

SF_to_SD_EDA

March 18, 2021

Objectives

Create a dataframe that obtains the seafood type to side dish associations and perform some exploratory data analysis on those associations.

```
[2]: import pandas as pd
import re

#Read the structured dataframe
df_final = pd.read_pickle('../Data/df_final.pkl')
df_final_cols = df_final.columns.values.tolist()

#Create a list of all the "SFx" columns, where x is a digit. This pulls in the
→columns containing the
#seafood type from the structured database.
#The regular expression looks for two digits after "SF" first, then it looks
→for one digit.
#This will only work for SF0-SF99, and will break if there are more than 99
→seafood dishes in one meal (unlikely)
sf_cols = []
for i in range(len(df_final_cols)):
    if (re.match(r"SF\d\d", df_final_cols[i])):
        sf_cols.append(re.findall(r"SF\d\d", df_final_cols[i])[0])
    elif (re.match(r"SF\d", df_final_cols[i])):
        sf_cols.append(re.findall(r"SF\d", df_final_cols[i])[0])

#Create a list of all the "SDx" columns, where x is a digit. This pulls in the
→columns containing the
#side dish type from the structured database.
#The regular expression looks for three digits after "SD" first, then it looks
→for two digits, then one.
#This will only work for SD0-SD999, and will break if there are more than 999
→seafood side dishes in one meal (unlikely)
sd_cols = []
for i in range(len(df_final_cols)):
    if (re.match(r"SD\d\d\d", df_final_cols[i])):
        sd_cols.append(re.findall(r"SD\d\d\d", df_final_cols[i])[0])
    elif (re.match(r"SD\d\d", df_final_cols[i])):
```

```

        sd_cols.append(re.findall(r"SD\d\d", df_final_cols[i])[0])
    elif (re.match(r"SD\d", df_final_cols[i])):
        sd_cols.append(re.findall(r"SD\d", df_final_cols[i])[0])

#Create the seafood to side dish associations by sequentially melting the
↳dataframe, pivoting on each SF column
df_sf_to_sd = pd.DataFrame()
#Loop over the SF columns to pivot the associations
for i in sf_cols:
    #Temporary storage of the Df columns. The columns will be used to filter
↳the DF
    #This line is used to reset at start of loop.
    df_cols_temp = []
    #Append the DF columns with the SD columns
    for j in range(len(sd_cols)):
        df_cols_temp.append(sd_cols[j])
    #Append the DF columns with the SF pivot column
    df_cols_temp.append(i)
    #Include only SF pivot column and all SD columns
    df_temp = df_final[df_cols_temp]
    #Melt the dataframe by the SF pivot column
    df_temp = pd.melt(df_temp, id_vars=i)
    #Rename the the SF pivot column to a common name to be used in all loop
↳iterations
    #Rename variable to the sd_n, to indicate the number of side dishes for the
↳seafood type
    df_temp.rename({i: 'sf_type', 'variable': 'sd_n', 'value': 'sd_name'},
↳axis=1, inplace=True)
    #Append the result of each loop to the final dataframe
    df_sf_to_sd = df_sf_to_sd.append(df_temp)
    df_sf_to_sd = df_sf_to_sd.reset_index(drop=True)

#Remove rows that are NaN
df_sf_to_sd = df_sf_to_sd[df_sf_to_sd['sf_type'].notna()]
df_sf_to_sd = df_sf_to_sd[df_sf_to_sd['sd_name'].notna()]

```

Part 1

What is the percentage distribution of associations among the different seafood types?

```

[3]: #Obtain number of unique side dishes for each seafood type
sf_type_unq_sd = df_sf_to_sd.groupby('sf_type')['sd_name'].nunique().
↳sort_values(ascending=False)

#Obtain association statistics for the unique seafood species
sf_to_sd_num = df_sf_to_sd.groupby(['sf_type', 'sd_name']).size().
↳reset_index(name='count').sort_values(by = 'count', ascending=False)

```

```

#Calculates the side dish total count for each seafood type, merges to
→association df
sd_tot = sf_to_sd_num.groupby('sf_type')['count'].sum()
sf_to_sd_num = pd.merge(sf_to_sd_num, sd_tot, how='left', on=['sf_type'])
#Calculate percetange of each side type as proportion to total side dishes for
→each seafood type
sf_to_sd_num['pct'] = (sf_to_sd_num['count_x']/sf_to_sd_num['count_y'])*100
#Sort by highest percentage
sf_to_sd_num = sf_to_sd_num.sort_values(by=['pct'], ascending=False)

sf_to_sd_num.head()

```

```

[3]:      sf_type      sd_name  count_x  count_y      pct
3888  turtle  mixed vegetables      1      2  50.000000
3839  turtle      water      1      2  50.000000
7011   ray      water      1      3  33.333333
6997   ray  margarine-like spread      1      3  33.333333
7012   ray      plantain      1      3  33.333333

```

The highest percentage of associations is among cases where the total side dish count for the seafood type is very low.

```

[4]: ## Sort by highest percentage
sf_to_sd_num_flt = sf_to_sd_num[sf_to_sd_num['count_y'] >= 20].head(25)
sf_to_sd_num_flt['sf_sd_pair'] = sf_to_sd_num_flt['sf_type'] + ' - ' +
→sf_to_sd_num_flt['sd_name']

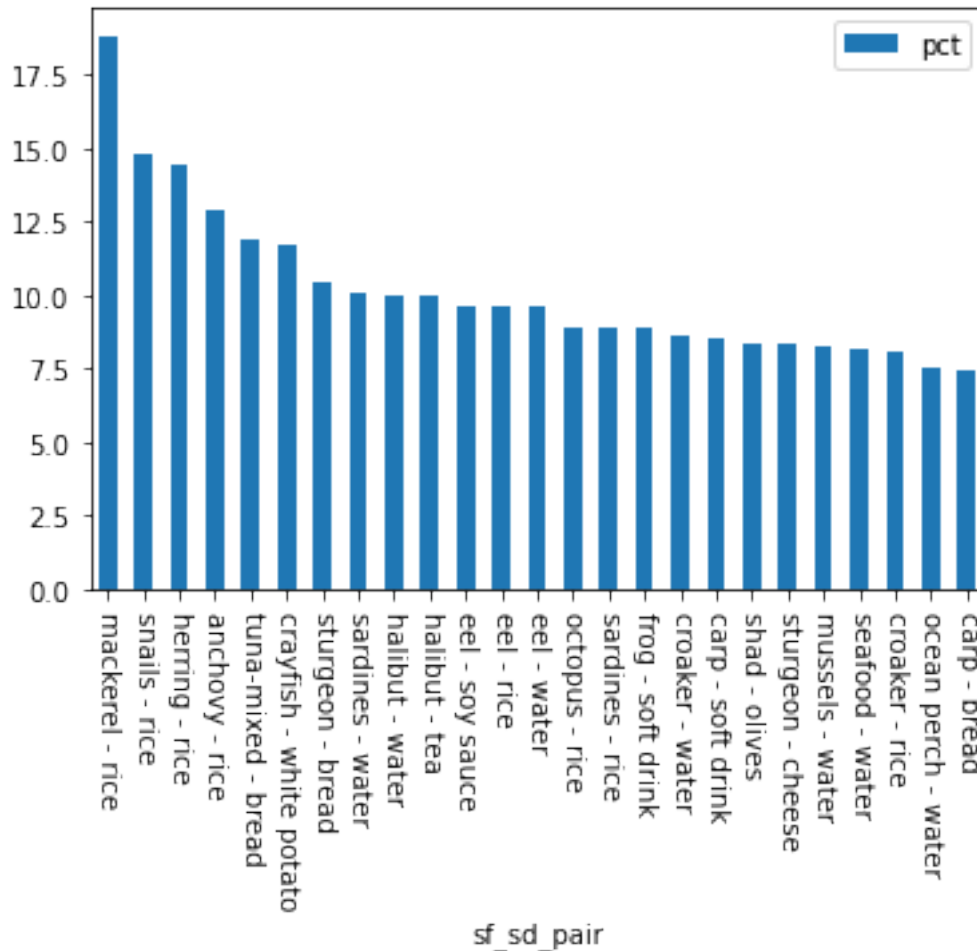
sf_to_sd_num_flt.plot.bar(x='sf_sd_pair', y='pct', rot=270)

```

```

[4]: <matplotlib.axes._subplots.AxesSubplot at 0x7fe889c902e0>

```



The plot above displays the top 15 associations based on percentage, for cases where there are at least 20 instances of side dishes.

```
[5]: sf_to_sd_num_grouped = sf_to_sd_num.groupby('sf_type')

for name,group in sf_to_sd_num_grouped:
    group = group.head(15)
    group['sf_sd_pair'] = group['sf_type'] + ' - ' + group['sd_name']
    group_plot = group.plot.barh(x='sf_sd_pair', y='pct', rot=0)
    group_plot.invert_yaxis()
```

