In [1]:
```
import collections
import numpy as np
import pandas as pd
import re

from argparse import Namespace
```

In [2]:
```
args = Namespace(
    raw_dataset_csv="./surnames.csv",
    train_proportion=0.7,
    val_proportion=0.15,
    test_proportion=0.15,
    output_munged_csv="./surnames_with_splits.csv",
    seed=1337
)
```

In [3]:
```
# Read raw data
surnames = pd.read_csv(args.raw_dataset_csv, header=0)
```

In [4]:
```
surnames.head()
```

Out[4]:

	surname	nationality
0	Woodford	English
1	Coté	French
2	Kore	English
3	Koury	Arabic
4	Lebzak	Russian

In [5]:
```
# Unique classes
set(surnames.nationality)
```

Out[5]:
```
{'Arabic',
 'Chinese',
 'Czech',
 'Dutch',
 'English',
 'French',
 'German',
 'Greek',
 'Irish',
 'Italian',
 'Japanese',
 'Korean',
 'Polish',
 'Portuguese',
 'Russian',
 'Scottish',
 'Spanish',
 'Vietnamese'}
```

In [6]:
```
# Splitting train by nationality
# Create dict
by_nationality = collections.defaultdict(list)
for _, row in surnames.iterrows():
    by_nationality[row.nationality].append(row.to_dict())
```

In [7]:
```
# Create split data
final_list = []
np.random.seed(args.seed)
for _, item_list in sorted(by_nationality.items()):
    np.random.shuffle(item_list)
    n = len(item_list)
    n_train = int(args.train_proportion*n)
    n_val = int(args.val_proportion*n)
    n_test = int(args.test_proportion*n)

    # Give data point a split attribute
    for item in item_list[:n_train]:
        item['split'] = 'train'
    for item in item_list[n_train:n_train+n_val]:
        item['split'] = 'val'
    for item in item_list[n_train+n_val:]:
        item['split'] = 'test'

    # Add to final list
    final_list.extend(item_list)
```

In [8]:
```
# Write split data to file
final_surnames = pd.DataFrame(final_list)
```

In [9]:
```
final_surnames.split.value_counts()
```

Out[9]:
```
train    7680
test     1660
val      1640
Name: split, dtype: int64
```

In [10]:
```
final_surnames.head()
```

Out[10]:

	surname	nationality	split
0	Totah	Arabic	train
1	Abboud	Arabic	train
2	Fakhoury	Arabic	train
3	Srour	Arabic	train
4	Sayegh	Arabic	train

In [11]:
```
# Write munged data to CSV
final_surnames.to_csv(args.output_munged_csv, index=False)
```

In [ ]:
```

```