('cayenne pepper', 'cayenne pepper'),
('celery', 'celery'),
('celery seed', 'celery seed'),
('cheddar cheese', 'cheddar cheese'),
('cheese', 'cheese'),
('cheese cracker cheez', 'cheese cracker cheez'),
('cherry tomato', 'cherry tomato'),
('chicken bouillon granule', 'chicken bouillon granule'),
('chicken breast skin', 'chicken breast skin'),
('chicken broth', 'chicken broth'),
('chicken leg quarter split drumstick thigh',
 'chicken leg quarter split drumstick thigh'),
('chili bean', 'chili bean'),
('chili mix', 'chili mix'),
('chili powder', 'chili powder'),
('cider vinegar', 'cider vinegar'),
('clove fresh garlic', 'clove fresh garlic'),
('clove garlic', 'clove garlic'),
('clove garlic unpeeled', 'clove garlic unpeeled'),
('cold buttermilk', 'cold buttermilk'),
('cold unsalted butter', 'cold unsalted butter'),
('cold water', 'cold water'),
('confectioner sugar', 'confectioner sugar'),
('container sour cream', 'container sour cream'),
('cooked chicken', 'cooked chicken'),
('cooking oil', 'cooking oil'),
('cooking sherry', 'cooking sherry'),
('cooking spray pam', 'cooking spray pam'),
('corn kernel', 'corn kernel'),
('cornstarch', 'cornstarch'),
('cream cheese', 'cream cheese'),
('cream chicken soup', 'cream chicken soup'),
('cream mushroom soup', 'cream mushroom soup'),
('crunchy peanut butter', 'crunchy peanut butter'),
('crust ready pie crust pillsbury', 'crust ready pie crust pillsbury'),
('curd cottage cheese', 'curd cottage cheese'),
('dark brown sugar', 'dark brown sugar'),
('dijon mustard', 'dijon mustard'),
('dry bread crumb', 'dry bread crumb'),
('dry chicken gravy mix', 'dry chicken gravy mix'),
('dry onion soup mix', 'dry onion soup mix'),
('dry sherry', 'dry sherry'),
('dry vermouth', 'dry vermouth'),
```

```
('dry white wine', 'dry white wine'),
('eaches skinless boneless chicken breast',
 'eaches skinless boneless chicken breast'),
('eaches skinless boneless chicken breast cube',
 'eaches skinless boneless chicken breast cube'),
('egg', 'egg'),
('egg beaten', 'egg beaten'),
('egg room temperature', 'egg room temperature'),
('egg yolk', 'egg yolk'),
('enchilada sauce', 'enchilada sauce'),
('extra virgin olive oil', 'extra virgin olive oil'),
('fennel seed', 'fennel seed'),
('feta cheese', 'feta cheese'),
('flank steak', 'flank steak'),
('flour tortilla', 'flour tortilla'),
('fluid enchilada sauce', 'fluid enchilada sauce'),
('french onion', 'french onion'),
('french roll split', 'french roll split'),
('fresh basil', 'fresh basil'),
('fresh basil leaf thin strip', 'fresh basil leaf thin strip'),
('fresh blueberry', 'fresh blueberry'),
('fresh cilantro', 'fresh cilantro'),
('fresh frozen green bean thawed', 'fresh frozen green bean thawed'),
('fresh frozen pea', 'fresh frozen pea'),
('fresh ginger root', 'fresh ginger root'),
('fresh italian parsley', 'fresh italian parsley'),
('fresh lemon juice', 'fresh lemon juice'),
('fresh mozzarella cube', 'fresh mozzarella cube'),
('fresh mushroom', 'fresh mushroom'),
('fresh parsley', 'fresh parsley'),
('fresh rosemary', 'fresh rosemary'),
('fresh thyme', 'fresh thyme'),
('fresh tomato', 'fresh tomato'),
('frozen green pea', 'frozen green pea'),
('frozen hash brown potato thawed', 'frozen hash brown potato thawed'),
('frozen pea', 'frozen pea'),
('frozen tater tot', 'frozen tater tot'),
('garlic', 'garlic'),
('garlic powder', 'garlic powder'),
('ginger', 'ginger'),
('goat cheese', 'goat cheese'),
('granny smith apple', 'granny smith apple'),
('grapeseed oil', 'grapeseed oil'),
```