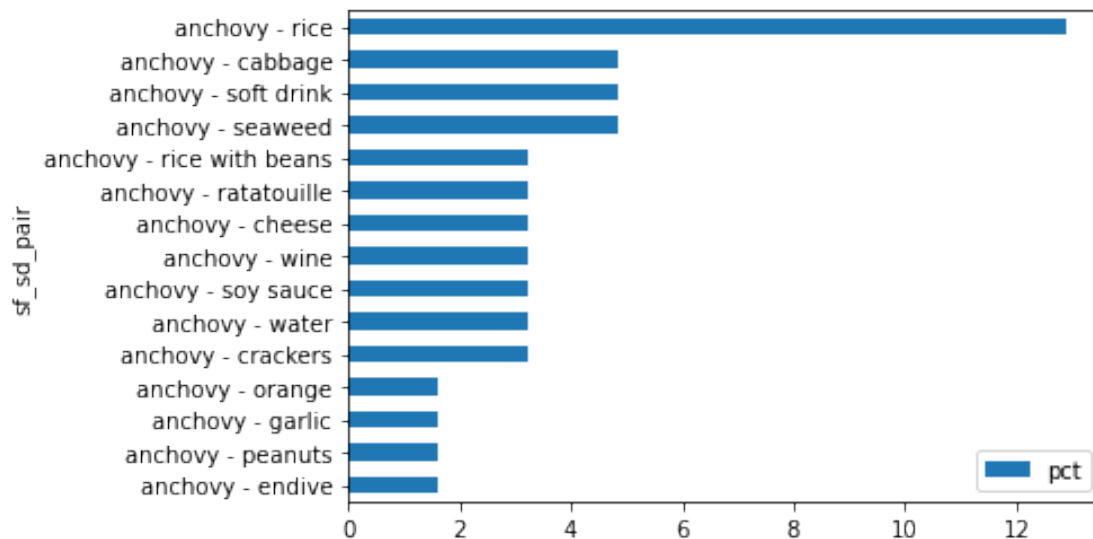
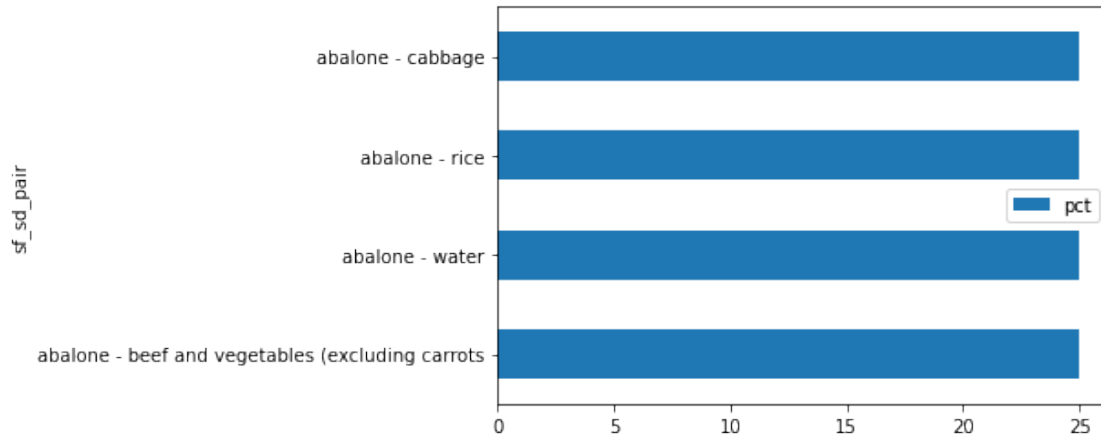
```
<ipython-input-5-f75ec8e27339>:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

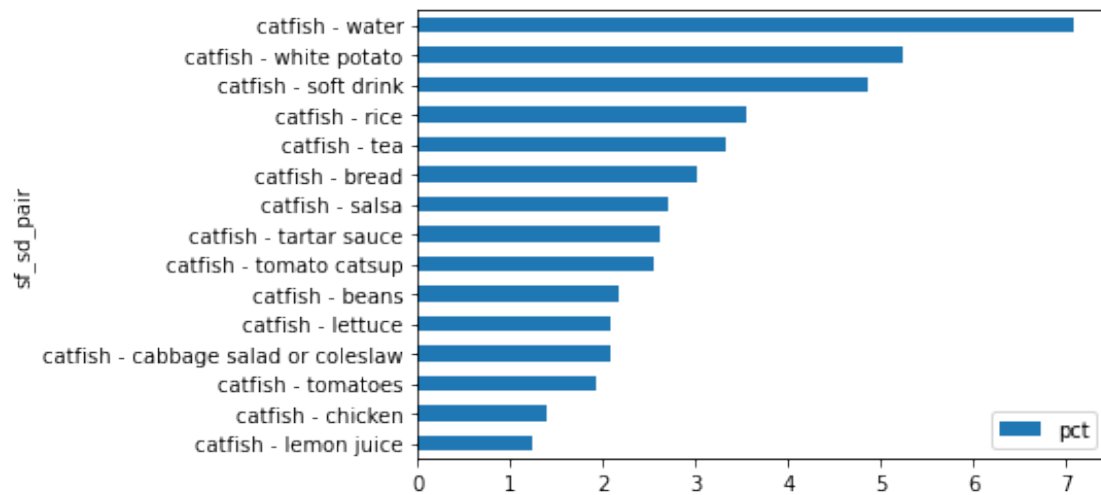
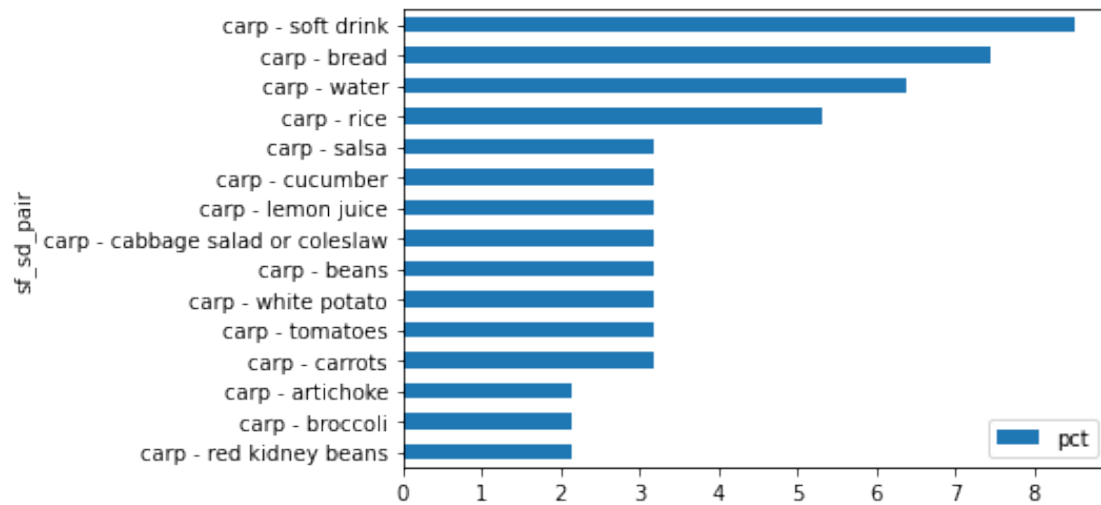
See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

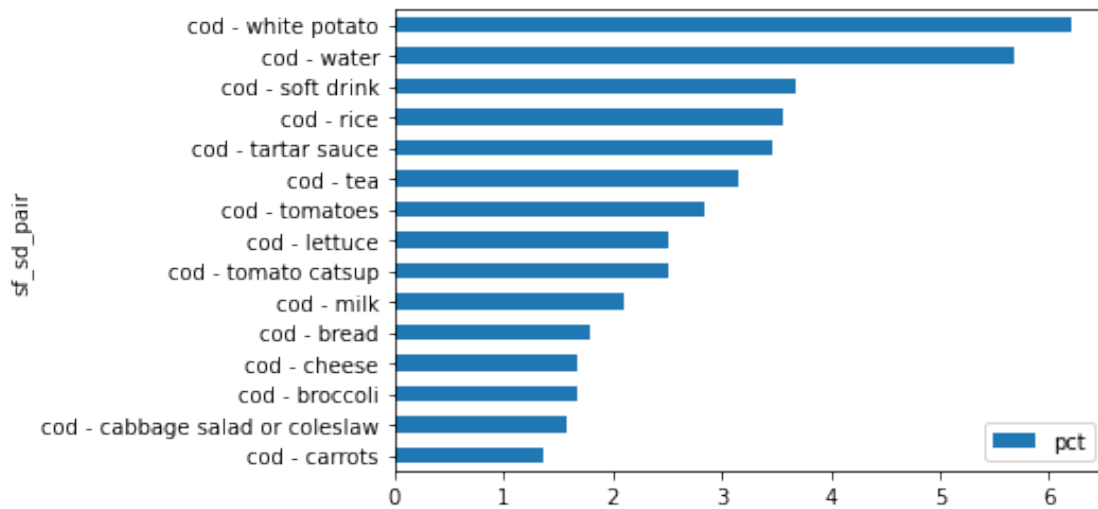
