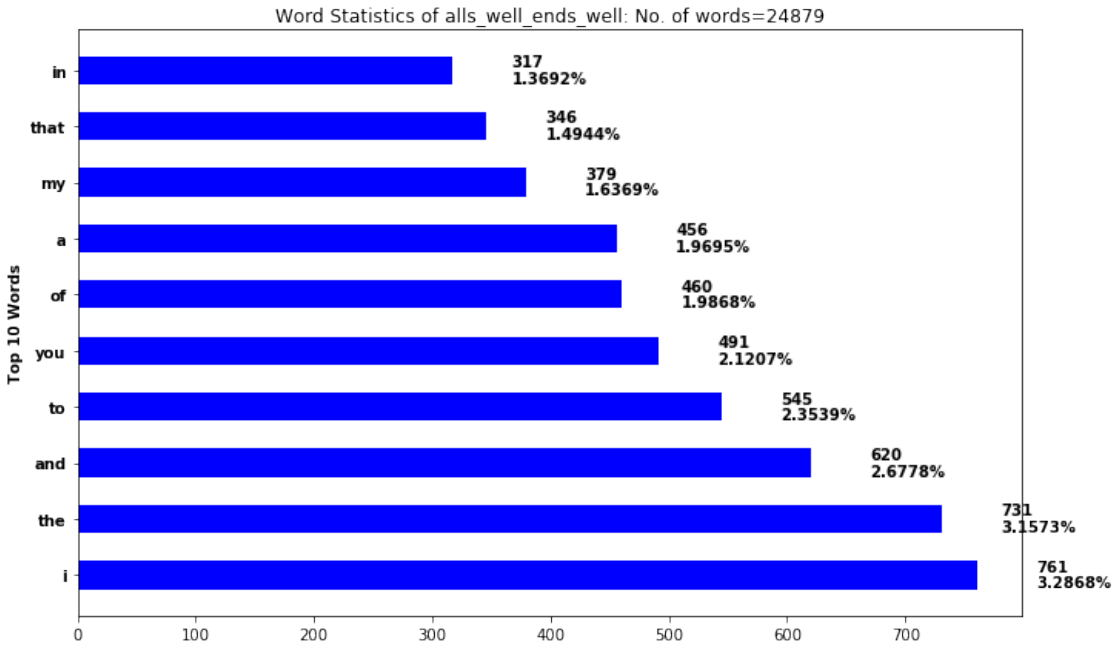
3-Statistics-of-Words

December 28, 2017

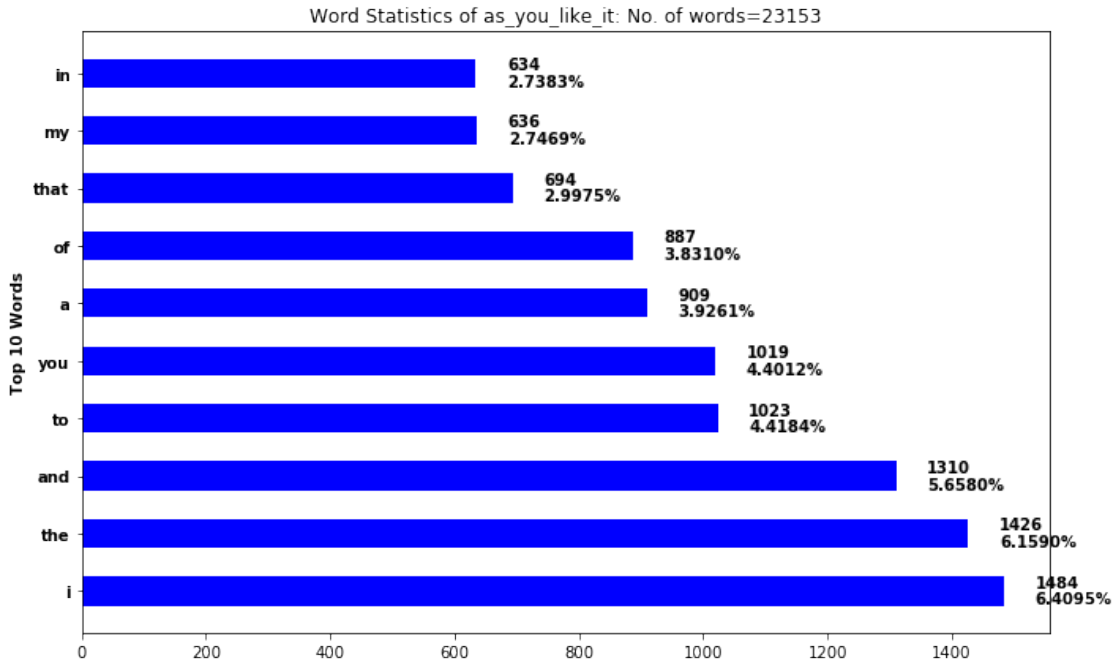
In [33]:
```
import numpy as np
import matplotlib.pyplot as plt
import matplotlib.mlab as mlab

import re
from collections import Counter

from os import listdir
from os.path import isfile, join
import codecs

path_shake = './complete_works/'
all_files = [f.replace('.txt','') for f in listdir(path_shake) if isfile(join(path_sha
n_files = len(all_files)
print("Total word numbers of Shakepeare's complete works: ", n_files)
print('\n')

# words for the complete work
words = []
N = 10 # top N
for i in range(n_files):
    file_i_path = './complete_works/'+ all_files[i]+'.txt'
    file_i = codecs.open(file_i_path, "r",encoding='utf-8', errors='ignore')
    text_i = file_i.read().lower()
    word_i = re.findall(r'\w+', text_i)
    words +=word_i
    word_i_count = len(word_i)
    vocab_i_count = len(set(word_i))
    top_N = Counter(words).most_common(N)
    print('-'*30 + str(i+1) + '-'*30)
    print(top_N)
    print('No. of words in %s is %d'%(all_files[i],word_i_count))
    print('No. of different words in %s is %d'%(all_files[i],vocab_i_count))

    top_word = []
    Nums= []
    for ele in top_N:
        top_word.append(ele[0])
```

```
            Nums.append(ele[1])

        # plot
        fig, ax = plt.subplots(figsize=(11, 7))
        index = np.arange(N)
        width = 0.5
        ax.barh(index, Nums,width,color="blue")
        for j, v in enumerate(Nums):
            ax.text(v + 50, j + .07, str(v), fontweight='bold')
            ax.text(v + 50, j - .22, str('%.4f'%(v/word_count*100))+'%', fontweight='bold
        ax.set_yticks(index)#+width/2)
        ax.set_yticklabels(top_word,fontweight='bold')
        plt.ylabel('Top %d Words'%N,fontweight='bold')
        plt.title('Word Statistics of %s: No. of words=%d'%(all_files[i],word_i_count))
        plt.show()

    words_count = len(words)
    vocabs_count = len(set(words))
    print('\nThe total number of words Shakespeare used in the complete work: ',words_cou
    print('\nThe total number of different words Shakespeare used in the complete work: '
```

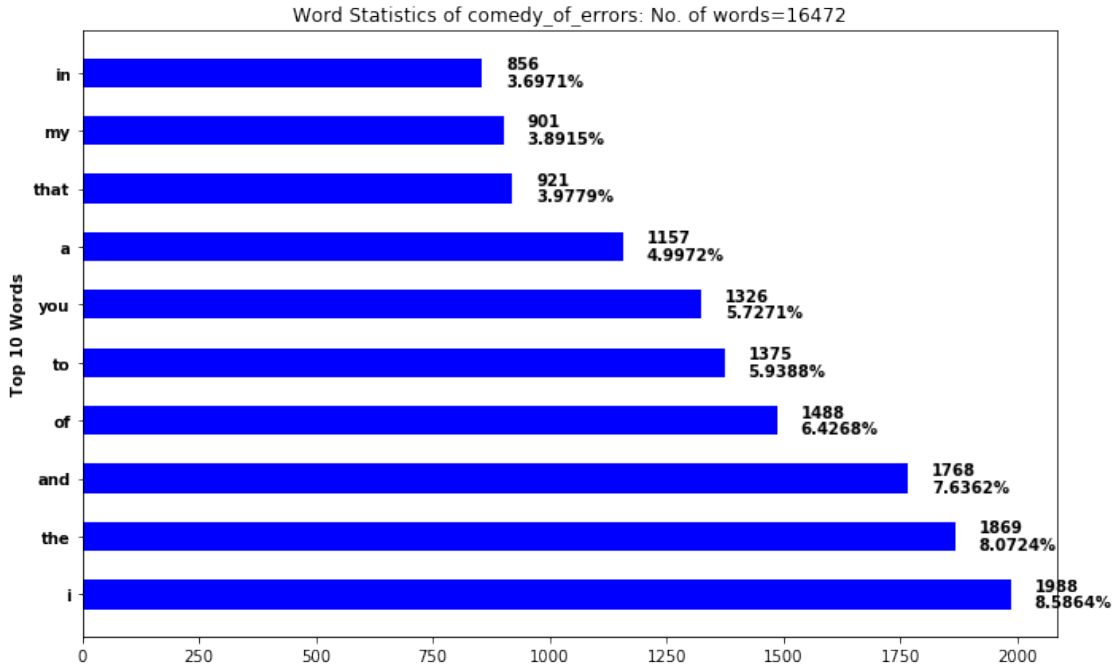
```
Total word numbers of Shakepeare's complete works:  42


------------------------------1------------------------------
[('i', 761), ('the', 731), ('and', 620), ('to', 545), ('you', 491), ('of', 460), ('a', 456), (
No. of words in alls_well_ends_well is 24879
No. of different words in alls_well_ends_well is 3381
```



