Data Structuring

February 12, 2021

Objectives

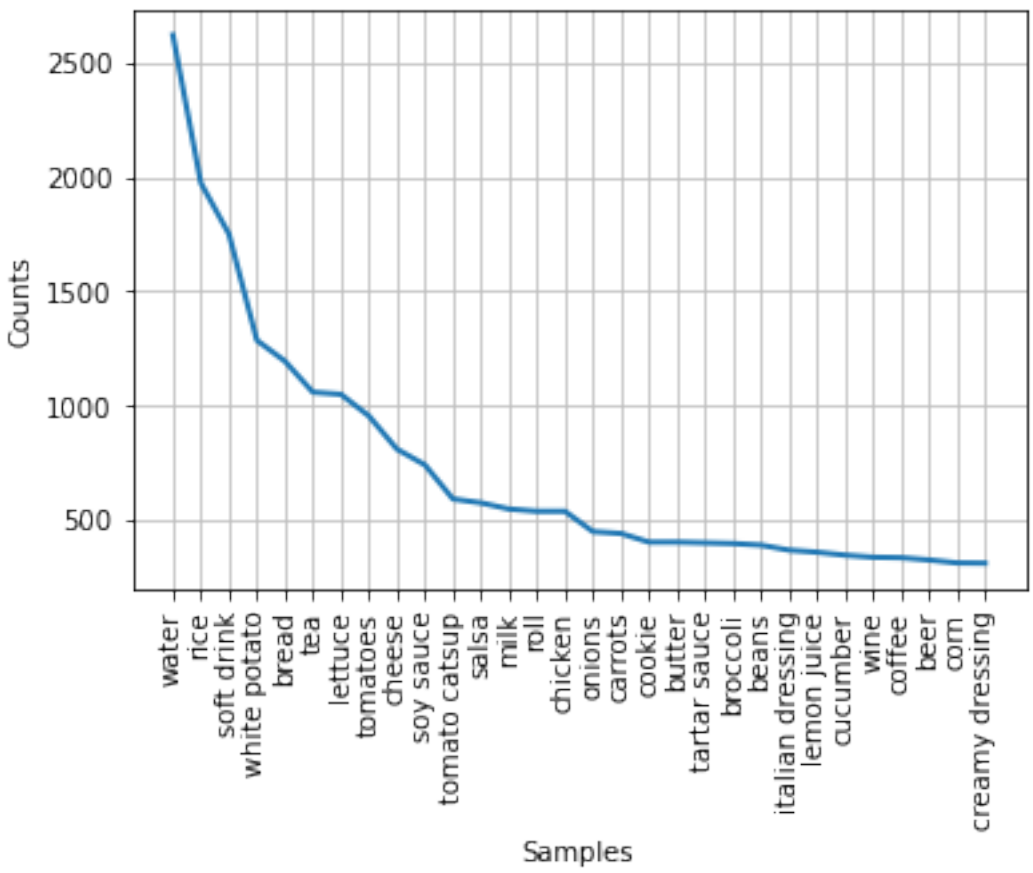
[1]:
```
import pandas as pd
import re
import nltk


#Read filtered dataframe
nhanes = pd.read_pickle('../../Data/nhanes_post.pkl')


#Obtain dataframe with seafood items
seafood_df = nhanes[nhanes['DR1I_PF_SEAFD_TOT'] > 0]
#Obtain dataframe with side dishes
side_dish_df = nhanes[nhanes['DR1I_PF_SEAFD_TOT'] == 0]

"""
Obtain initial test corpus for the whole meal, seafood item only, and side␣
 ↪dishes only
Obtains the first word in the text description string before a comma, if comma␣
 ↪exists.
Obtains the whole string in the text description if comma is not present.
"""
food_type_cps = nhanes['DESCRIPTION'].apply(lambda x: re.search(r'^([^,])+', x).
 ↪group(0) if re.search((r','), x) else x)
seafood_cps = seafood_df['DESCRIPTION'].apply(lambda x: re.search(r'^([^,])+',␣
 ↪x).group(0) if re.search((r','), x) else x)
side_dish_cps = side_dish_df['DESCRIPTION'].apply(lambda x: re.
 ↪search(r'^([^,])+', x).group(0) if re.search((r','), x) else x)

#Obtain and plot frequency distribution of the side dish words
side_dish_fdist = nltk.FreqDist(side_dish_cps)
side_dish_fdist.plot(30)
```



[1]: <matplotlib.axes._subplots.AxesSubplot at 0x7f8c63ac9250>

Conclusions