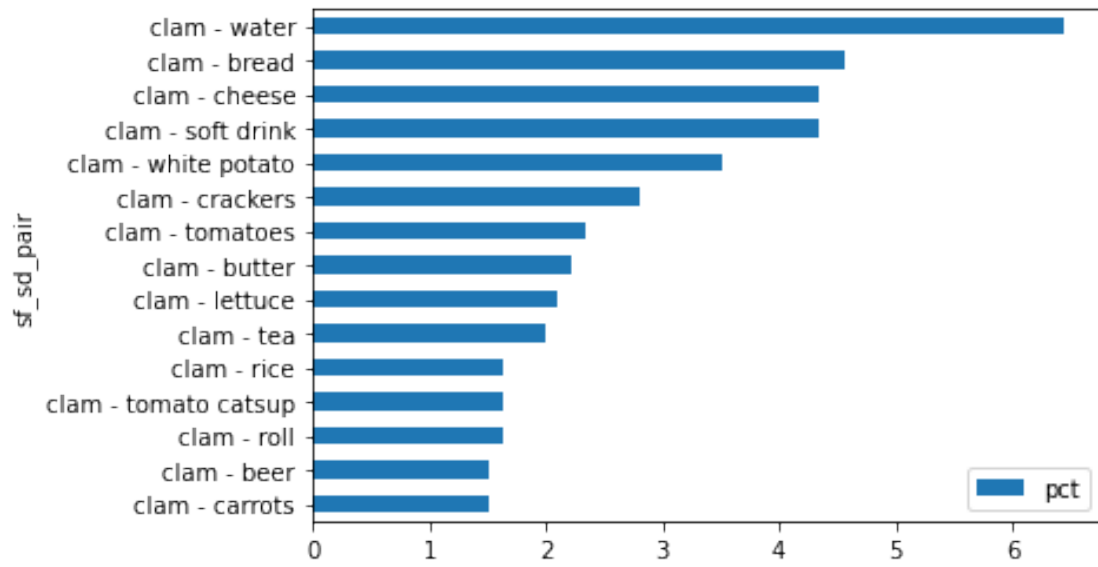
```
group['sf_sd_pair'] = group['sf_type'] + ' - ' + group['sd_name']
```

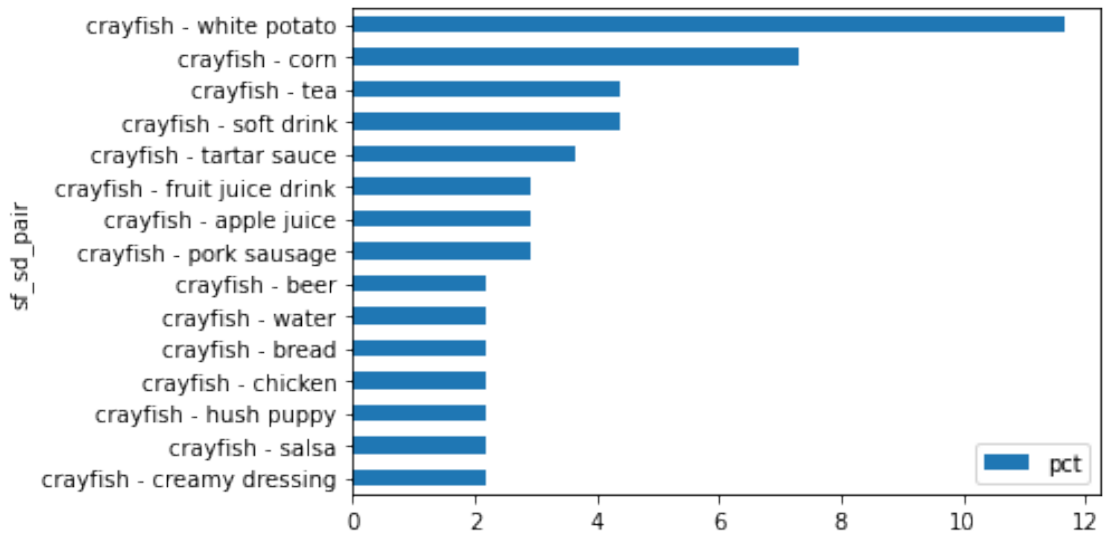
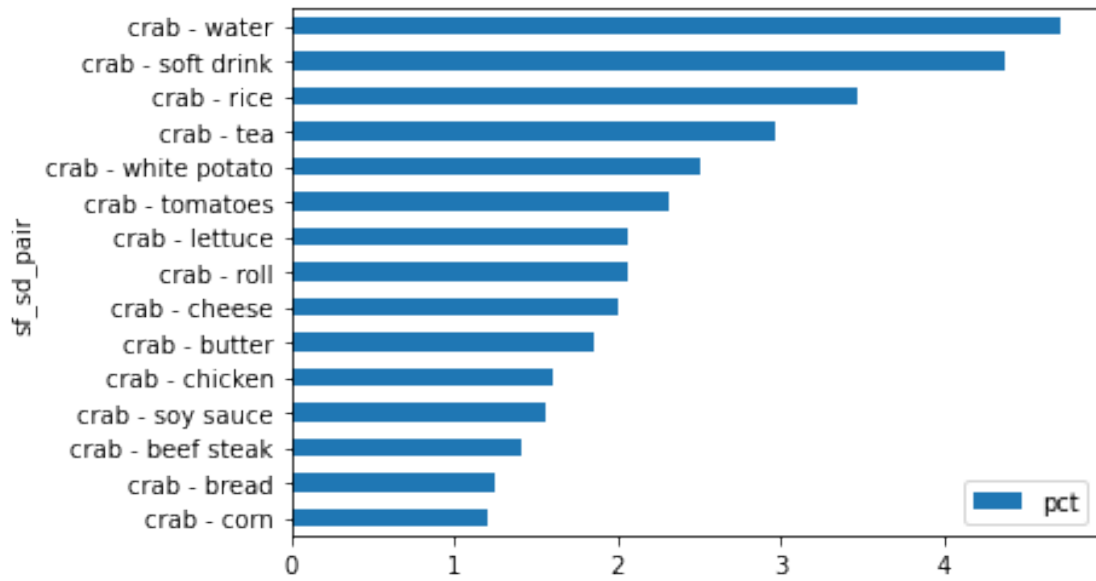
/Users/jori/anaconda2/lib/python3.8/site-packages/pandas/plotting/_matplotlib/core.py:328: RuntimeWarning: More than 20 figures have been opened. Figures created through the pyplot interface (`matplotlib.pyplot.figure`) are retained until explicitly closed and may consume too much memory. (To control this warning, see the rcParam `figure.max_open_warning`).

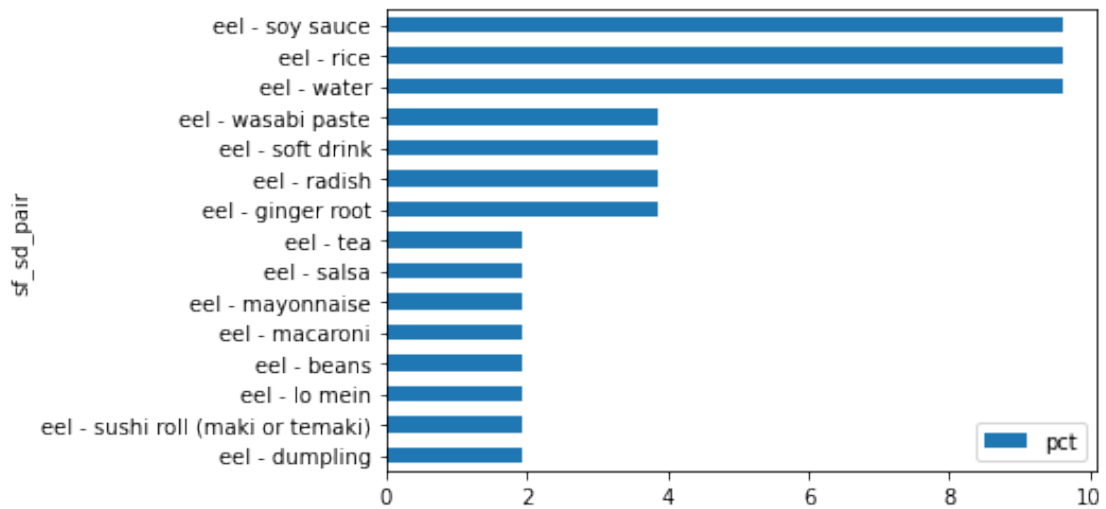
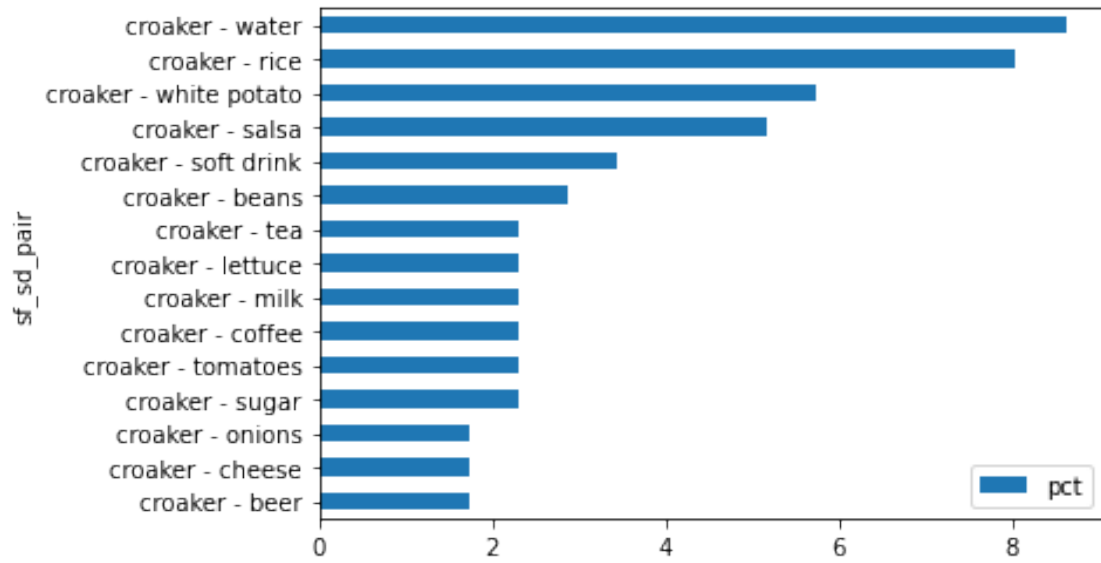
```
fig = self.plt.figure(figsize=self.figsize)
```

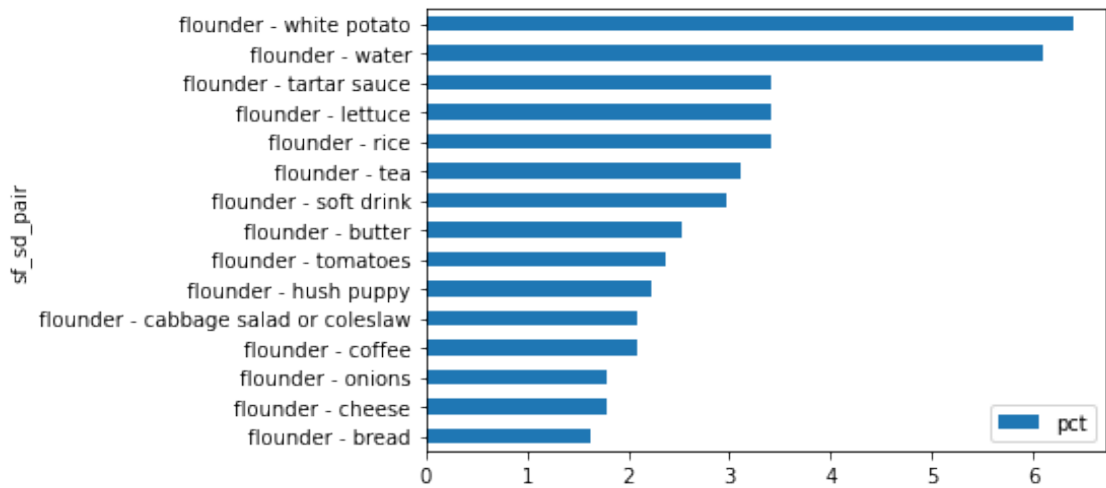
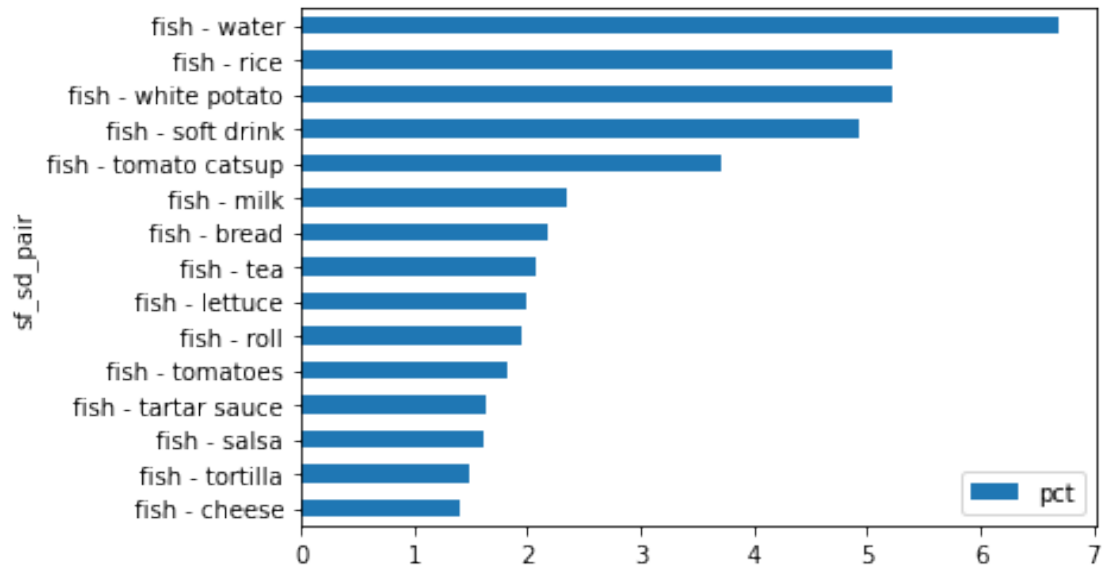


