<https://frenzy86.s3.eu-west-2.amazonaws.com/python/nlp/email_re.xlsx>

import re

import pandas as pd

df = pd.read\_excel('email\_re.xlsx')

lista = df['content'].tolist()

lista = [str(i) for i in lista] # element all string inside list

text = ''.join(lista).lower()          # make a unique string and LOWER!!!!!!!

pattern = r'[a-z0-9]+@[a-z]+\.[a-z]{2,3}'

#pattern = r'\S+@\S+\.[a-z]{2,3}' text = text.replace("("," ")

#pattern = r'\w+@\w+\.[a-z]{2,3}'

email = re.findall(pattern,text)

email

from collections import Counter

# words = ['a', 'b', 'c', 'a']

# Counter(words).keys() # equals to list(set(words))

# Counter(words).values() # counts the elements' frequency

print(Counter(email).keys()) # equals to list(set(words))

print(Counter(email).values()) # counts the elements' frequency

k = list(Counter(email).keys())

v = list(Counter(email).values())

tot = pd.DataFrame(list(zip(k, v)),

                 columns =['email', 'count'])

tot = tot.sort\_values('count',ascending=False)

tot

set\_ = {"mala", "banana", "arancia","arancia"}

print(set\_)

mylist = ['aa', 'bb', 'bb', 'aa', 'c', 'd', 'e']

myset = set(mylist)

myset

email\_final = set(email) #unique valuse inside set

len(email\_final)

pallen = re.findall('pallen@enron.com',text) # because needs string not list

len(pallen)

retwell = re.findall('retwell@mail.san',text) # because needs string not list

len(retwell)

with open('result.txt', 'w') as f:

    for i in email\_final:

        f.write("%s\n" % i)

tot.to\_csv('result.csv',index=False)

reviews\_text = df["Review Text"]

reviews\_text = reviews\_text.dropna()

reviews\_text.head(10)

reviews\_text.loc[4] # first review id4.

new\_reviews = pd.concat([reviews\_text, df["Star Rating"]], axis=1).dropna() ## clean dataframe with dropna

new\_reviews.head(10)

reviews\_text.map(len)

new\_reviews['Review Length'] = reviews\_text.map(len)

new\_reviews

new\_reviews.groupby(['Star Rating']).count()

new\_reviews.mean()

new\_reviews.describe()