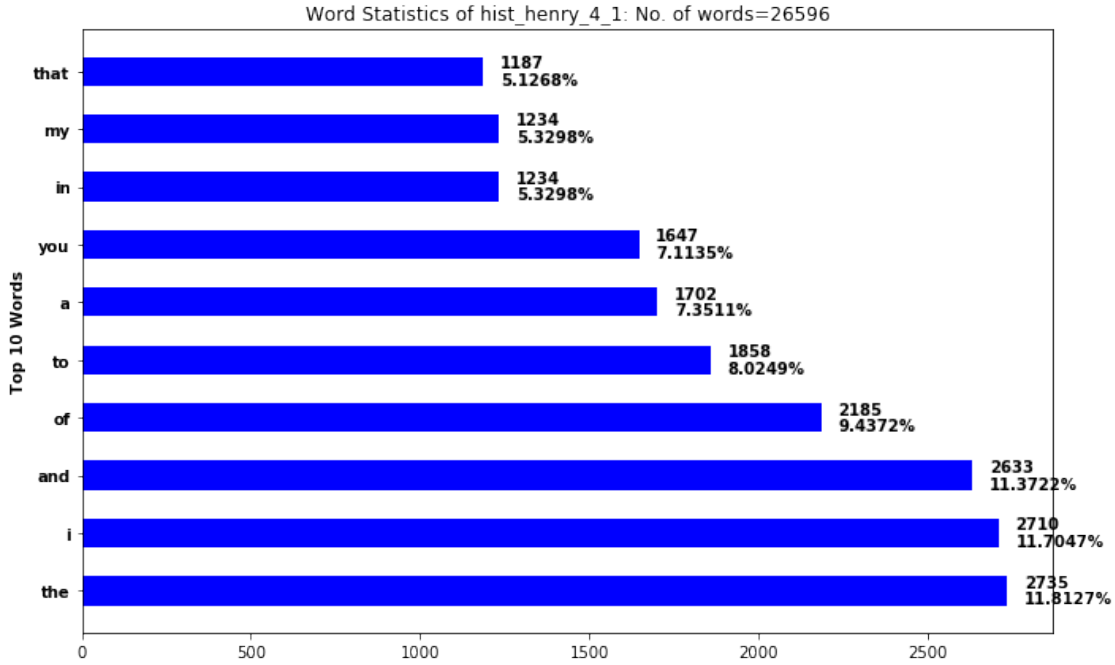
```
------------------------------2------------------------------
[('i', 1484), ('the', 1426), ('and', 1310), ('to', 1023), ('you', 1019), ('a', 909), ('of', 88
No. of words in as_you_like_it is 23153
No. of different words in as_you_like_it is 3166
```



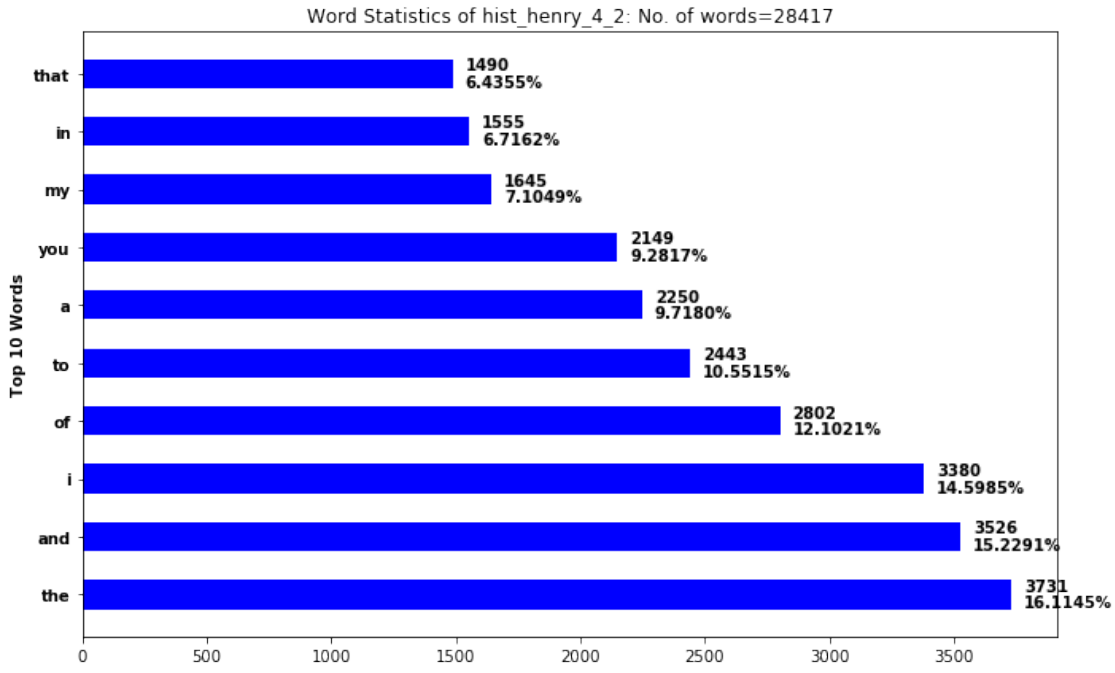

```
------------------------------3------------------------------
[('i', 1988), ('the', 1869), ('and', 1768), ('of', 1488), ('to', 1375), ('you', 1326), ('a', 1
No. of words in comedy_of_errors is 16472
No. of different words in comedy_of_errors is 2440
```



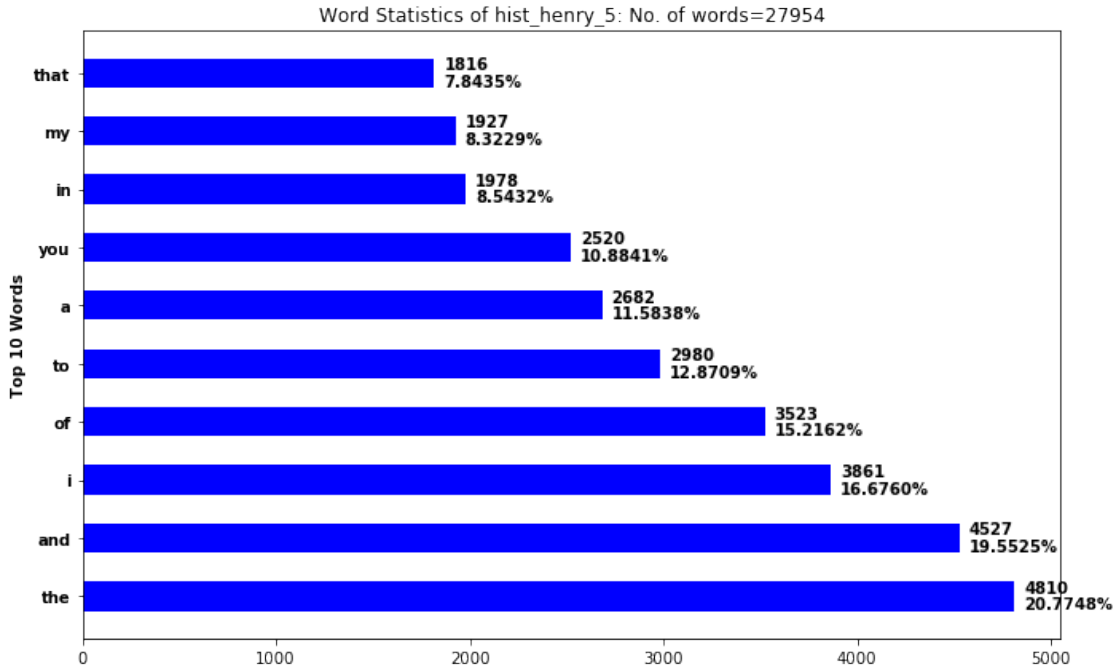

```
--------------------------------4--------------------------------
[('the', 2735), ('i', 2710), ('and', 2633), ('of', 2185), ('to', 1858), ('a', 1702), ('you', 16
No. of words in hist_henry_4_1 is 26596
No. of different words in hist_henry_4_1 is 3724
```



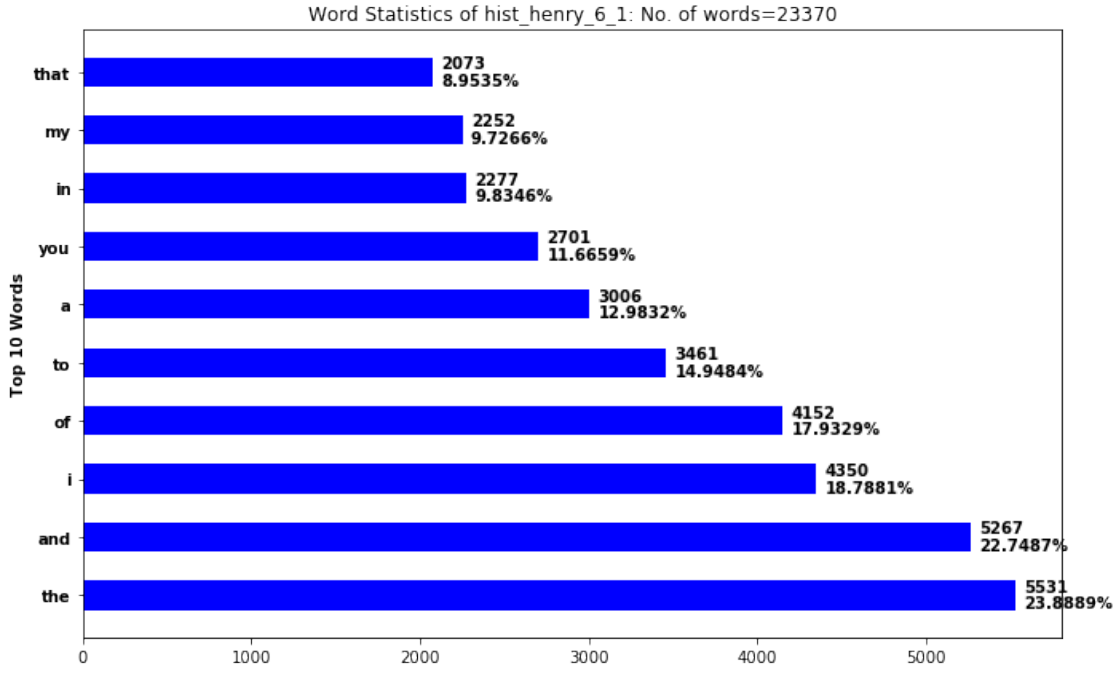
```
------------------------------5------------------------------
[('the', 3731), ('and', 3526), ('i', 3380), ('of', 2802), ('to', 2443), ('a', 2250), ('you', 2
No. of words in hist_henry_4_2 is 28417
No. of different words in hist_henry_4_2 is 3961
```