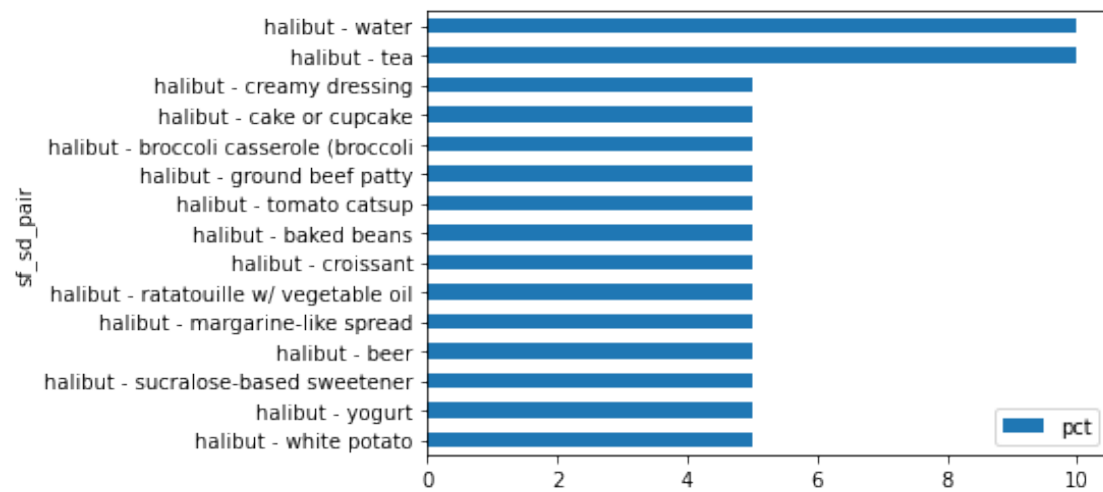
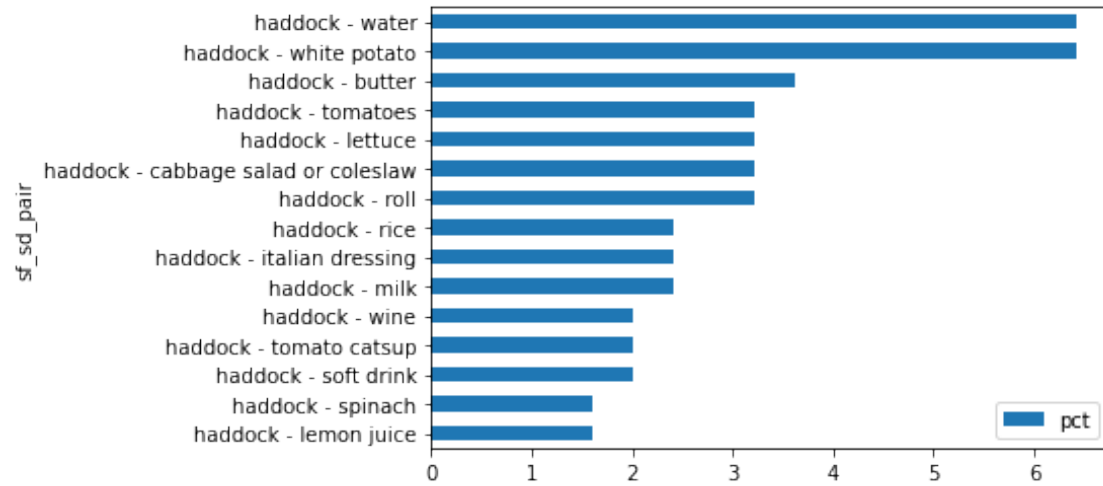
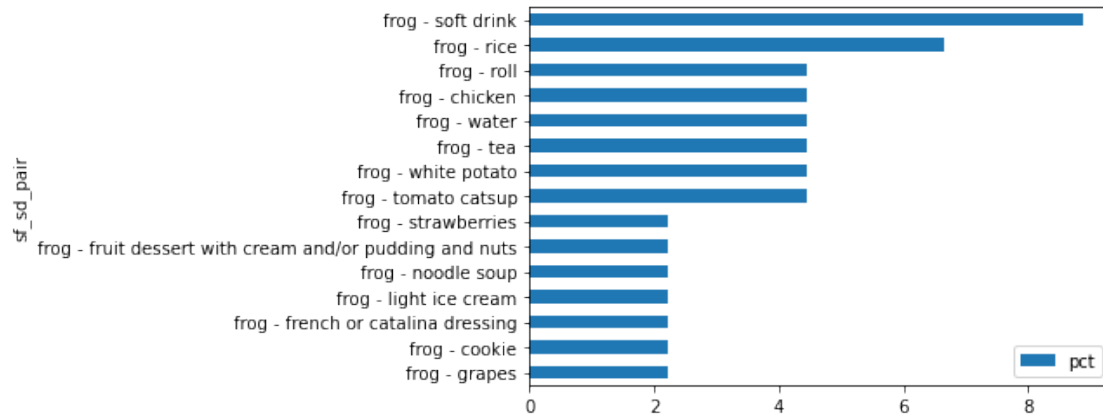


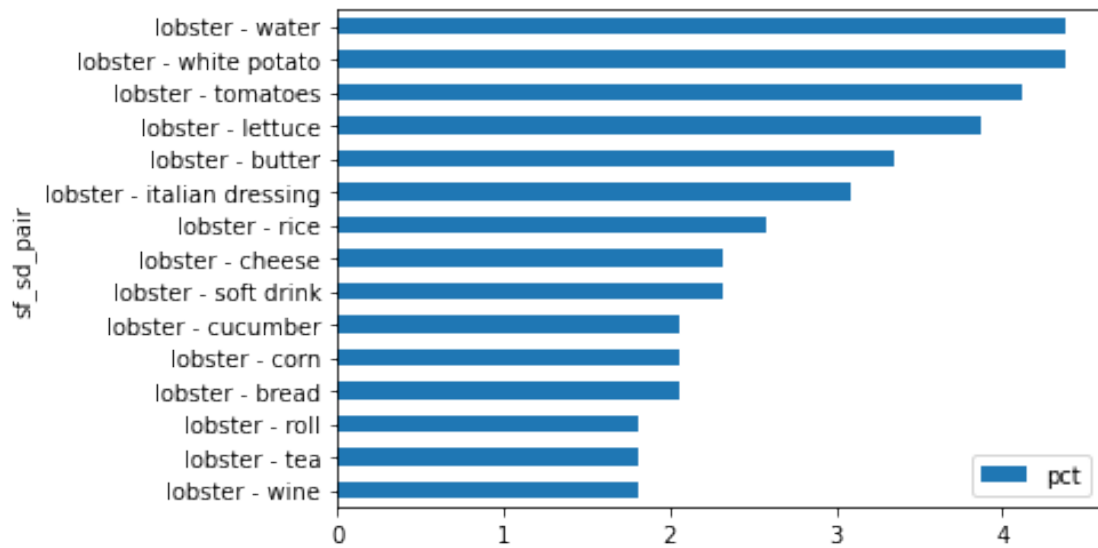
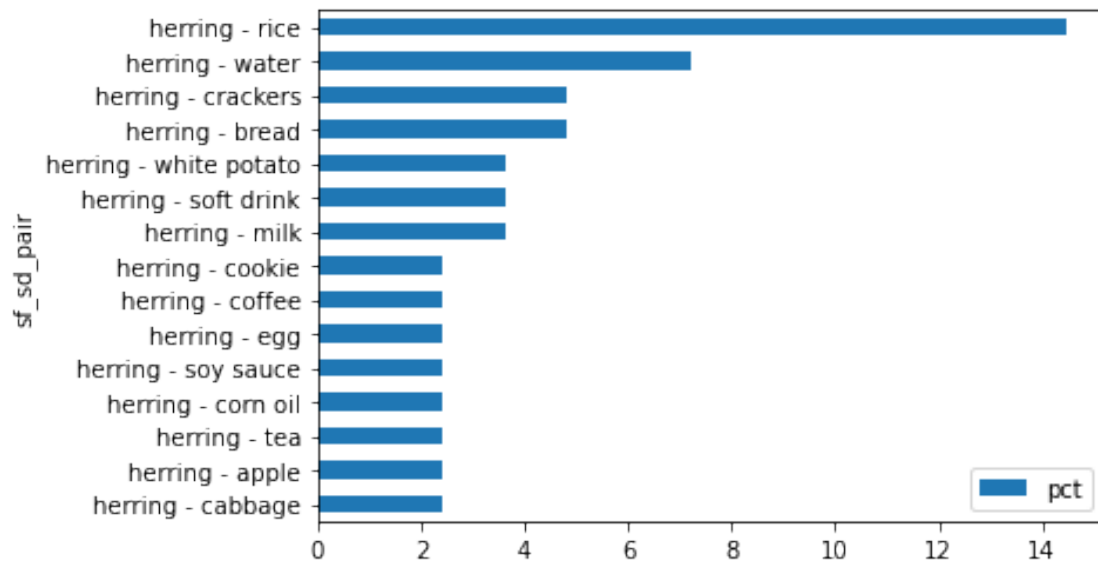


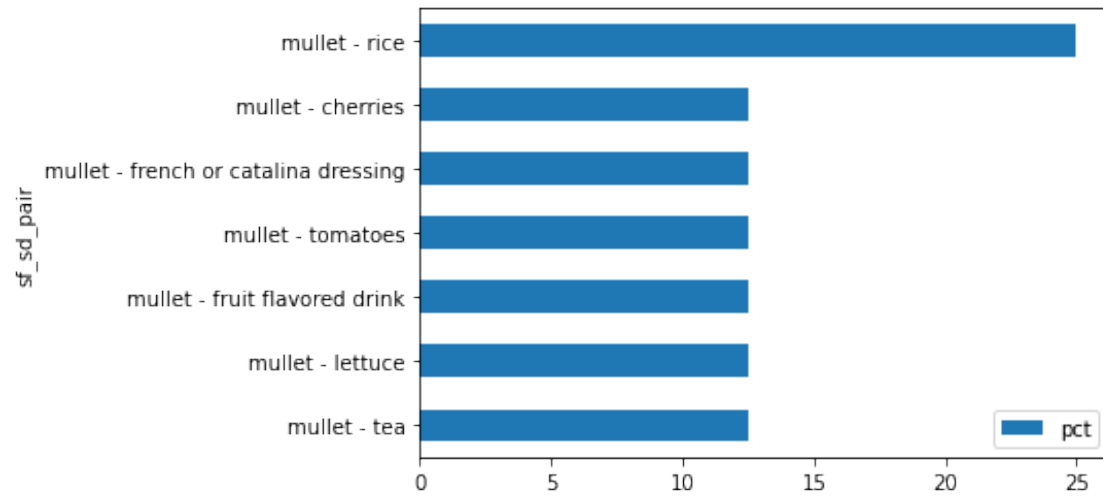
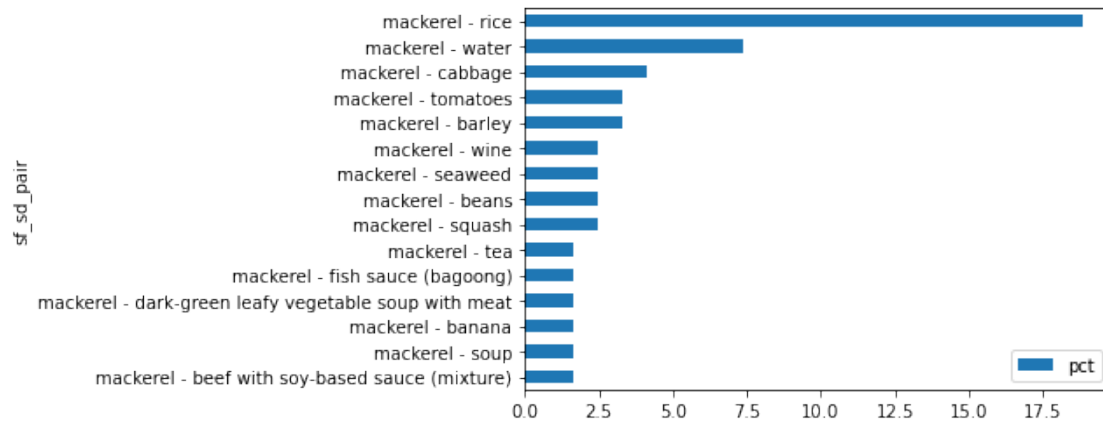


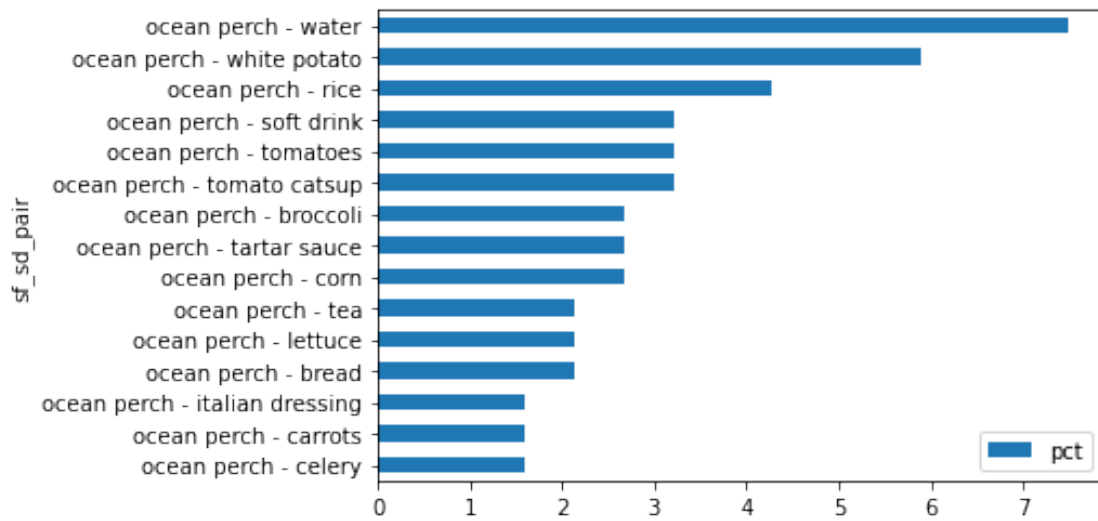
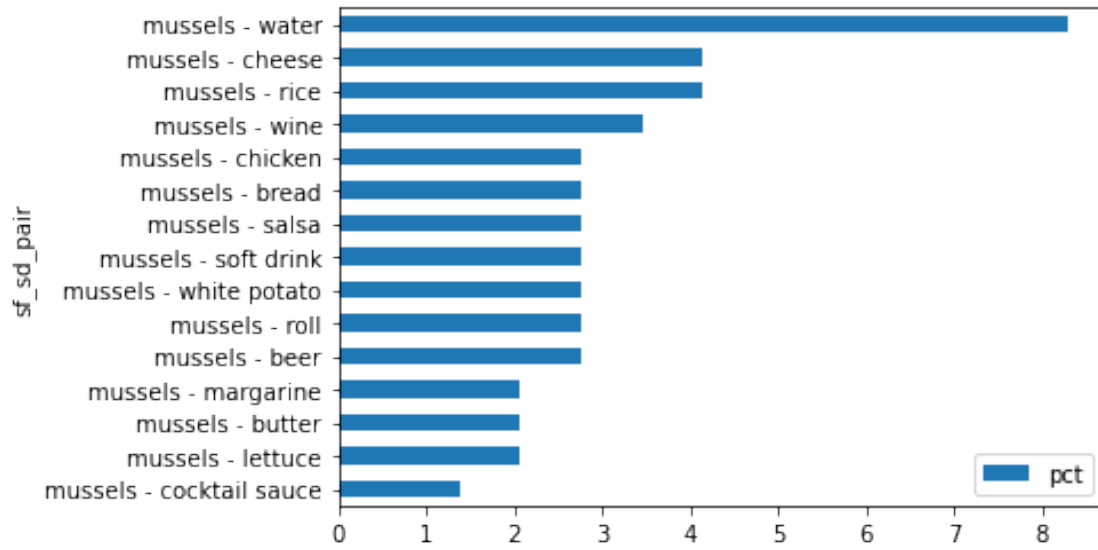


