NLP 447 Programing Project 1 Strategy

- **Parse input.txt to Color and Object arrays**. The element in Color array is a letter, the element in Object array is a word. The target is to use this input to predict the whole word of color

```
color, objects
```

```
(['b', 'b', 'g', 'g', 'r', 'b', 'w'],
 ['ribbon', 'hair', 'hair', 'eye', 'lip', 'box', 'teeth'])
```

- **Use *matplotlib.colors.CSS4_COLORS* as color set**. All the prediction color are from this list, which contains 148 colors.

```
[ ]  import matplotlib.colors as mcolors
```

```
[191] len(mcolors.CSS4_COLORS)
```

```
148
```

```
colors_set = list(mcolors.CSS4_COLORS)
colors_set
```

```
['aliceblue',
 'antiquewhite',
 'aqua',
 'aquamarine',
 'azure',
 'beige',
 'bisque',
 'black',
 'blanchedalmond',
 'blue',
```

- Use **FreqDist** from *nltk.probability* to **count frequencies** of **2grams from Brwon.words()** **{(k[0], k[1]): value, …..}**

```
fdist.items()
```

```
dict_items([(('The', 'Fulton'), 1), (('Fulton', 'County'), 6), (('County', 'Grand'), 1),
```

- **Test 1:**
  1. For each element of Color and Object arrays compare with **FreqDist** keys, if object[i] = k[1], k[0] in the colors_set, and color[i] = k[0][0] (the first letter of k[0]). Then append this **FreqDist** key and frequency (k, v) to candidate's dictionary.
  2. For above condition is False, just add the original object[i], color[i] to candidates dictionary, count frequency as 1.
  3. Compute the frequency for each candidate, pick the keys from dictionary with highest value

- **Test 2: (Bonus)**
  4. If the prediction color is not in the colors_set, then check if prediction color (probably is still a letter) is in any color word of colors_set, and this color word is also appear in **FreqDist** key, add this (k[0], k[1]: value) to a dictionary called ‘others’, get the k[0], k[1] which has highest value
  5. Update this key and value (k[0], k[1]: value) to the candidate dictionary
  6. Predict the output accordingly

```
(base) dhcp-10-5-46-84:project1 wayoo$ python3 p1_ywang340.py input.txt output.txt
['blue ribbon', 'black hair', 'gray hair', 'green eye', 'red lip', 'beige box', 'white teeth', 'green frog']
```