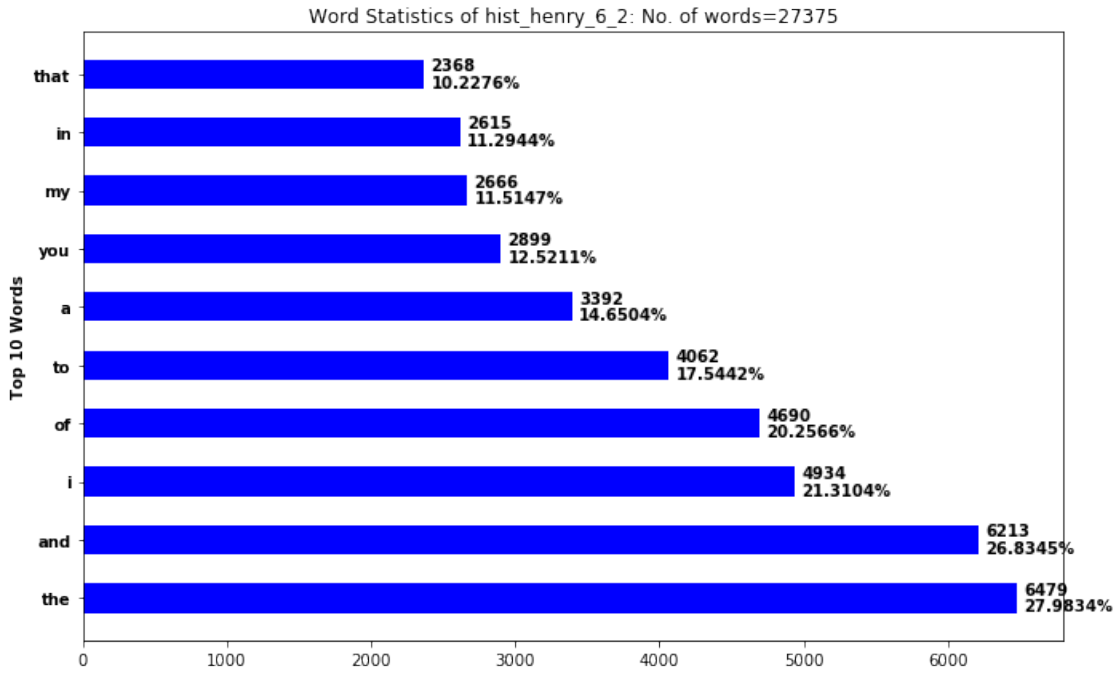

```
------------------------------6------------------------------
[('the', 4810), ('and', 4527), ('i', 3861), ('of', 3523), ('to', 2980), ('a', 2682), ('you', 25
No. of words in hist_henry_5 is 27954
No. of different words in hist_henry_5 is 4396
```



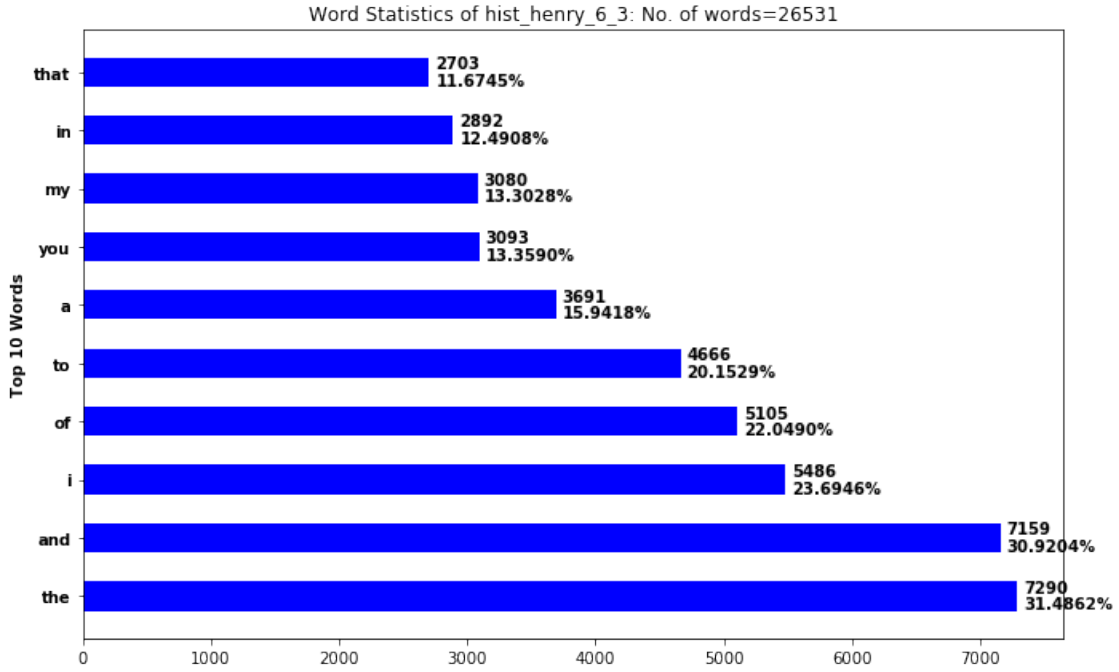

```
--------------------------------7--------------------------------
[('the', 5531), ('and', 5267), ('i', 4350), ('of', 4152), ('to', 3461), ('a', 3006), ('you', 2
No. of words in hist_henry_6_1 is 23370
No. of different words in hist_henry_6_1 is 3726
```



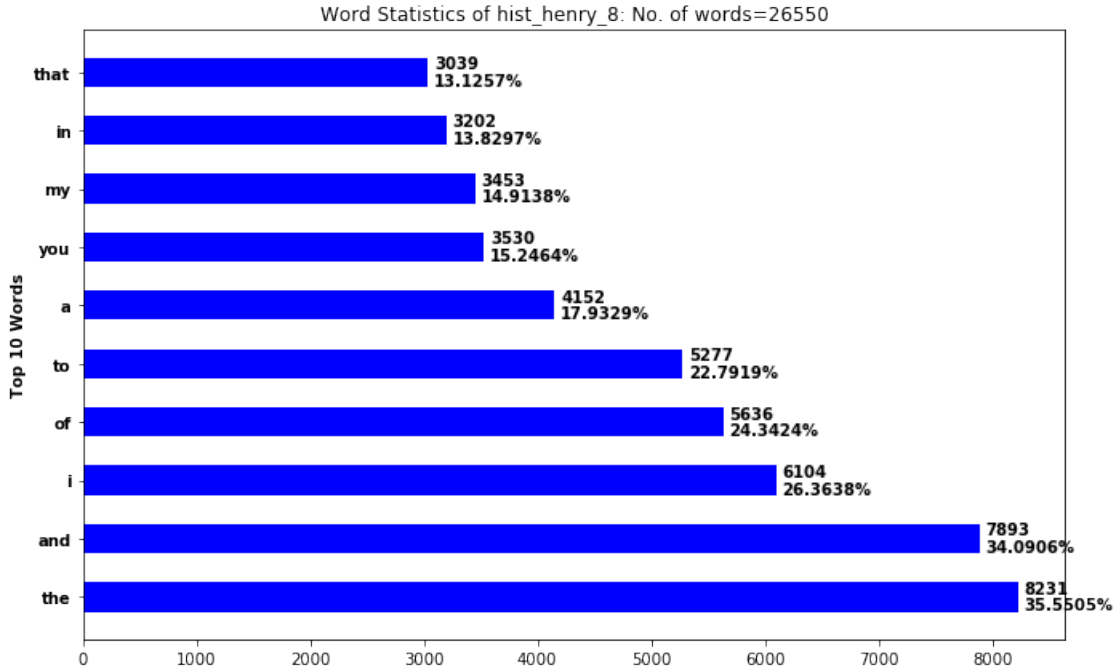
```
------------------------------8------------------------------
[('the', 6479), ('and', 6213), ('i', 4934), ('of', 4690), ('to', 4062), ('a', 3392), ('you', 28
No. of words in hist_henry_6_2 is 27375
No. of different words in hist_henry_6_2 is 3928
```



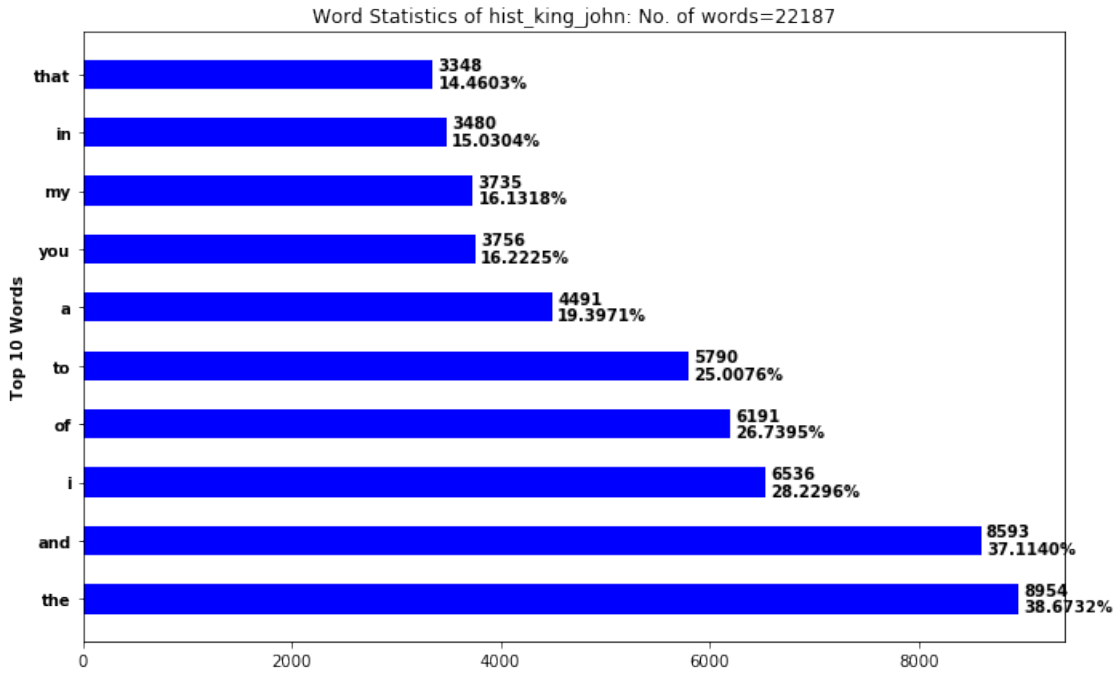

```
------------------------------9------------------------------
[('the', 7290), ('and', 7159), ('i', 5486), ('of', 5105), ('to', 4666), ('a', 3691), ('you', 30
No. of words in hist_henry_6_3 is 26531
No. of different words in hist_henry_6_3 is 3445
```