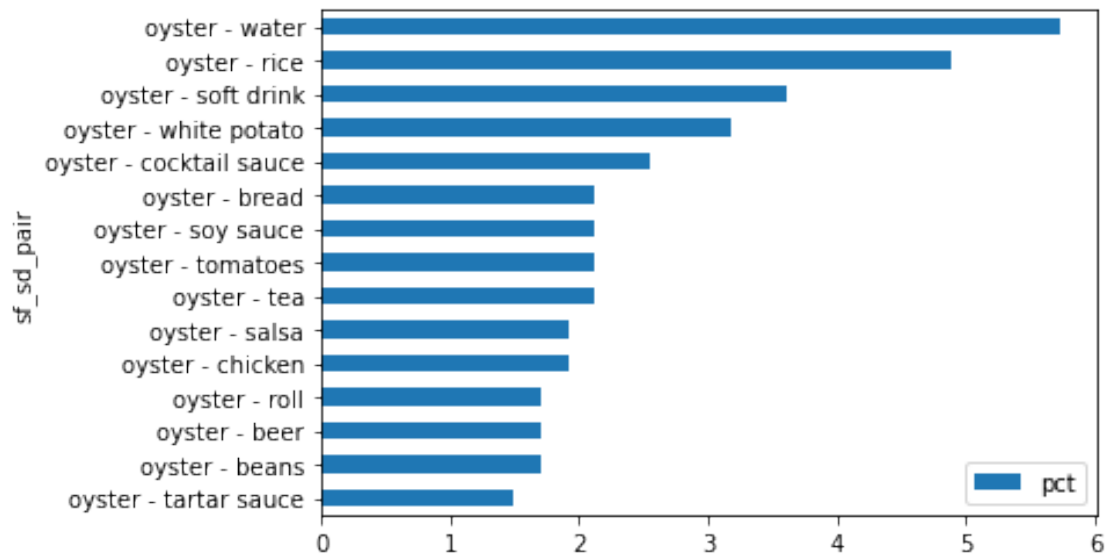
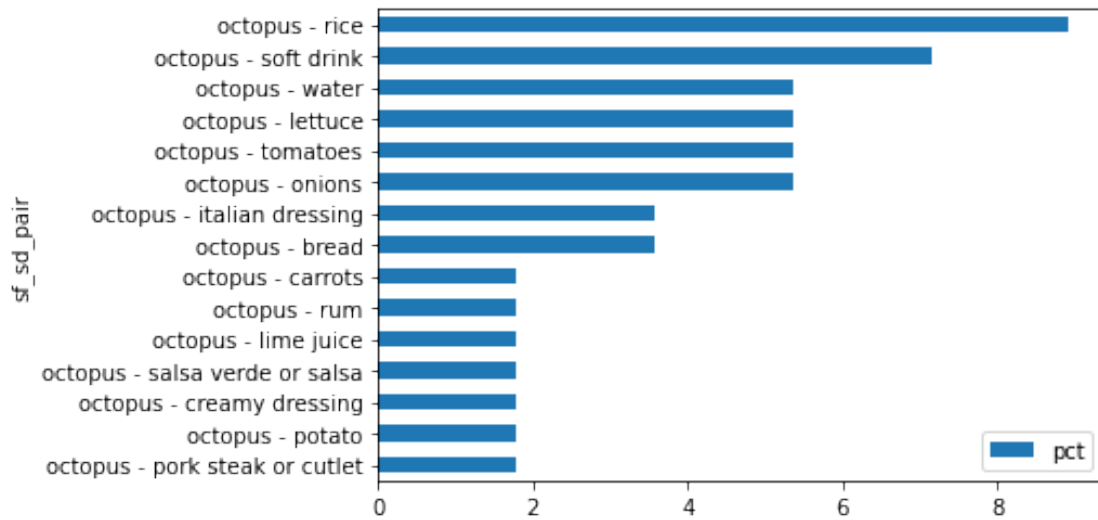


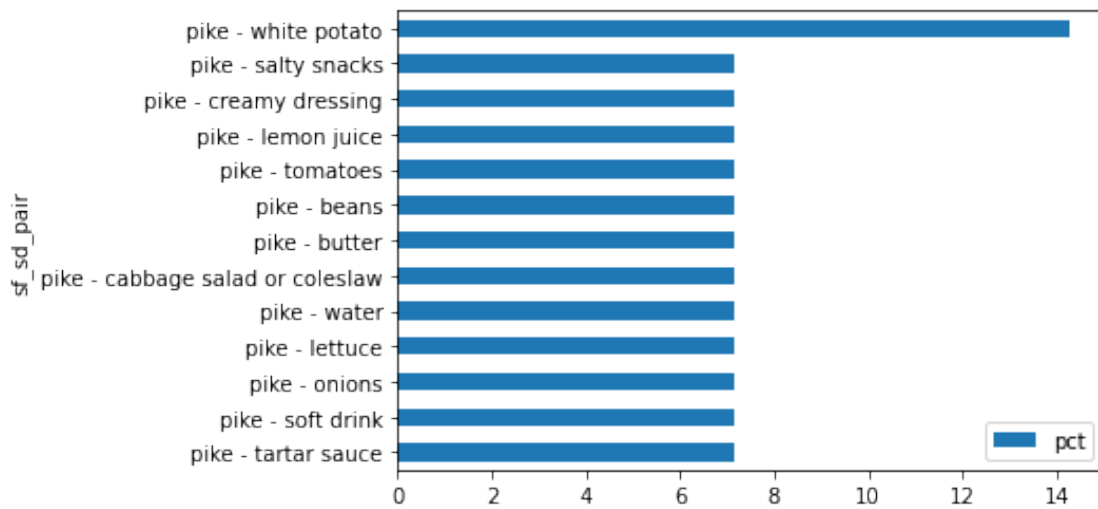
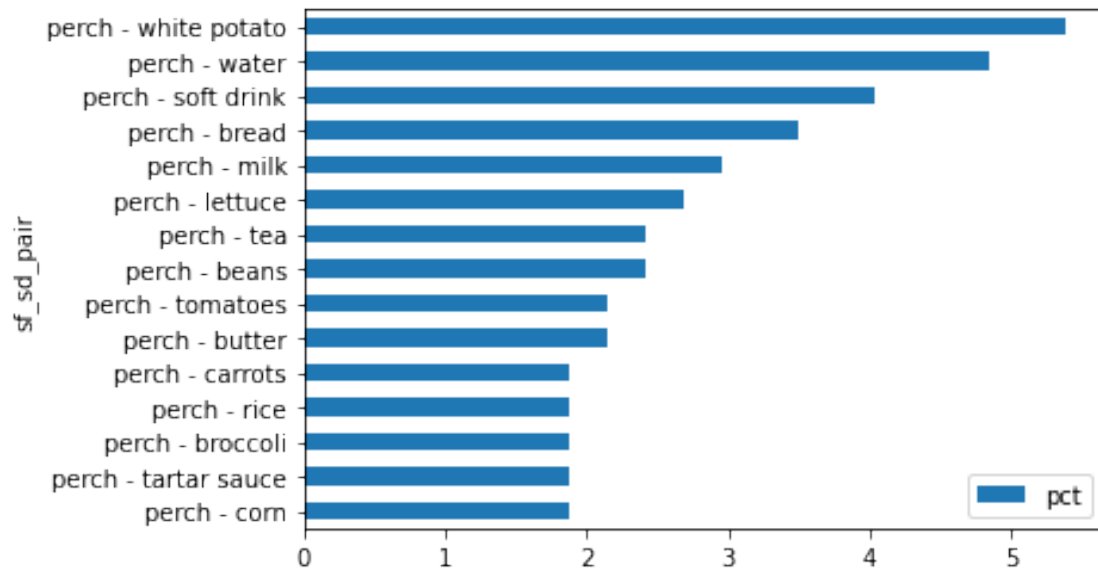


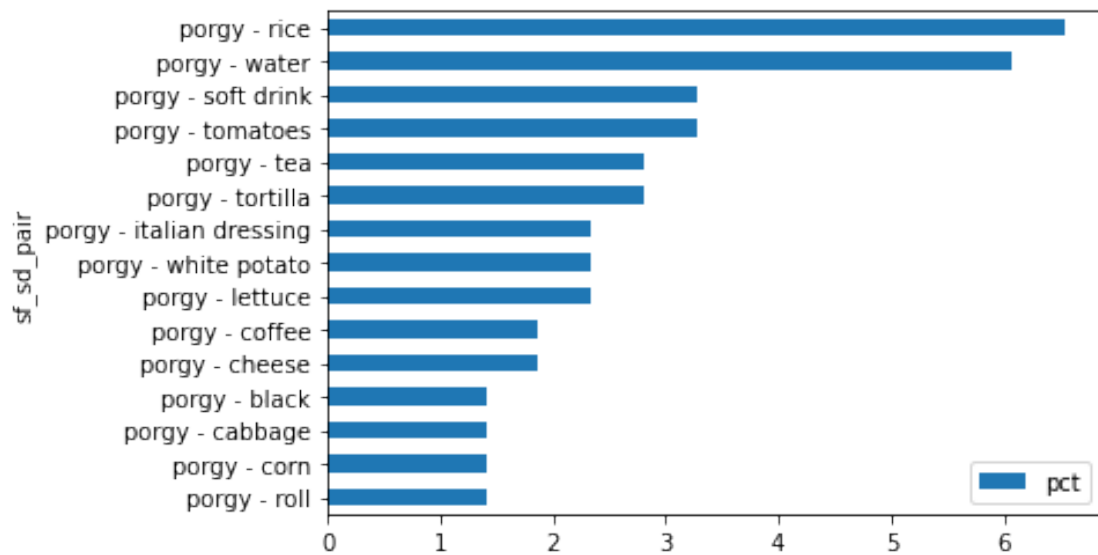
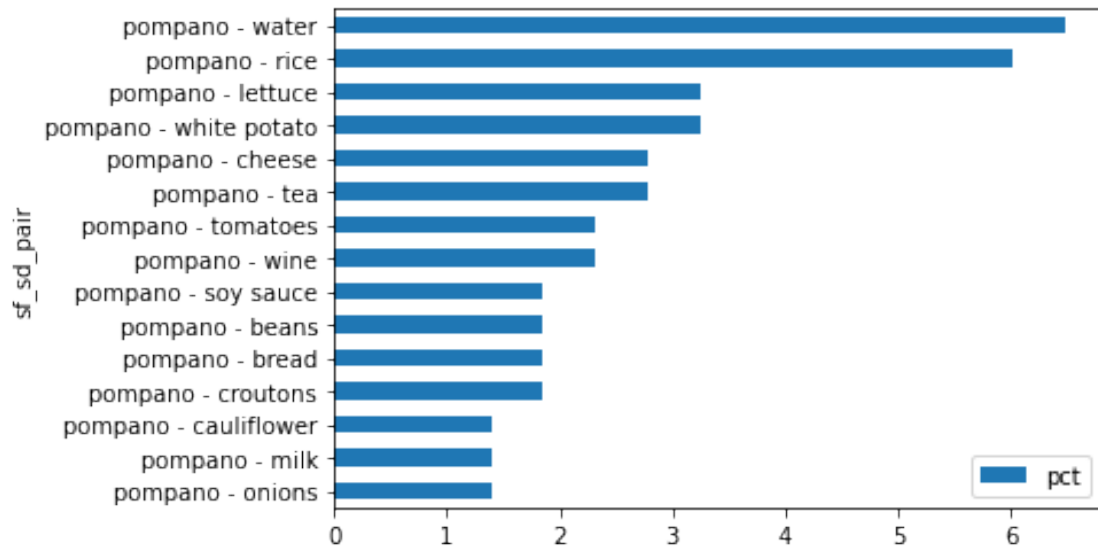


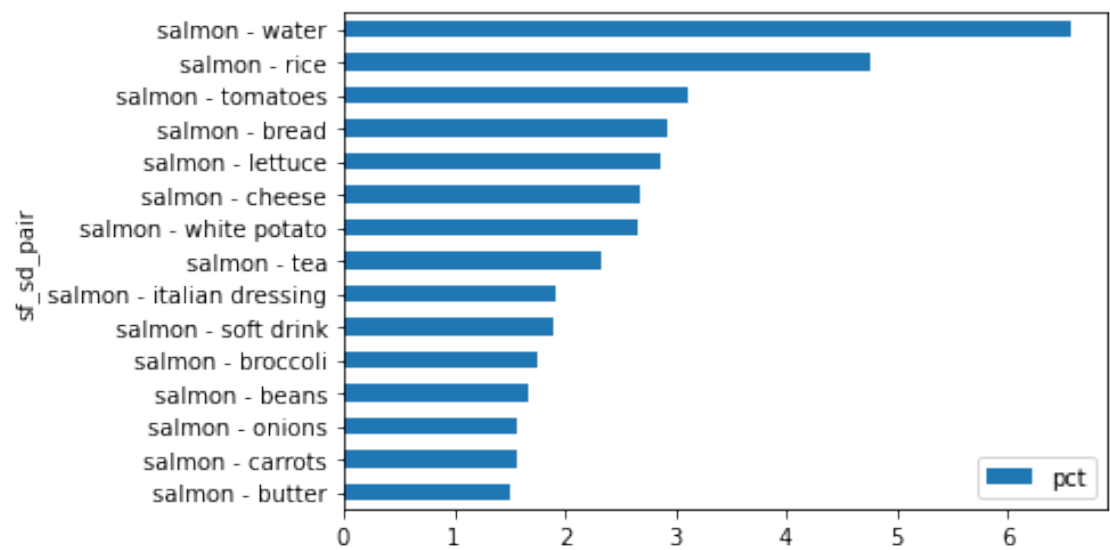
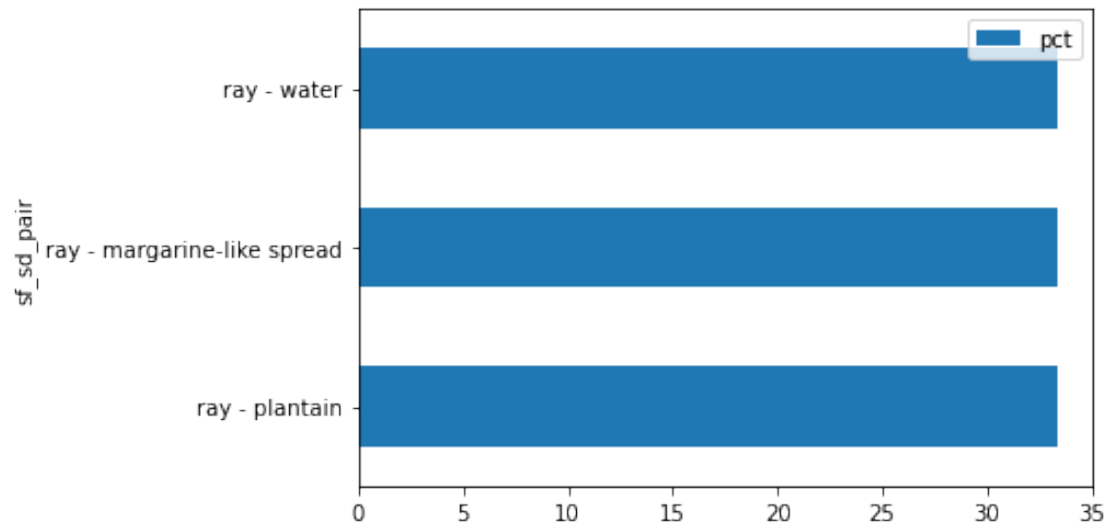


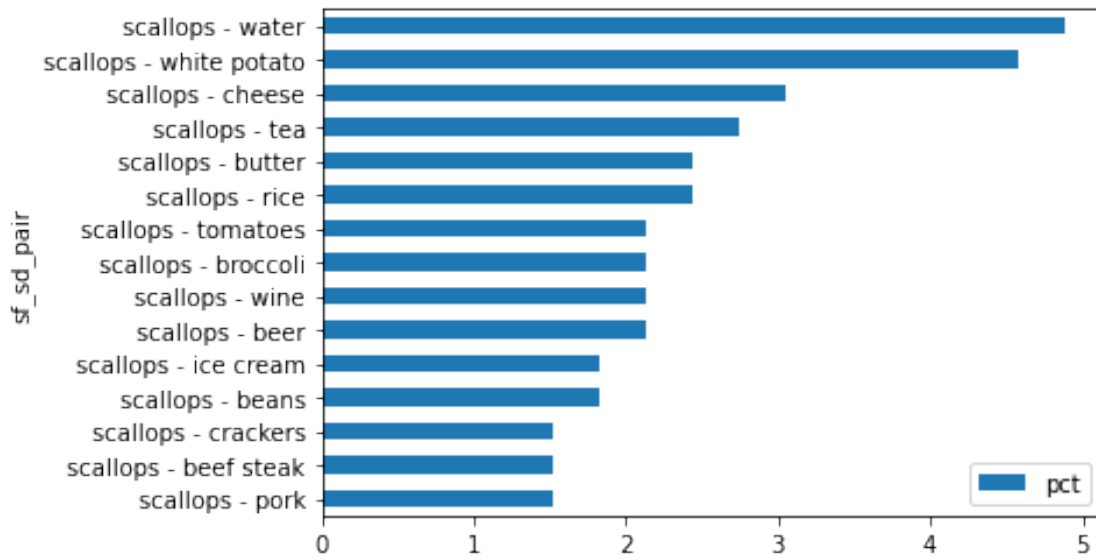
