Descriptive Statistics

In [17]:
```
import string
from string import punctuation

# defining function to calculate number of tokens, unique tokens, number of characters, and lexical diversity.

def descriptive_stats(tokens, top_num_tokens = 5, verbose=True) :
    """
        Given a list of tokens, print number of tokens, number of unique tokens,
        number of characters, lexical diversity (https://en.wikipedia.org/wiki/Lexical_diversity),
        and num_tokens most common tokens. Return a list with the number of tokens, number
        of unique tokens, lexical diversity, and number of characters.
    """

    # Fill in the correct values here.

    num_tokens = len(tokens)
    num_unique_tokens = len(set(tokens))
    lexical_diversity = num_unique_tokens/num_tokens
    num_characters = len("".join(tokens))

    if verbose :
        print(f"There are {num_tokens} tokens in the data.")
        print(f"There are {num_unique_tokens} unique tokens in the data.")
        print(f"There are {num_characters} characters in the data.")
        print(f"The lexical diversity is {lexical_diversity:.3f} in the data.")
```

In [18]:
```
# descriptive statistics for gavin newsom
descriptive_stats(df.loc[df['id'] == 'gavinnewsom']['clean_text'], verbose = True)
```

```
There are 3249 tokens in the data.
There are 3001 unique tokens in the data.
There are 392152 characters in the data.
The lexical diversity is 0.924 in the data.
```

In [19]:
```
# descriptive statistics for brian kemp
descriptive_stats(df.loc[df['id'] == 'briankempga']['clean_text'], verbose = True)
```

```
There are 3247 tokens in the data.
There are 3212 unique tokens in the data.
There are 446019 characters in the data.
The lexical diversity is 0.989 in the data.
```