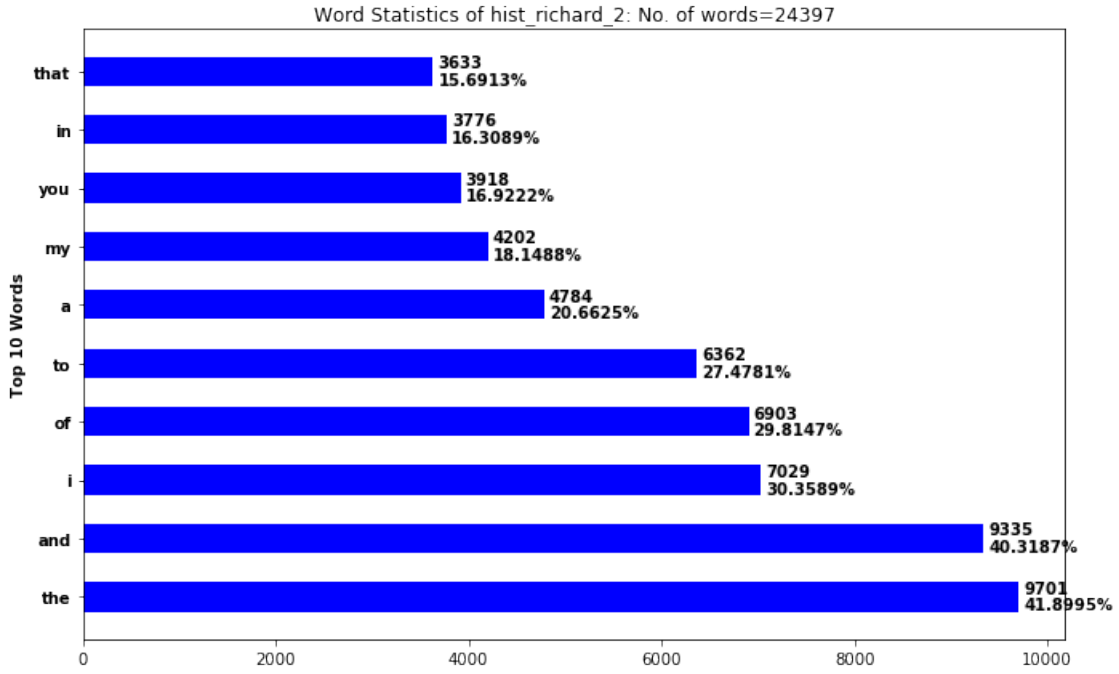
----------10----------
[('the', 8231), ('and', 7893), ('i', 6104), ('of', 5636), ('to', 5277), ('a', 4152), ('you', 35
No. of words in hist_henry_8 is 26550
No. of different words in hist_henry_8 is 3509



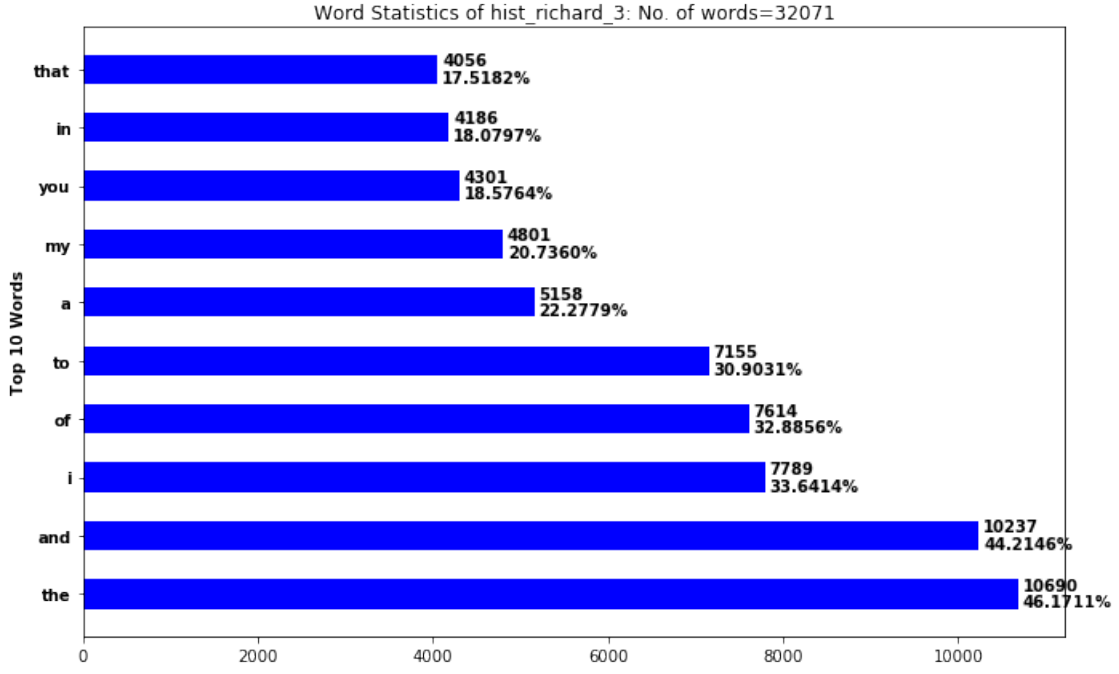
```
------------------------------11------------------------------
[('the', 8954), ('and', 8593), ('i', 6536), ('of', 6191), ('to', 5790), ('a', 4491), ('you', 3
No. of words in hist_king_john is 22187
No. of different words in hist_king_john is 3435
```



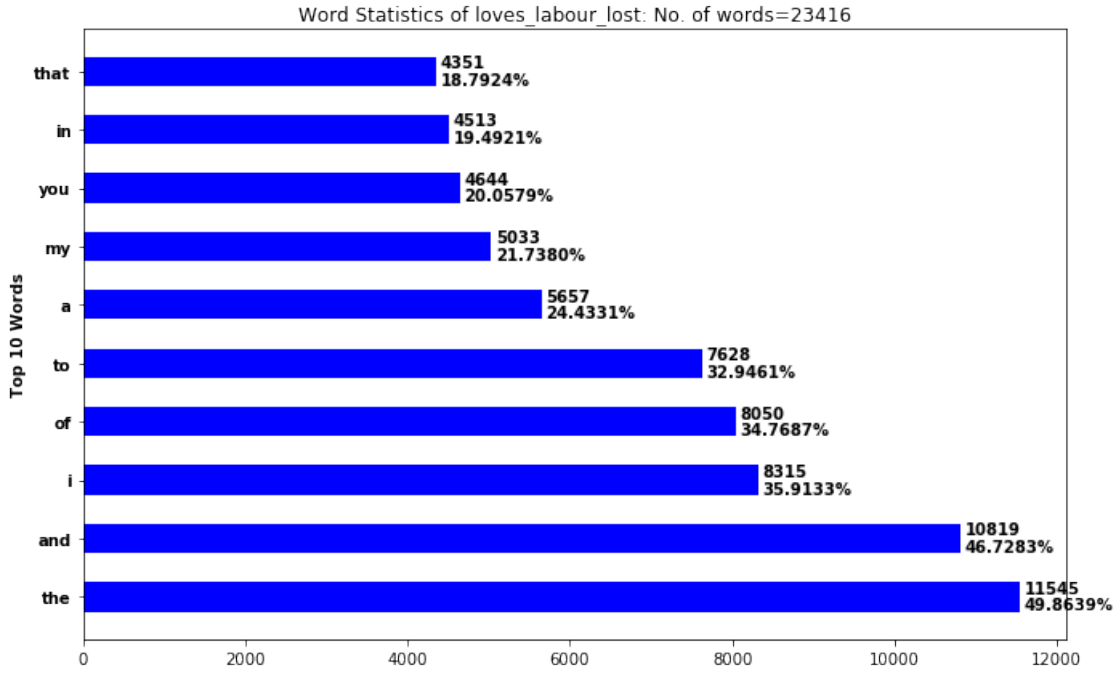

```
------------------------------12------------------------------
[('the', 9701), ('and', 9335), ('i', 7029), ('of', 6903), ('to', 6362), ('a', 4784), ('my', 42
No. of words in hist_richard_2 is 24397
No. of different words in hist_richard_2 is 3513
```



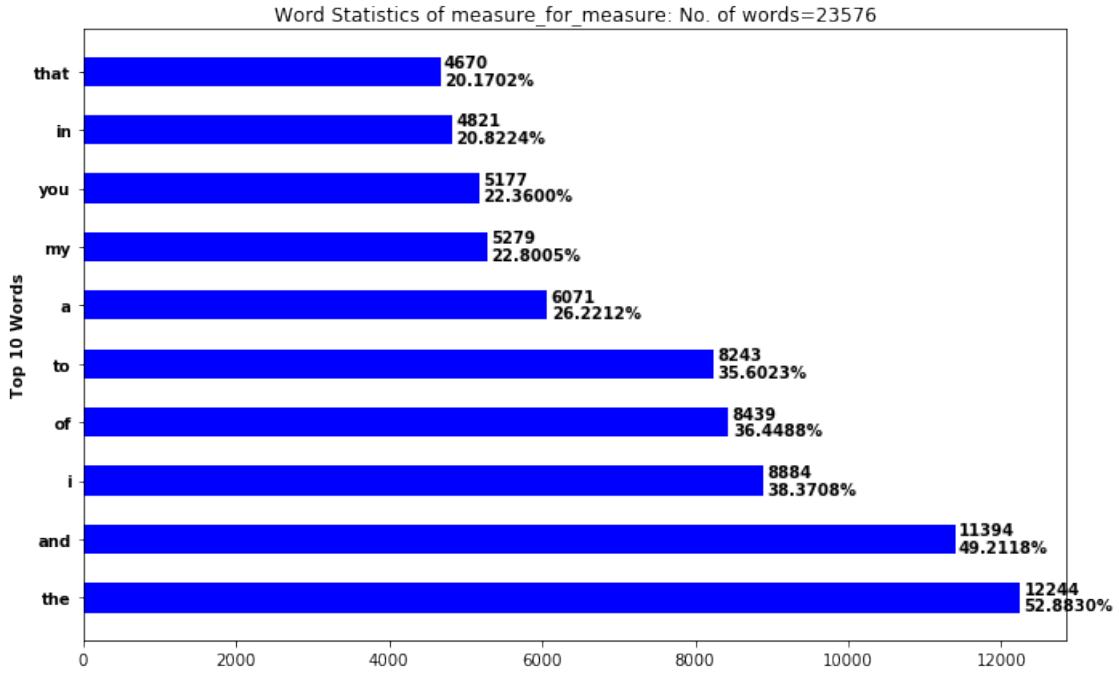

```
------------------------------13------------------------------
[('the', 10690), ('and', 10237), ('i', 7789), ('of', 7614), ('to', 7155), ('a', 5158), ('my', 4
No. of words in hist_richard_3 is 32071
No. of different words in hist_richard_3 is 3889
```



```
------------------------------14------------------------------
[('the', 11545), ('and', 10819), ('i', 8315), ('of', 8050), ('to', 7628), ('a', 5657), ('my', 5
No. of words in loves_labour_lost is 23416
No. of different words in loves_labour_lost is 3628
```