```
('green bell pepper', 'green bell pepper'),
('green bell pepper top seed', 'green bell pepper top seed'),
('green chilies', 'green chilies'),
('green onion', 'green onion'),
('ground beef', 'ground beef'),
('ground black pepper', 'ground black pepper'),
('ground cayenne pepper', 'ground cayenne pepper'),
('ground cinnamon', 'ground cinnamon'),
('ground coriander', 'ground coriander'),
('ground cumin', 'ground cumin'),
('ground ginger', 'ground ginger'),
('ground pork', 'ground pork'),
('ground sirloin', 'ground sirloin'),
('ham', 'ham'),
('heavy cream', 'heavy cream'),
('hoisin sauce', 'hoisin sauce'),
('honey', 'honey'),
('hot cooked brown rice', 'hot cooked brown rice'),
('hot dog bun', 'hot dog bun'),
('hot water', 'hot water'),
('instant yeast', 'instant yeast'),
('italian', 'italian'),
('italian bread crumb', 'italian bread crumb'),
('italian herb', 'italian herb'),
('ketchup', 'ketchup'),
('kosher salt', 'kosher salt'),
('kosher salt ground black pepper', 'kosher salt ground black pepper'),
('lard', 'lard'),
('lasagna noodle', 'lasagna noodle'),
('lean ground beef', 'lean ground beef'),
('lemon juice', 'lemon juice'),
('lemon juiced', 'lemon juiced'),
('lemon pepper', 'lemon pepper'),
('light brown sugar', 'light brown sugar'),
('lime juice', 'lime juice'),
('lime juiced', 'lime juiced'),
('linguine pasta', 'linguine pasta'),
('maple syrup', 'maple syrup'),
('margarine', 'margarine'),
('marsala wine', 'marsala wine'),
('mayonnaise', 'mayonnaise'),
('medium banana', 'medium banana'),
('mexican cheese blend', 'mexican cheese blend'),
```

```
('mild taco mix', 'mild taco mix'),
('milk', 'milk'),
('mozzarella cheese', 'mozzarella cheese'),
('mushroom', 'mushroom'),
('oil', 'oil'),
('olive oil', 'olive oil'),
('onion', 'onion'),
('onion powder', 'onion powder'),
('onion ring', 'onion ring'),
('onion salt', 'onion salt'),
('orange juice', 'orange juice'),
('orange zest', 'orange zest'),
('oregano', 'oregano'),
('overripe banana', 'overripe banana'),
('oyster sauce', 'oyster sauce'),
('packet taco mix', 'packet taco mix'),
('panko bread crumb', 'panko bread crumb'),
('paprika', 'paprika'),
('parmesan cheese', 'parmesan cheese'),
('parsley', 'parsley'),
('peanut oil', 'peanut oil'),
('pecan', 'pecan'),
('pepper', 'pepper'),
('pine nut', 'pine nut'),
('plain bread crumb', 'plain bread crumb'),
('pork sausage', 'pork sausage'),
('potato', 'potato'),
('prepared mustard', 'prepared mustard'),
('prepared tomato sauce', 'prepared tomato sauce'),
('prepared yellow mustard', 'prepared yellow mustard'),
('provolone cheese', 'provolone cheese'),
('purpose flour', 'purpose flour'),
('quick cooking oat', 'quick cooking oat'),
('raisin', 'raisin'),
('ramen noodle flavor packet', 'ramen noodle flavor packet'),
('ranch', 'ranch'),
('ranch mix', 'ranch mix'),
('recipe pastry crust pie', 'recipe pastry crust pie'),
('red bell pepper', 'red bell pepper'),
('red onion', 'red onion'),
('red pepper', 'red pepper'),
('red pepper flake', 'red pepper flake'),
('red potato', 'red potato'),
```