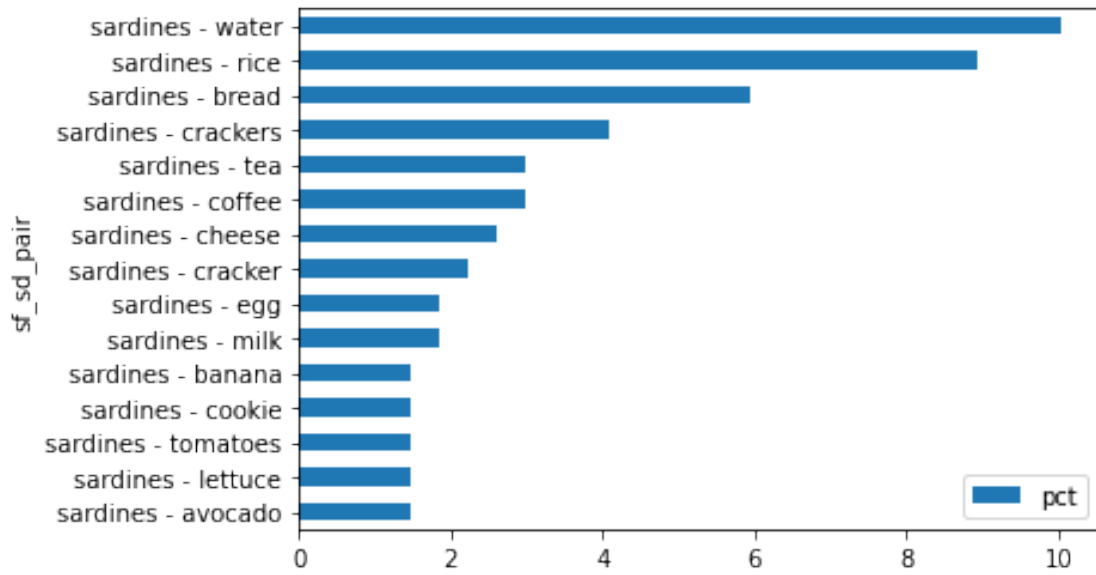


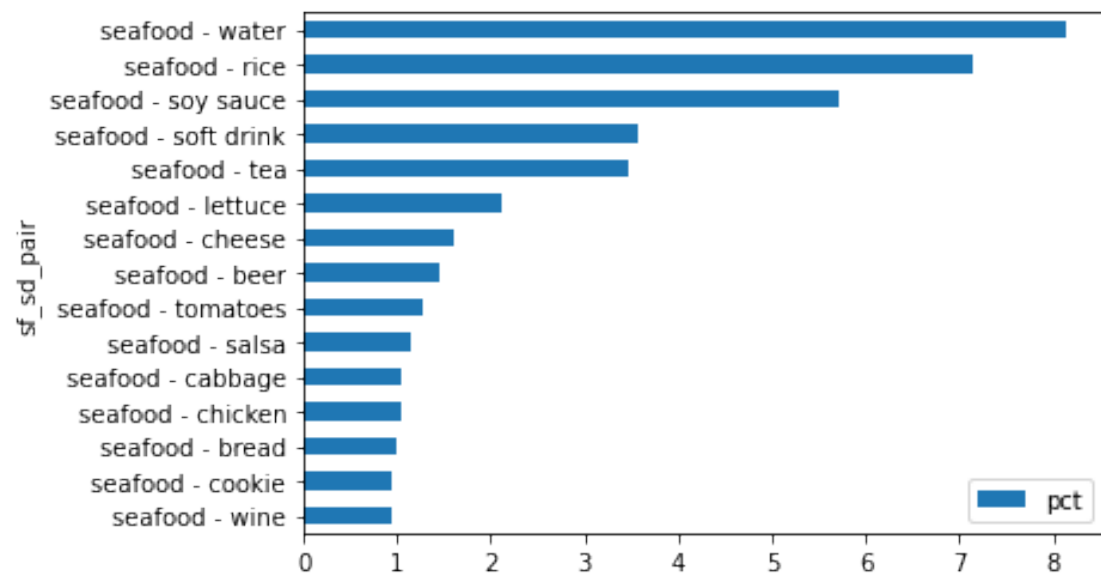
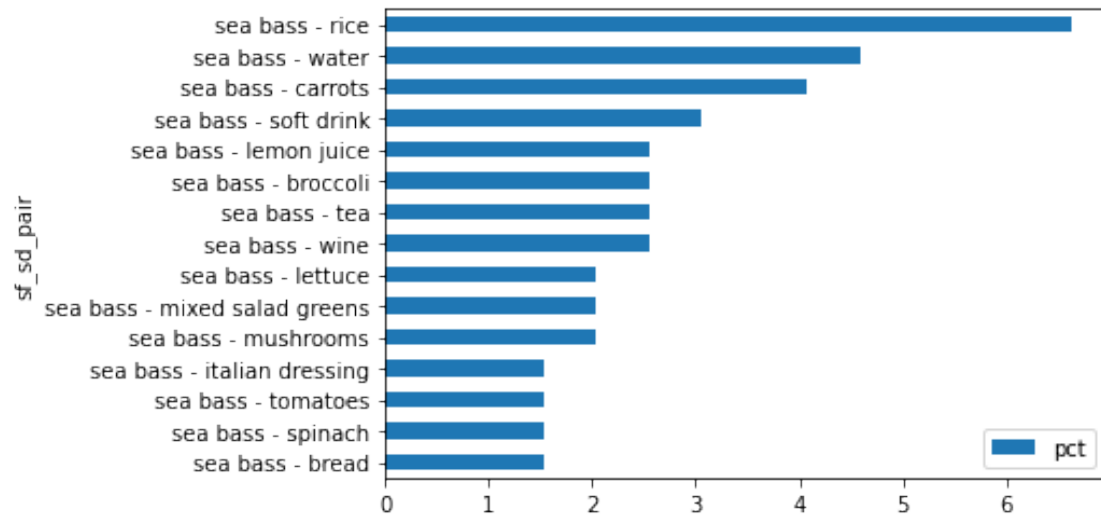


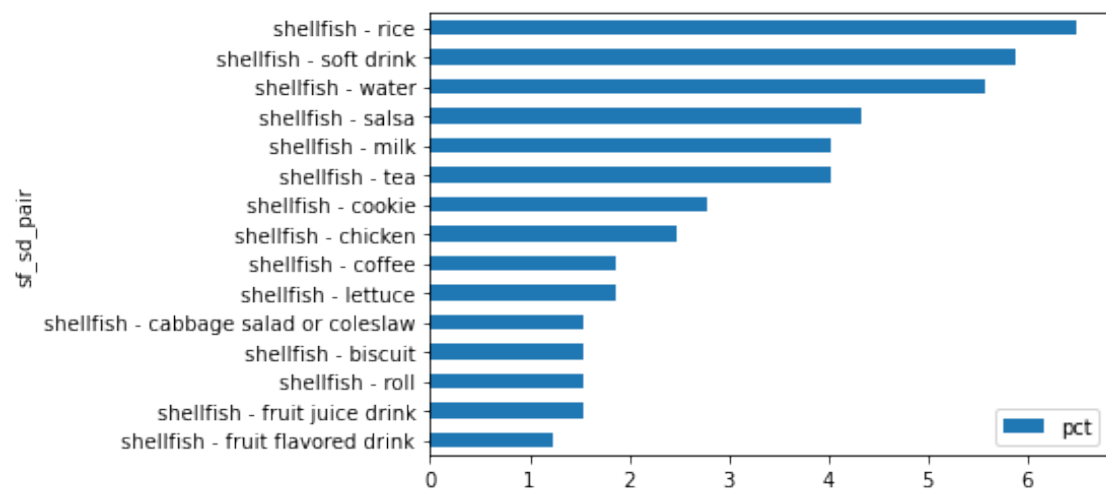
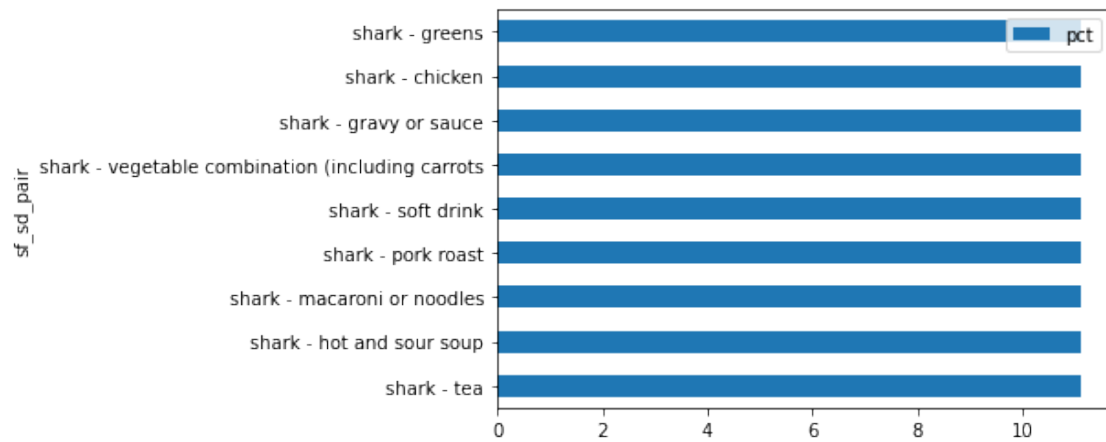
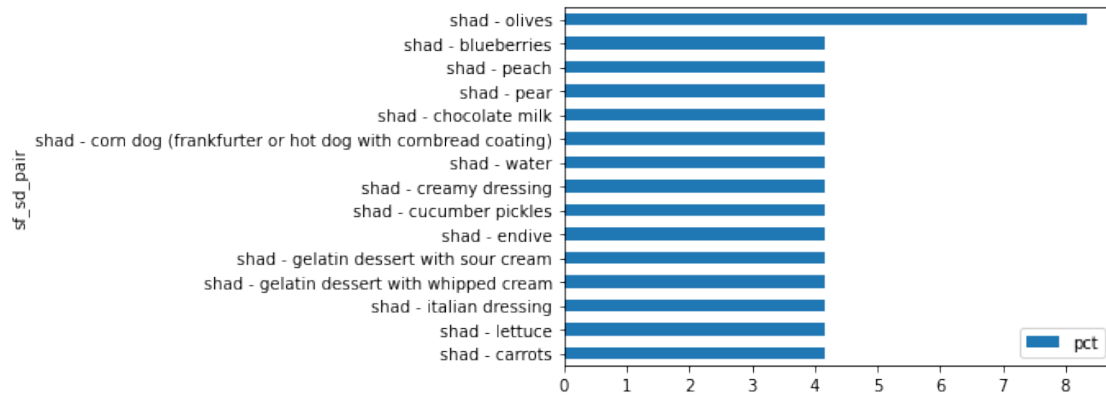


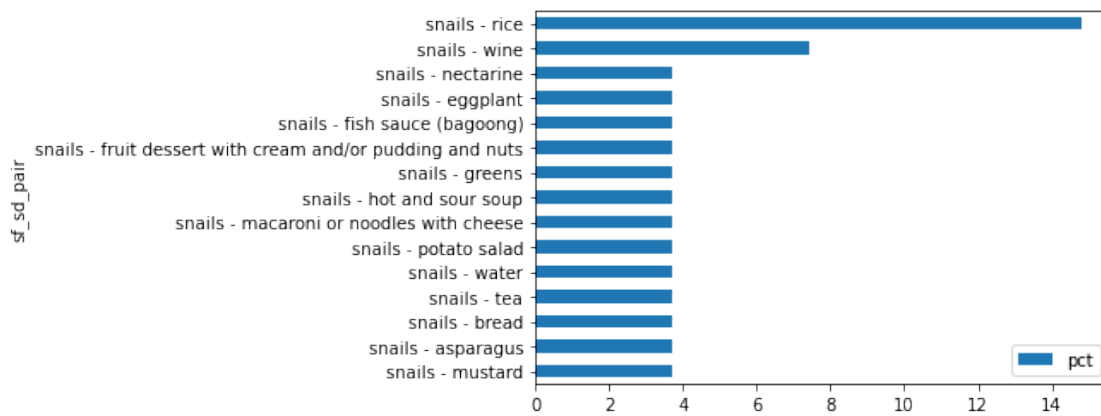
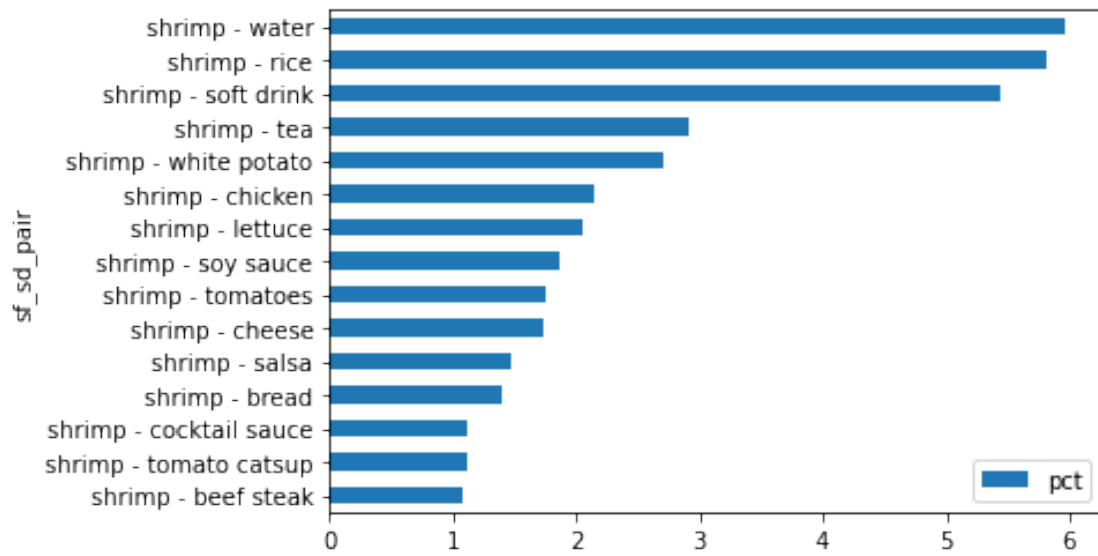