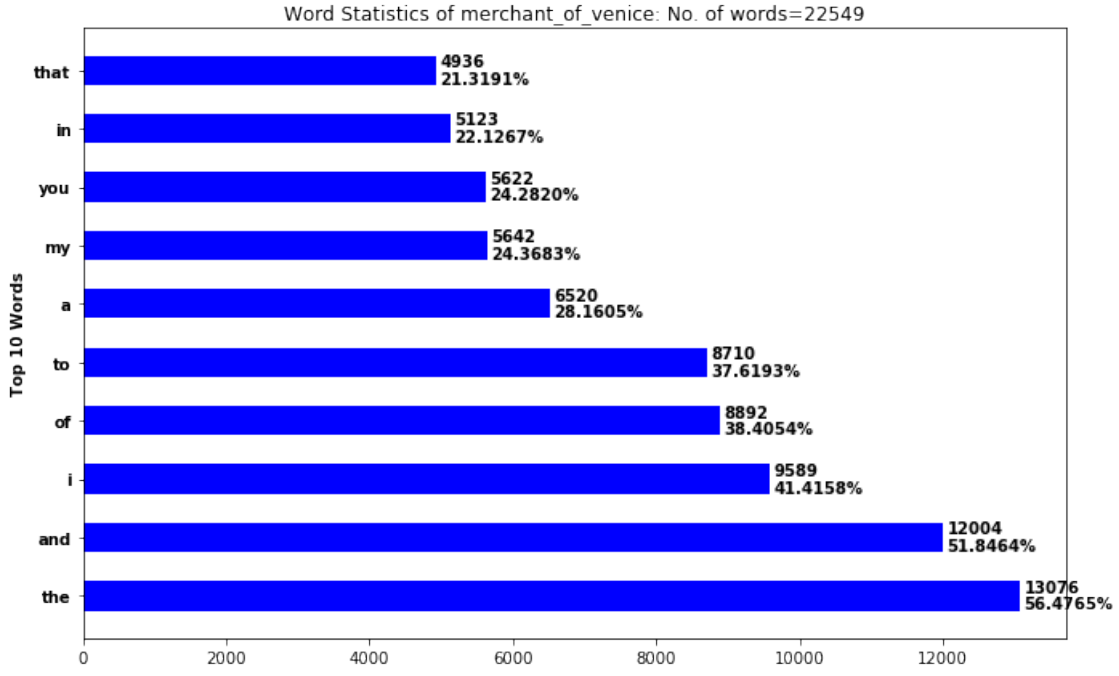



```
------------------------------15------------------------------
[('the', 12244), ('and', 11394), ('i', 8884), ('of', 8439), ('to', 8243), ('a', 6071), ('my', 5
No. of words in measure_for_measure is 23576
No. of different words in measure_for_measure is 3223
```

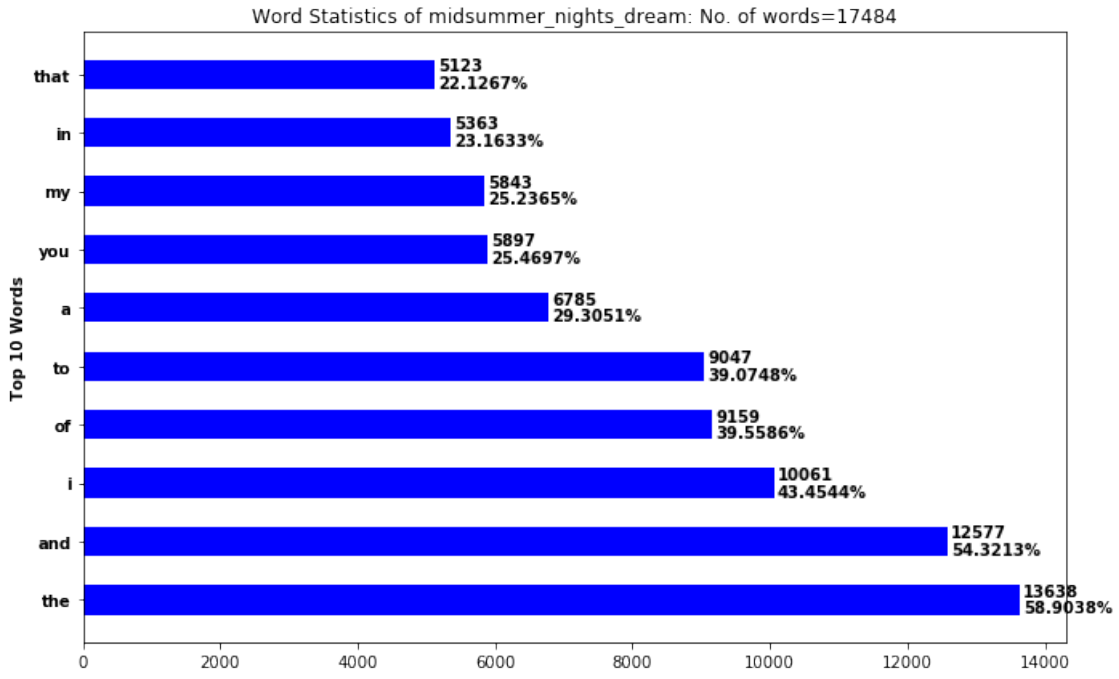


----------16----------

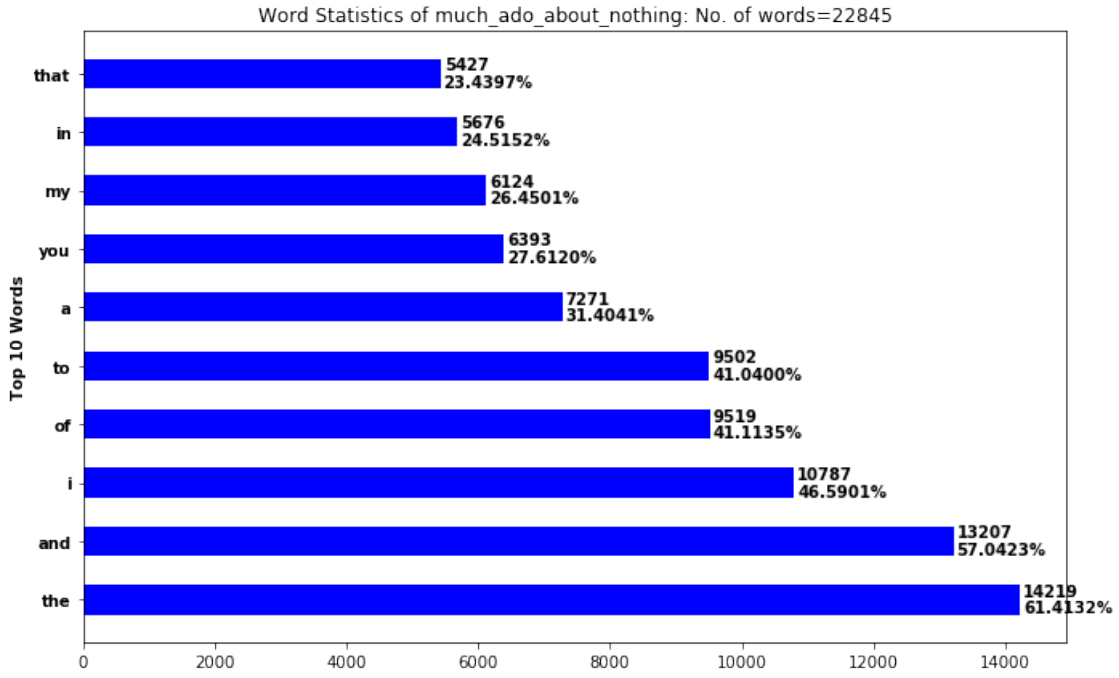
```
[('the', 13076), ('and', 12004), ('i', 9589), ('of', 8892), ('to', 8710), ('a', 6520), ('my', 5
No. of words in merchant_of_venice is 22549
No. of different words in merchant_of_venice is 3158
```



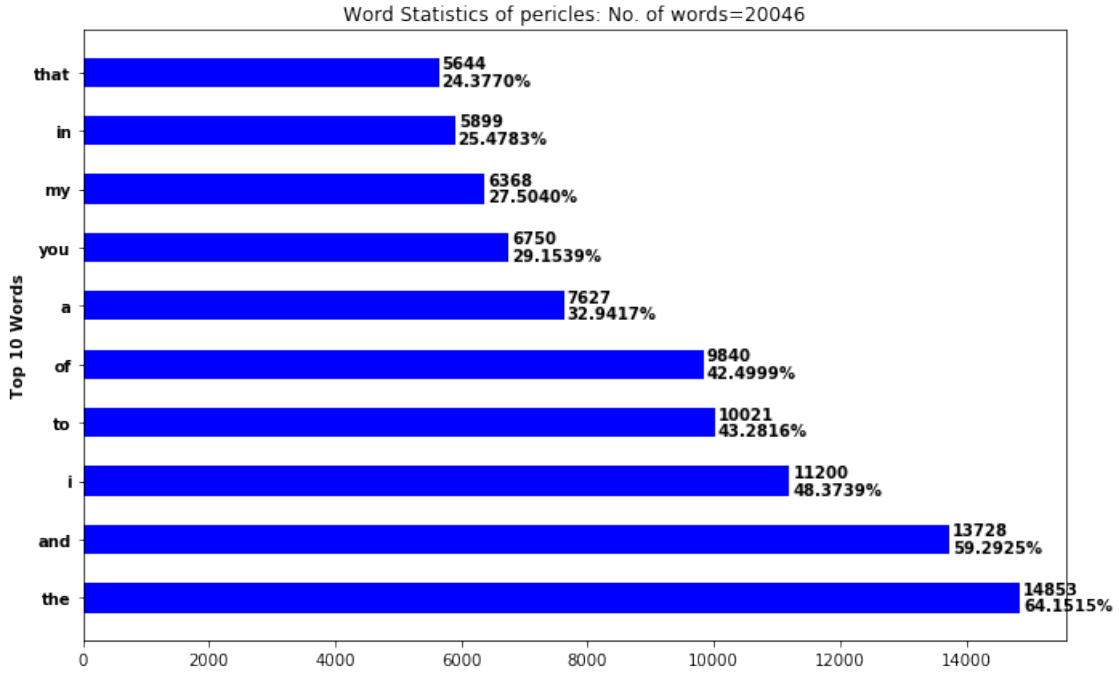
```
------------------------------17------------------------------
[('the', 13638), ('and', 12577), ('i', 10061), ('of', 9159), ('to', 9047), ('a', 6785), ('you'
No. of words in midsummer_nights_dream is 17484
No. of different words in midsummer_nights_dream is 2908
```