```
('rice vinegar', 'rice vinegar'),
('rice wine vinegar', 'rice wine vinegar'),
('ricotta cheese', 'ricotta cheese'),
('ripe banana', 'ripe banana'),
('rom plum tomato', 'rom plum tomato'),
('rom tomato', 'rom tomato'),
('rump roast', 'rump roast'),
('russet potato', 'russet potato'),
('salsa', 'salsa'),
('salt', 'salt'),
('salt ground black pepper', 'salt ground black pepper'),
('salt pepper', 'salt pepper'),
('sea salt', 'sea salt'),
('seedless grape', 'seedless grape'),
('self flour', 'self flour'),
('semisweet chocolate chip', 'semisweet chocolate chip'),
('sesame oil', 'sesame oil'),
('sesame seed', 'sesame seed'),
('skinless boneless chicken breast', 'skinless boneless chicken breast'),
('skinless boneless chicken breast meat',
 'skinless boneless chicken breast meat'),
('skinless boneless chicken breast thick',
 'skinless boneless chicken breast thick'),
('skinless chicken thigh', 'skinless chicken thigh'),
('snow pea', 'snow pea'),
('sodium soy sauce', 'sodium soy sauce'),
('sour cream', 'sour cream'),
('soy sauce', 'soy sauce'),
('spaghetti', 'spaghetti'),
('spinach', 'spinach'),
('stalk celery', 'stalk celery'),
('strip orange zest', 'strip orange zest'),
('sweet italian sausage', 'sweet italian sausage'),
('sweet onion', 'sweet onion'),
('swiss cheese', 'swiss cheese'),
('thin asparagus spear', 'thin asparagus spear'),
('thyme', 'thyme'),
('tomato', 'tomato'),
('tomato paste', 'tomato paste'),
('tomato sauce', 'tomato sauce'),
('tomato soup', 'tomato soup'),
('tortilla chip', 'tortilla chip'),
('unbaked pie crust', 'unbaked pie crust'),
```

```
('unsalted butter', 'unsalted butter'),
('unsalted butter freezer thin', 'unsalted butter freezer thin'),
('unsweetened cocoa powder', 'unsweetened cocoa powder'),
('vanilla extract', 'vanilla extract'),
('vegetable oil', 'vegetable oil'),
('walnut', 'walnut'),
('warm milk', 'warm milk'),
('warm water', 'warm water'),
('water', 'water'),
('water chestnut', 'water chestnut'),
('white cake mix', 'white cake mix'),
('white onion', 'white onion'),
('white pepper', 'white pepper'),
('white sugar', 'white sugar'),
('white vinegar', 'white vinegar'),
('white wine', 'white wine'),
('white wine vinegar', 'white wine vinegar'),
('whole chicken bone skin meat', 'whole chicken bone skin meat'),
('whole red chilies', 'whole red chilies'),
('worcestershire sauce', 'worcestershire sauce'),
('yellow mustard', 'yellow mustard'),
('yellow onion', 'yellow onion'),
('yukon gold potato', 'yukon gold potato'),
('zucchini', 'zucchini')}
```

As you can see in the list, some of the items are repeated with a different name (for e.g. clove garlic and clove fresh garlic are the same ingredient). With further modifications in the code, we can correct these problems as well. However, I want to keep this code as general as possible.

In [13]:
```
#Saving the file in the folder
df.to_csv(r'cleanData.csv', index=False)
```

Task 3 : Calculating

Finally, I grouped the data by the ingredients to get the number of times the ingredient has occured in my data. Next, I divided this count by the number of unique recipies I have in my list. This gave me the proportion of the total unique recipies that have the ingredient in them.

In [15]:
```
df_res = df.groupby('ingredient').count().sort_values(by='url', ascending=False)[['name']].reset_index()
df_res = df_res.rename({'name':'count', 'ingredient':'word'}, axis=1)
total = len(df.name.unique())
df_res['proportion'] = df_res['count'].apply(lambda x:x/total)
df_top10 = df_res.head(10)
df_top10
```

Out[15]:

	word	count	proportion
0	salt	38	0.550725
1	purpose flour	31	0.449275
2	butter	31	0.449275
3	white sugar	29	0.420290
4	egg	20	0.289855
5	water	16	0.231884
6	onion	15	0.217391
7	brown sugar	14	0.202899
8	clove garlic	13	0.188406
9	ground black pepper	13	0.188406

These results make sense as well.

Salt is used in almost all the recipies (expect for deserts), which is why its present in a very high proporton of 0.55.

Next in the line are flour, butter and white sugar - all three which are again very popular in all the dishes.

In [16]:
```
#Saving the file in the folder
df_top10.to_csv(r'results.csv', index=False)
```