Created by Axel Garcia

In [ ]:
```
import pandas as pd
import glob
```

Grabb merged .txt file containing all committee info. Remove unnecessary columns, we only need committee id and party columns. Additionally, drop all rows with NA.

In [ ]:
```
path = r'/Users/axelgarcia/Documents/CSE 184/usafec.nosync/Data/committees.csv' # use your path
df = pd.read_csv(path, header=0, sep = ",",error_bad_lines=False)
cols = [1,2,3,4,5,6,7,8,9,11,12,13,14]
df.drop(df.columns[cols],axis=1,inplace=True)
df.dropna(inplace =True)
df
```

Remove all rows with unknown party affiliation

In [ ]:
```
df = df[~df["10"].str.contains("UNK")]
df
```

Remove junk party affiliations.

In [ ]:
```
df = df[~df["10"].str.contains('  0')]
df = df[~df["10"].str.contains('\.')]
```

Add column names.

In [ ]:
```
df.columns = ['cmte_id', 'party']
```

In [ ]:
```
df
```

Map known typos to correct values.

In [ ]:
```
replacement = {
    "Rep": "REP",
    "r": "REP",
    "R": "REP",
    "GOP": "REP",
    "rep": "REP",
    "d" : "DEM",
    "D" : "DEM",
    'dem' : "DEM",
    'Dem' : 'DEM',
    'NDP' : "DEM",
    'N':"NON",
    "NPA": "NON",
    "NOP":"NON",
    "NNE": "NON",
}

df = df.replace(replacement)
```

Finally, drop duplicates, keeping the latest party affiliation. Then save this data.

In [ ]:
```
df.drop_duplicates(subset='cmte_id', keep="last",inplace =True)
```

In [ ]:
```
df.to_csv('/Users/axelgarcia/Documents/CSE 184/usafec.nosync/Data/cleanCommittees.csv', index=False)
```