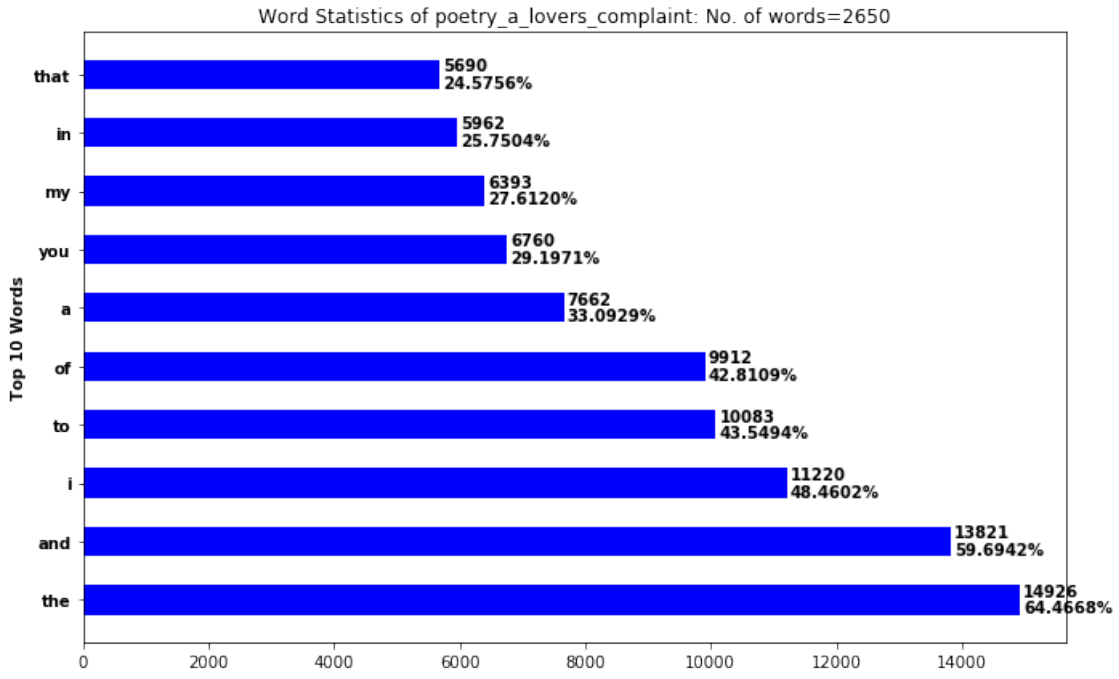

```
------------------------------18------------------------------
[('the', 14219), ('and', 13207), ('i', 10787), ('of', 9519), ('to', 9502), ('a', 7271), ('you'
No. of words in much_ado_about_nothing is 22845
No. of different words in much_ado_about_nothing is 2899
```



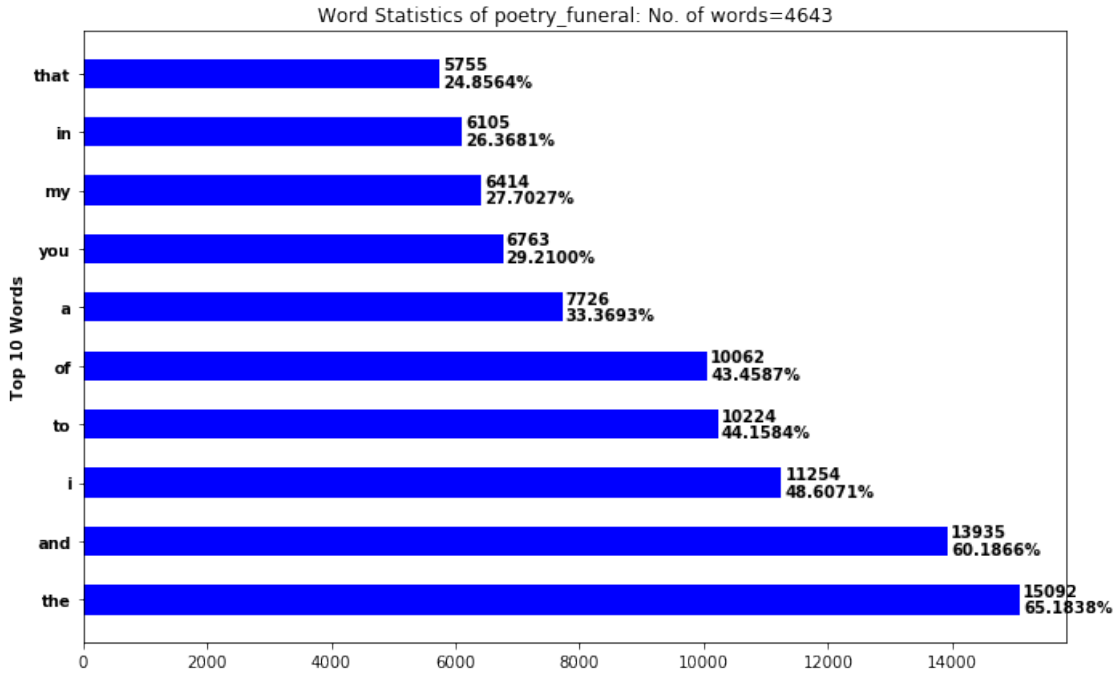

```
------------------------------19------------------------------
[('the', 14853), ('and', 13728), ('i', 11200), ('to', 10021), ('of', 9840), ('a', 7627), ('you
No. of words in pericles is 20046
No. of different words in pericles is 3141
```



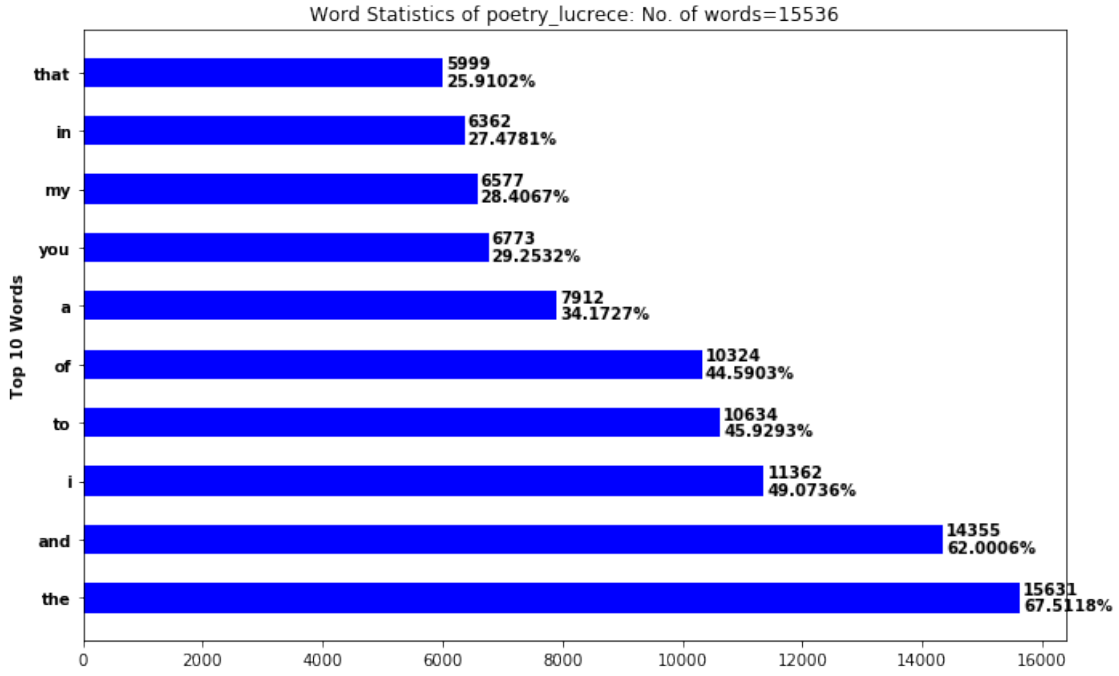
```
------------------------------20------------------------------
[('the', 14926), ('and', 13821), ('i', 11220), ('to', 10083), ('of', 9912), ('a', 7662), ('you
No. of words in poetry_a_lovers_complaint is 2650
No. of different words in poetry_a_lovers_complaint is 1077
```



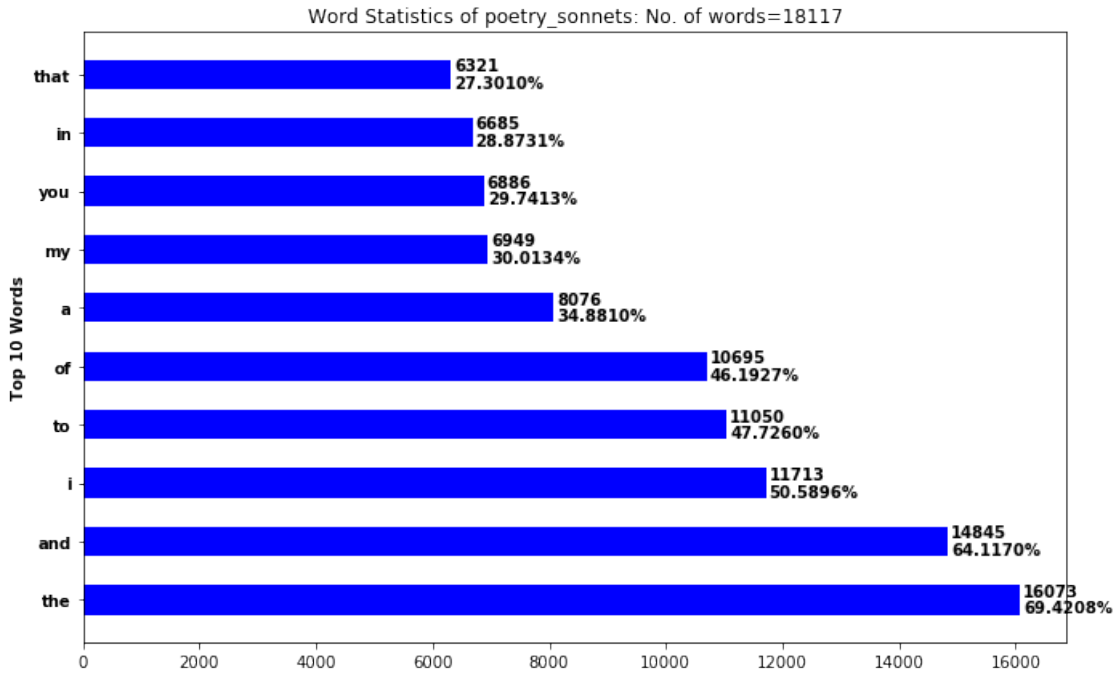

```
------------------------------21------------------------------
[('the', 15092), ('and', 13935), ('i', 11254), ('to', 10224), ('of', 10062), ('a', 7726), ('you
No. of words in poetry_funeral is 4643
No. of different words in poetry_funeral is 1554
```



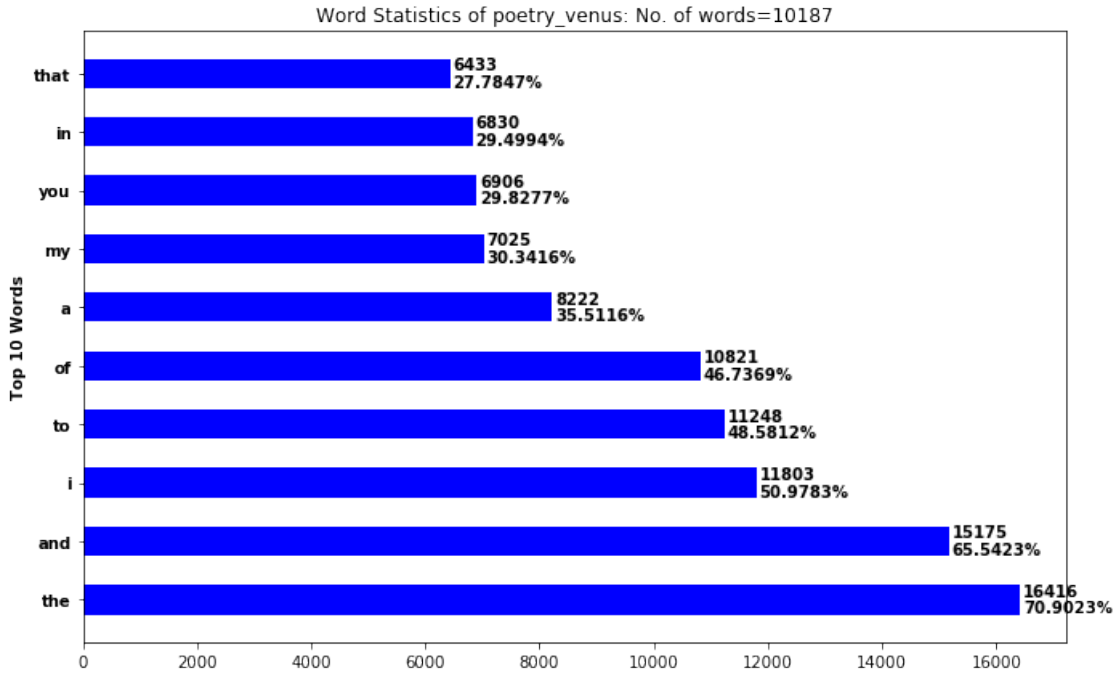

```
------------------------------22------------------------------
[('the', 15631), ('and', 14355), ('i', 11362), ('to', 10634), ('of', 10324), ('a', 7912), ('you
No. of words in poetry_lucrece is 15536
No. of different words in poetry_lucrece is 3403
```



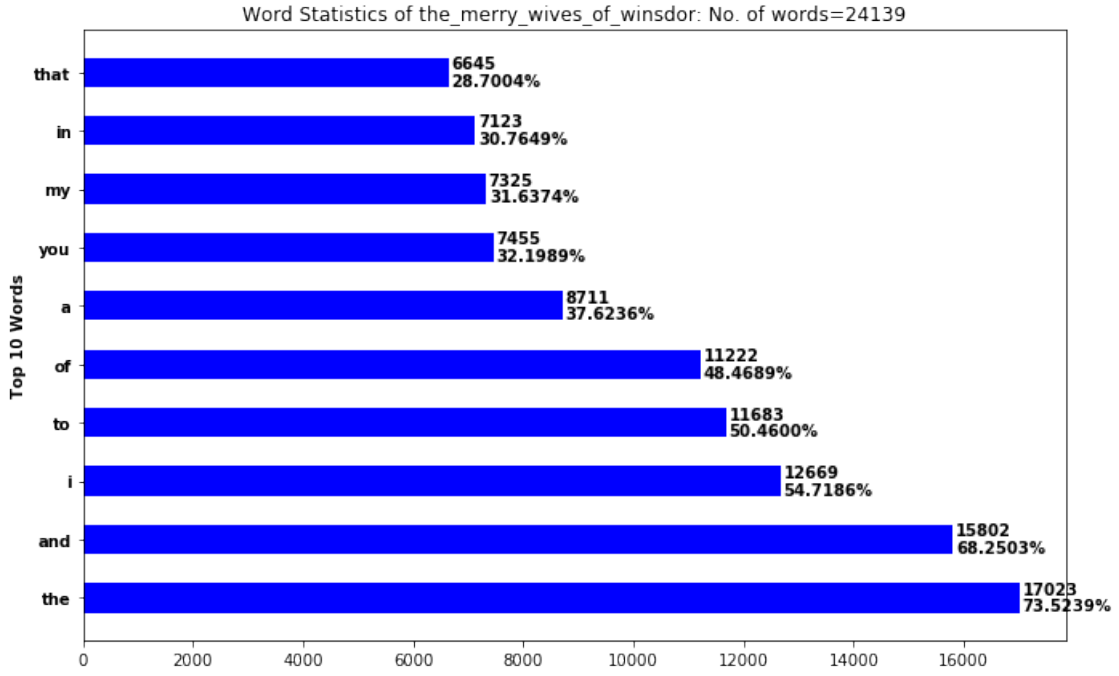
```
------------------------------23------------------------------
[('the', 16073), ('and', 14845), ('i', 11713), ('to', 11050), ('of', 10695), ('a', 8076), ('my
No. of words in poetry_sonnets is 18117
No. of different words in poetry_sonnets is 3251
```



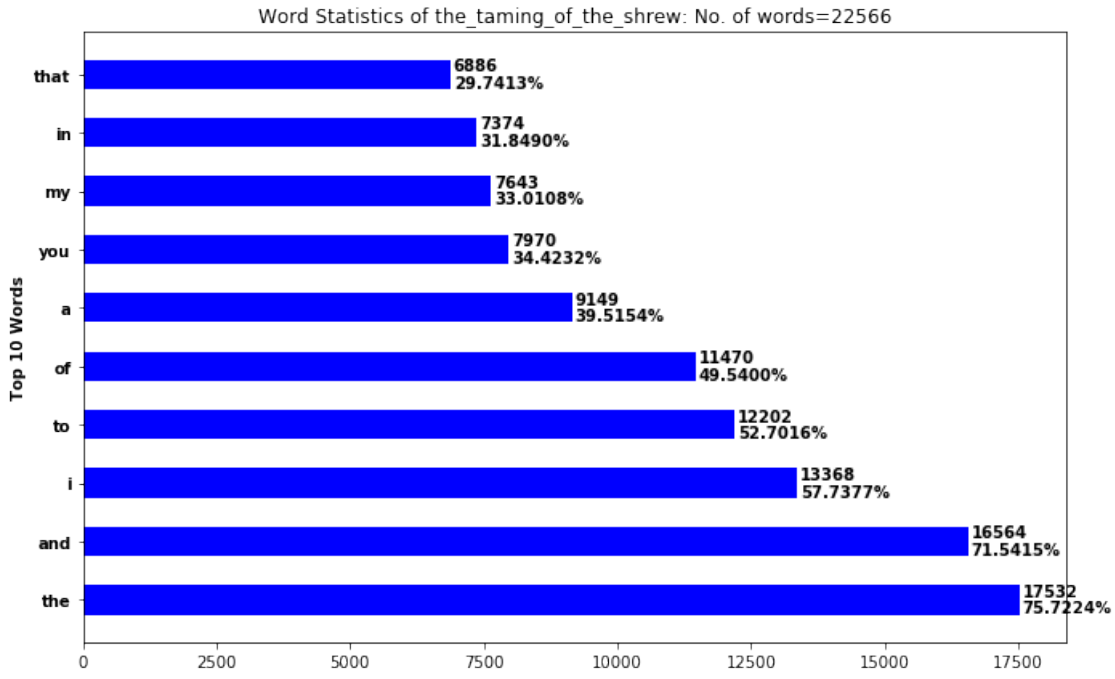

```
------------------------------24------------------------------
[('the', 16416), ('and', 15175), ('i', 11803), ('to', 11248), ('of', 10821), ('a', 8222), ('my
No. of words in poetry_venus is 10187
No. of different words in poetry_venus is 2534
```



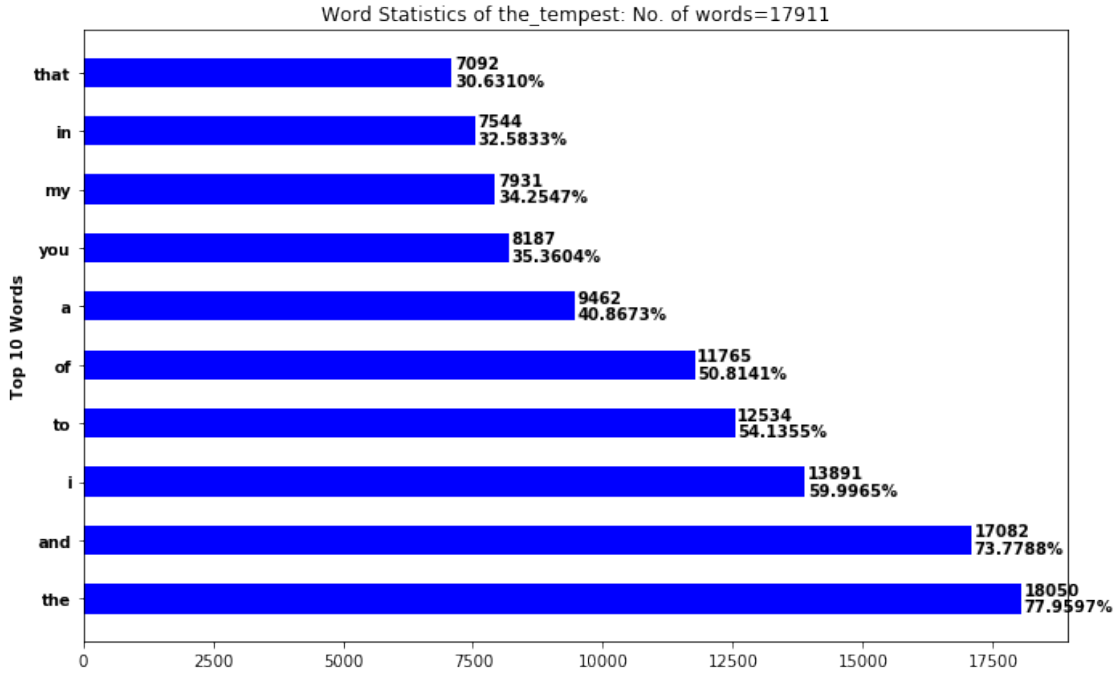

```
------------------------------25------------------------------
[('the', 17023), ('and', 15802), ('i', 12669), ('to', 11683), ('of', 11222), ('a', 8711), ('you
No. of words in the_merry_wives_of_winsdor is 24139
No. of different words in the_merry_wives_of_winsdor is 3177
```