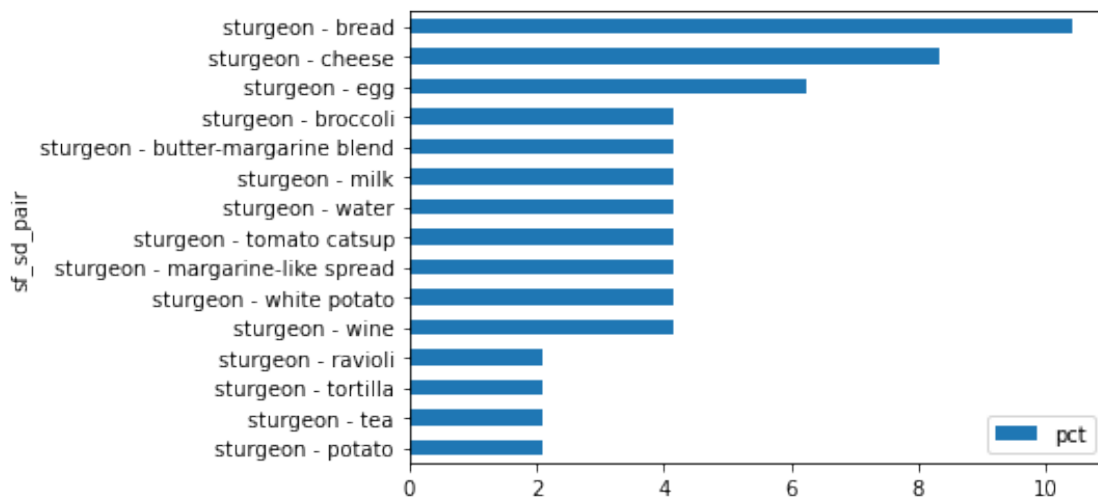
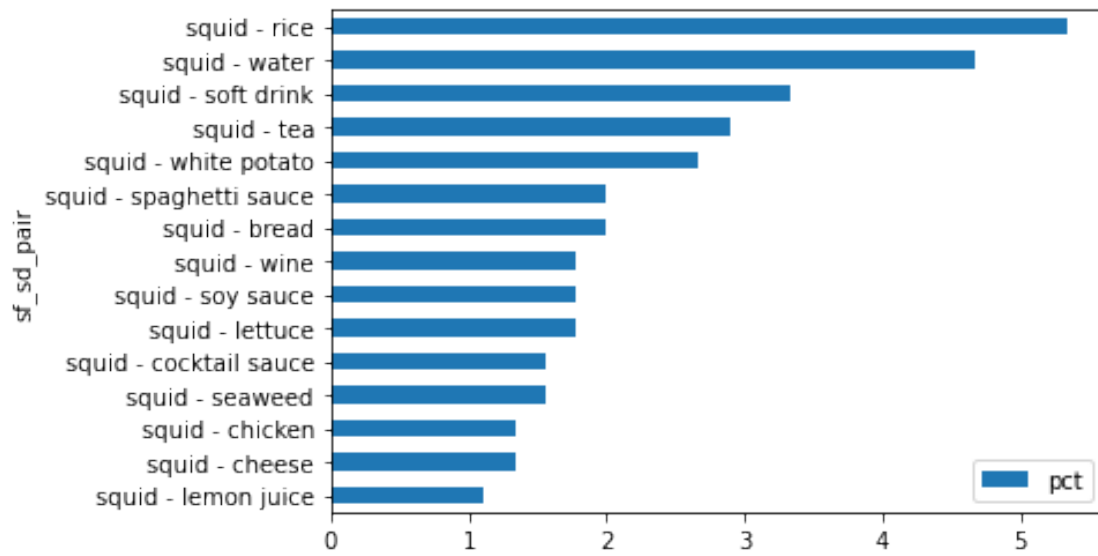


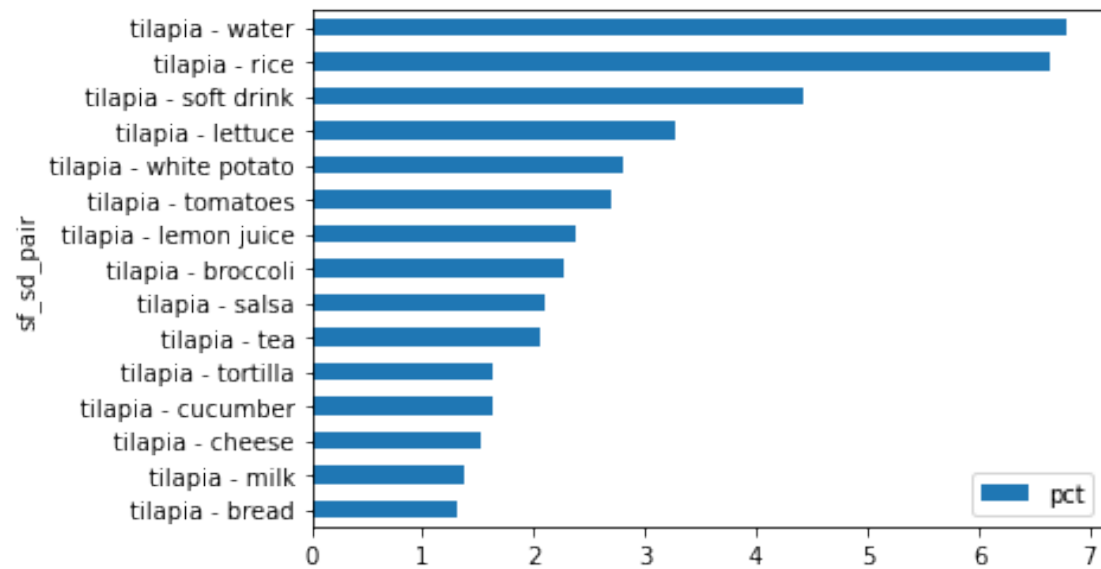
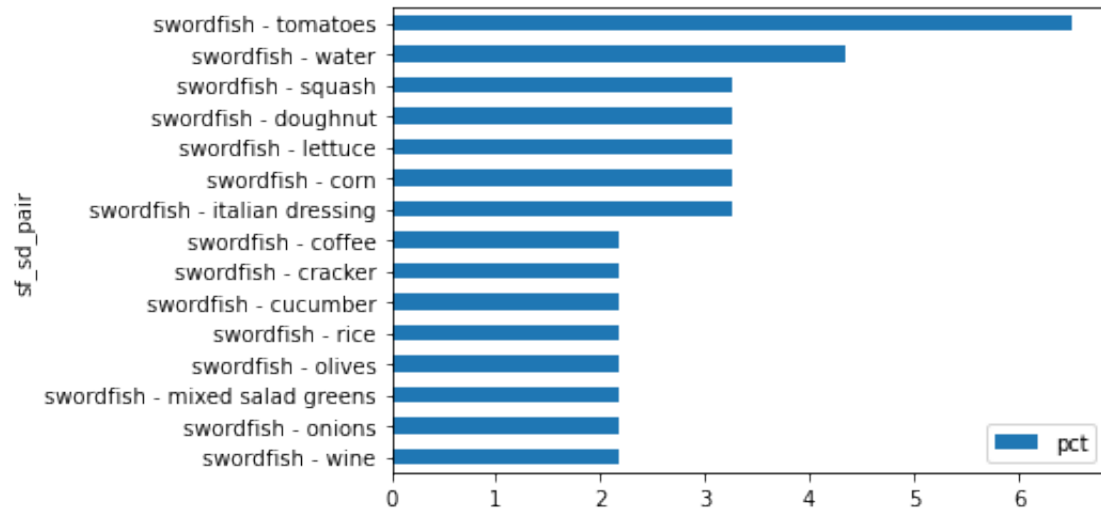


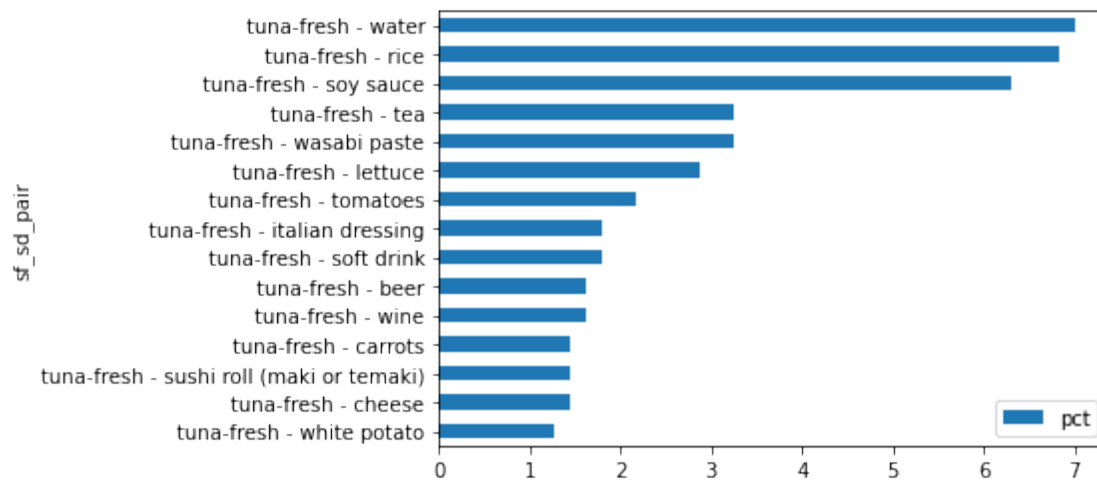
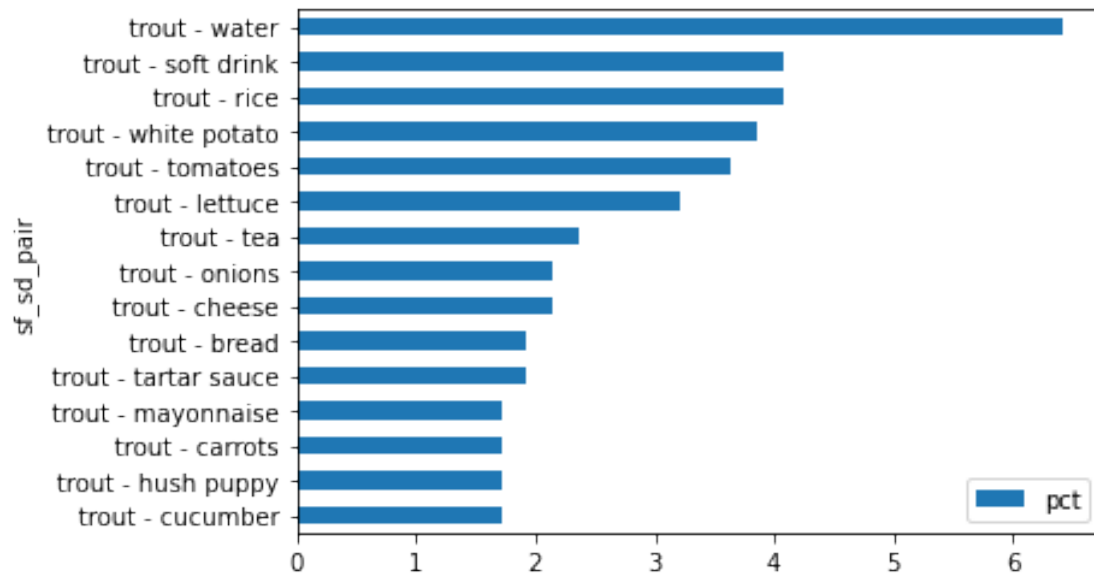


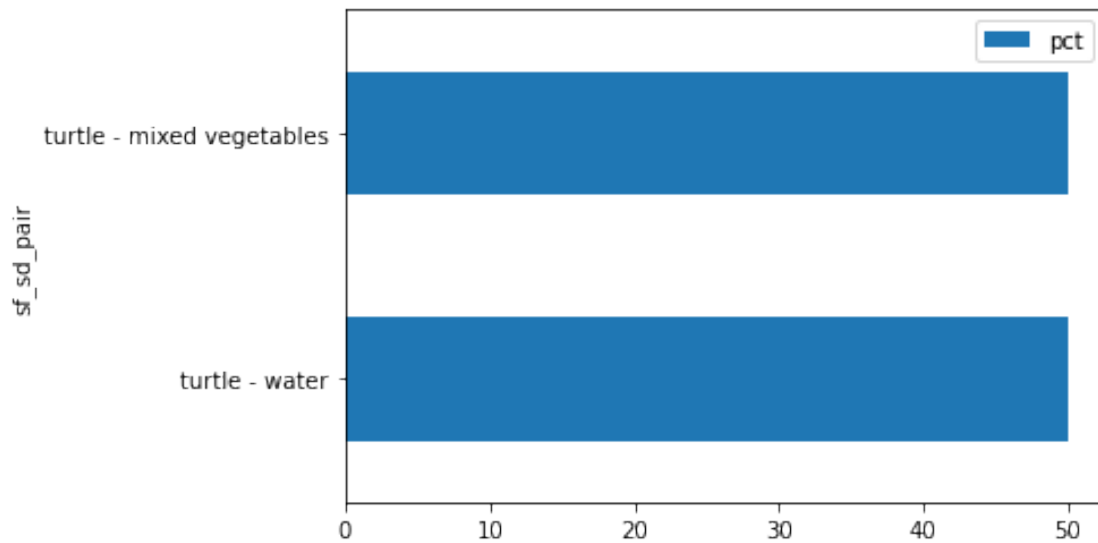
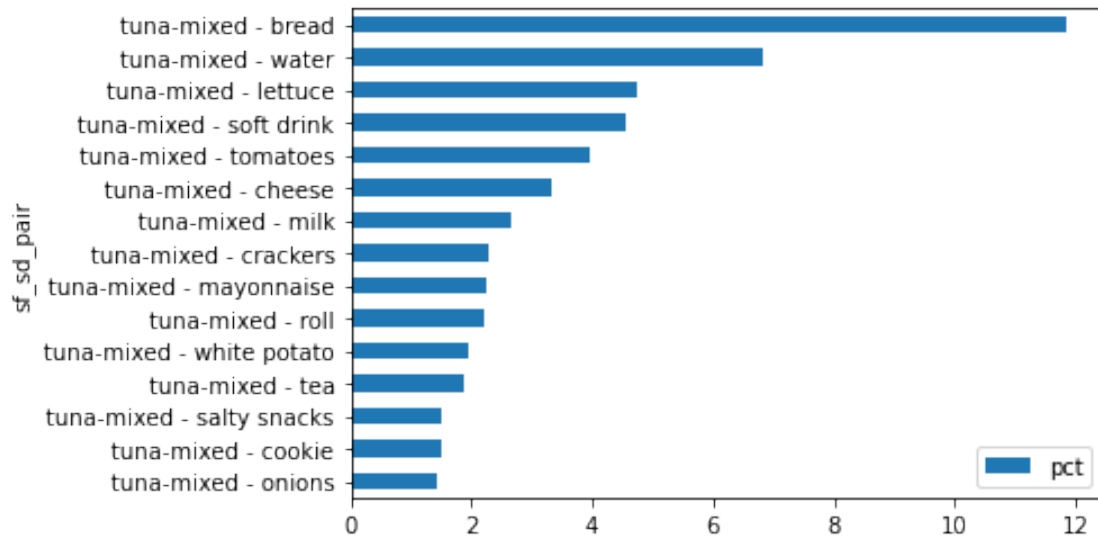


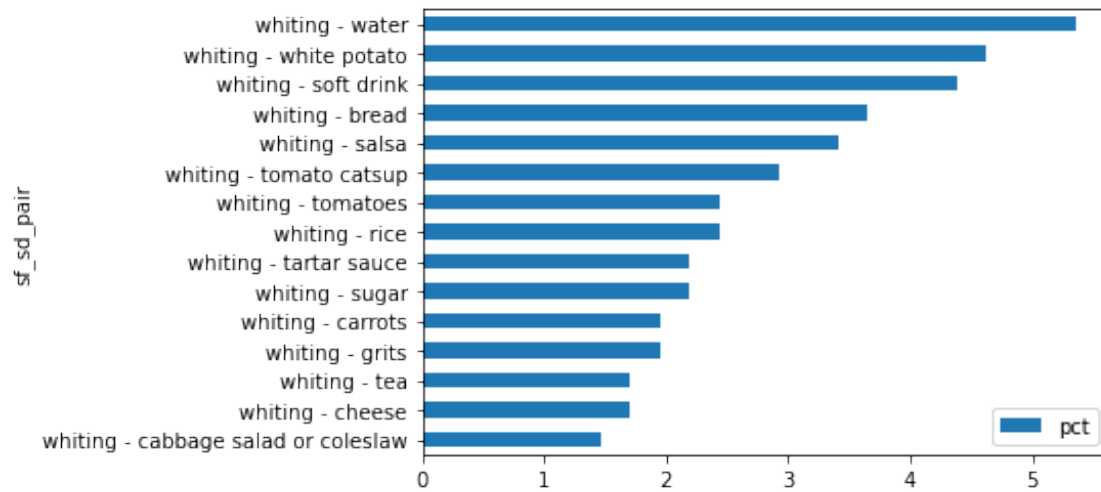












[]: