counting_words_better

January 6, 2025

[1]:
```
import string
from collections import Counter
import nltk #natural lang. toolkit
from nltk.corpus import stopwords
import os
```

Get stopwords

[3]:
```
#nltk.download('stopwords')
stop = stopwords.words('english')
```

Create function to get all txt files in a specified directory

[5]:
```
def get_path():
    current = os.getcwd()

    txt_files = []
    direct = current
    for file in os.listdir(direct):
        if file.endswith('.txt'):
            txt_files.append(file)

    return txt_files
```

[7]:
```
all_files = get_path()
```

[9]:
```
all_files
```

[9]:
```
['4300.txt',
 'new_file.txt',
 '1322.txt',
 '161.txt',
 '2701.txt',
 '98.txt',
 '1342.txt',
 '1400.txt',
 '84.txt',
 '768.txt',
 '730.txt']
```

REGULAR CODE (NO FUNCTION)

[88]:
```
clean_words = []
total_words = {'word_count': int()}
no_stop = []

with open('730.txt', 'r') as f:

    file_ = f.read()

words = file_.lower().split()
punc = list(string.punctuation)


for word in words:
    for char in word:
        special_case = ["--", "-"]

    for i in special_case:
        if i in word:
            word = word.replace(i,"\n").strip()

        else:
            if any(symb in char for symb in punc):
                word = word.replace(char,'').strip()

    clean_words.append(word)

for i in clean_words:
    total_words['word_count']+=1

#print(total_words)

for word in clean_words:
    if not word in stop:
        if word:
            no_stop.append(word)

word_counts = Counter(no_stop)

most_common = sorted(word_counts.items(), key = lambda x: x[1], reverse =␣
 ↪True)[0]
```

[89]:
```
most_common
```

[89]:
```
('said', 1229)
```

FUNCTION ASSIGNMENT

[145]:
```
def get_count(txtfile):
    clean_words = []
    total_words = {'word_count': int()}
    no_stop = []

    with open(txtfile, 'r') as f:

        file_ = f.read()

    words = file_.lower().split()
    punc = list(string.punctuation)


    for word in words:
        for char in word:
            special_case = ["--", "-"]

        for i in special_case:
            if i in word:
                word = word.replace(i,"\n")

            else:
                if any(symb in char for symb in punc):
                    word = word.replace(char,'')

        clean_words.append(word)

    for i in clean_words:
        total_words['word_count']+=1

#print(total_words)

    for word in clean_words:
        if not word in stop:
            if word:
                no_stop.append(word)

    word_counts = Counter(no_stop)

    most_common = sorted(word_counts.items(), key = lambda x: x[1], reverse =␣
 ↪True)[0]

    return txtfile, total_words, most_common
```

[147]:
```
get_count('1322.txt')
```

[147]:
```
('1322.txt', {'word_count': 121693}, ('see', 419))
```

[149]:
```
get_count('730.txt')
```

[149]:
```
('730.txt', {'word_count': 157993}, ('said', 1229))
```

[151]:
```
get_count('4300.txt')
```

[151]:
```
('4300.txt', {'word_count': 264969}, ('said', 1207))
```

[189]:
```
def get_counts(allfiles):

    for file in allfiles:
        clean_words = []
        total_words = {'word_count': int()}
        no_stop = []

        print(file)
        with open(file, 'r') as f:

            file_ = f.read()

        words = file_.lower().split()
        punc = list(string.punctuation)


        for word in words:
            for char in word:
                special_case = ["--", "-"]

            for i in special_case:
                if i in word:
                    word = word.replace(i,"\n")

                else:
                    if any(symb in char for symb in punc):
                        word = word.replace(char,'')

            clean_words.append(word)

        for i in clean_words:
            total_words['word_count']+=1
        print(total_words)


        for word in clean_words:
            if not word in stop:
                if word:
                    no_stop.append(word)
```

```
        word_counts = Counter(no_stop)

        most_common = sorted(word_counts.items(), key = lambda x: x[1], reverse
↪= True)[0]

        print(most_common)
```

[191]:
```
get_counts(all_files)
```

```
4300.txt
{'word_count': 264969}
('said', 1207)
new_file.txt
{'word_count': 8}
('<_io.textiowrapper', 1)
1322.txt
{'word_count': 121693}
('see', 419)
161.txt
{'word_count': 118573}
('elinor', 597)
2701.txt
{'word_count': 212107}
('whale', 917)
98.txt
{'word_count': 135846}
('said', 658)
1342.txt
{'word_count': 121559}
('mr', 766)
1400.txt
{'word_count': 184450}
('said', 1337)
84.txt
{'word_count': 74968}
('one', 198)
768.txt
{'word_count': 115947}
('would', 441)
730.txt
{'word_count': 157993}
('said', 1229)
```

[ ]: