Data Cleaning and Transformation with Diseases & Symptom Dataset

In [ ]:

```
import pandas as pd
import numpy as np
```

loading data

In [ ]:

```
data = pd.read_csv('datasets/dataset.csv')
severity = pd.read_csv('datasets/Symptom-severity.csv')
descriptions = pd.read_csv('datasets/symptom_Description.csv')
precautions = pd.read_csv('datasets/symptom_precaution.csv')
```

shape and columns of each dataset

In [ ]:

```
print(data.columns)
print(data.shape)
```

In [ ]:

```
print(severity.columns)
print(severity.shape)
```

In [ ]:

```
print(descriptions.columns)
print(descriptions.shape)
```

In [ ]:

```
print(precautions.columns)
print(precautions.shape)
```

viewing data

In [ ]:

```
data.head(3)
```

In [ ]:

```
severity.head(3)
```

In [ ]:

```
descriptions.head(3)
```

In [ ]:

```
precautions.head(3)
```

summary statistics

In [ ]:

```
data.describe()
```

In [ ]:

```
severity.describe()
```

descriptions.describe()

In [ ]:

```
precautions.describe()
```

Data Transformation

- combining all datasets into one

repalcing null values with 0

In [ ]:

```
data.fillna(0, inplace=True)
data.head(2)
```

adding all symptoms to a single list

In [ ]:

```
ordered_symptoms = []

for i in range(len(data['Disease'])):
    temp_list = []
    for k in range(1,17):
        if data.iloc[i][k] == 0:
            break
        temp_list.append(data.iloc[i][k])
    ordered_symptoms.append(temp_list)
```

capitalize Diseases

In [ ]:

```
data['Disease']=data['Disease'].str.capitalize()
```

Using sorting methods to sort all columns by disease

In [ ]:

```
# lit of all diseases alphabetically

diseases = sorted(data['Disease'].unique())
```

In [ ]:

```
# sorting columns of dataset
descs = descriptions.sort_values(by='Disease')
pre_c = precautions.sort_values(by='Disease')
```

adding all precautions to one list

In [ ]:

```
ordered_cautions = []
for i in range(len(pre_c['Disease'])):
    temp_list = []
    for k in range(1,5):
        temp_list.append(pre_c.iloc[i][k])
    ordered_cautions.append(temp_list)
```

creating a dictionary with diesease and symptoms

In [ ]:

```
#Dictionary to hold disease (keys) and its values (a list of the symptoms)
disease_dict = {}

for i in range(len(data['Disease'])):
    symptoms_list = []
    for k in range(len(data.columns)):
        if data.iloc[i][k] == 0 or data.iloc[i][k] in disease_dict.keys():
            continue
        symptoms_list.append(data.iloc[i][k])
    disease_dict[data['Disease'][i]] = symptoms_list
```

sorting disease and symptom

In [ ]:

```
# sorting diseases
sorted_keys = sorted(disease_dict.keys())
```

In [ ]:

```
# sorting symptoms in accordance with key(disease)
symptoms_list = []
for i in range(len(sorted_keys)):
    symptoms_list.append(disease_dict[sorted_keys[i]])
```

Assembling everything into a single dataframe

In [ ]:

```
df = pd.DataFrame({"Diseases":diseases,"Descriptions":descs['Description'],
"Precautions":ordered_cautions, "Symptoms":symptoms_list})
```

setting index values

In [ ]:

```
index=np.arange(1,len(df)+1)
df.set_index(index, inplace=True)
```

End Result

In [ ]:

```
df
```