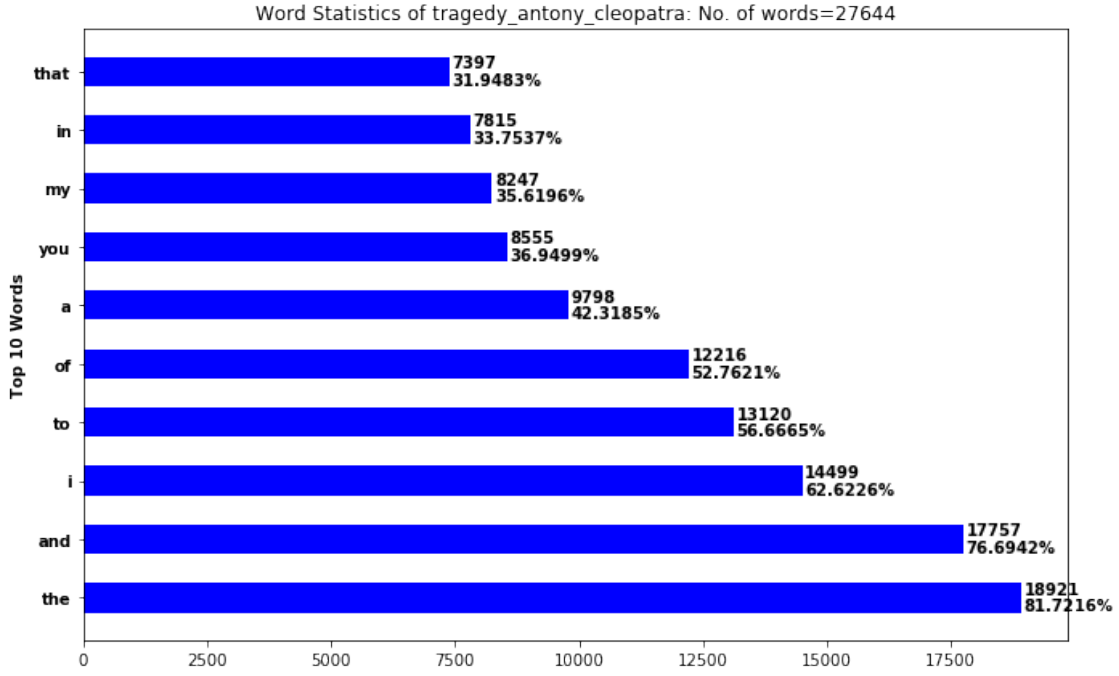
```
------------------------------26------------------------------
[('the', 17532), ('and', 16564), ('i', 13368), ('to', 12202), ('of', 11470), ('a', 9149), ('you
No. of words in the_taming_of_the_shrew is 22566
No. of different words in the_taming_of_the_shrew is 3144
```



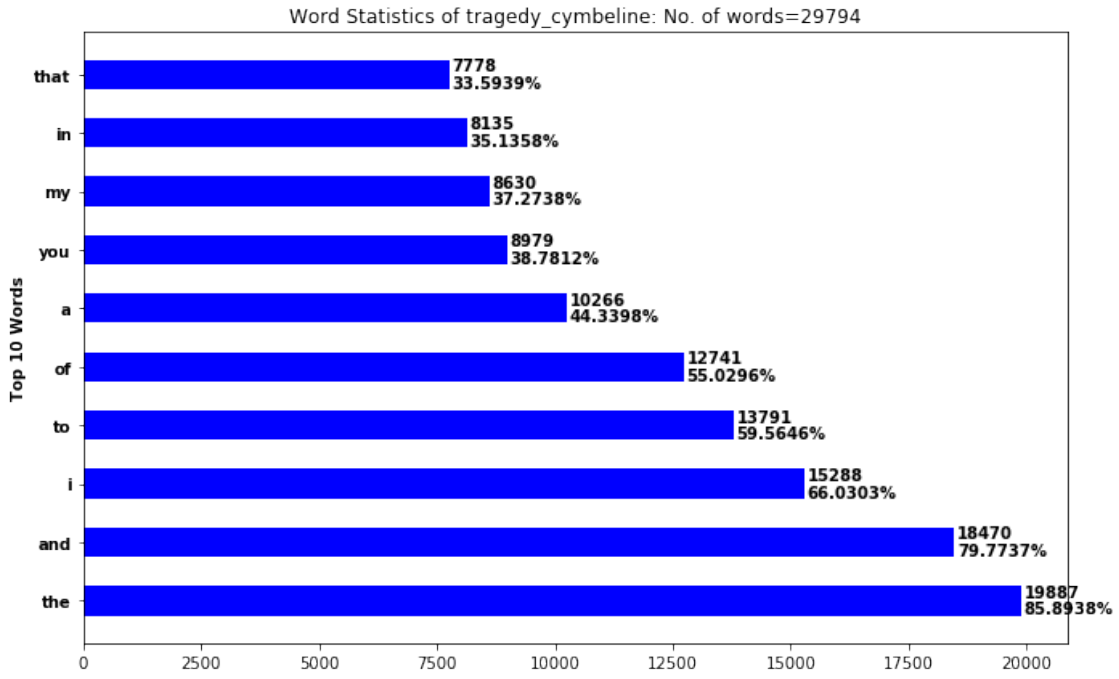

```
------------------------------27------------------------------
[('the', 18050), ('and', 17082), ('i', 13891), ('to', 12534), ('of', 11765), ('a', 9462), ('you
No. of words in the_tempest is 17911
No. of different words in the_tempest is 3076
```



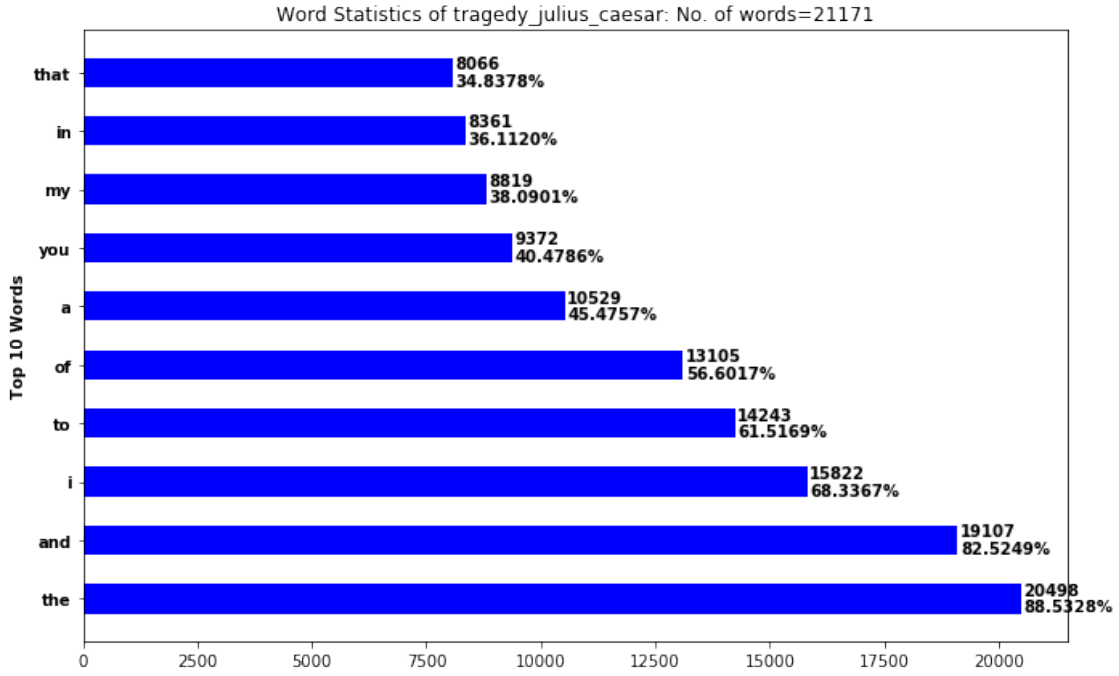

```
------------------------------28------------------------------
[('the', 18921), ('and', 17757), ('i', 14499), ('to', 13120), ('of', 12216), ('a', 9798), ('you
No. of words in tragedy_antony_cleopatra is 27644
No. of different words in tragedy_antony_cleopatra is 3773
```



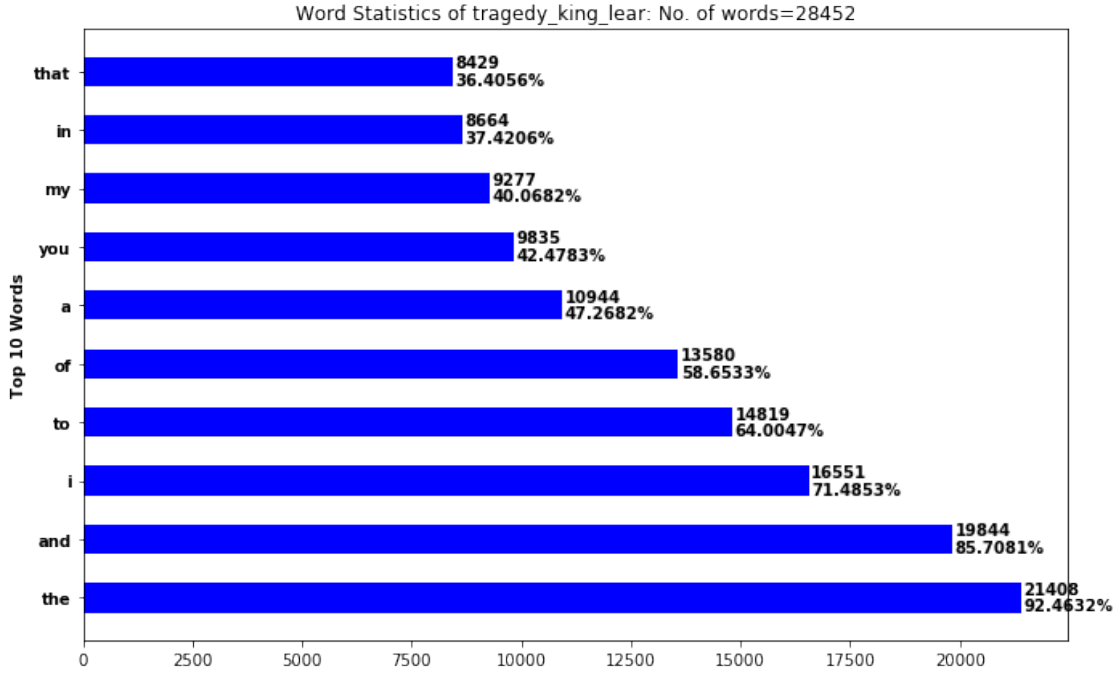
```
------------------------------29------------------------------
[('the', 19887), ('and', 18470), ('i', 15288), ('to', 13791), ('of', 12741), ('a', 10266), ('y
No. of words in tragedy_cymbeline is 29794
No. of different words in tragedy_cymbeline is 4053
```



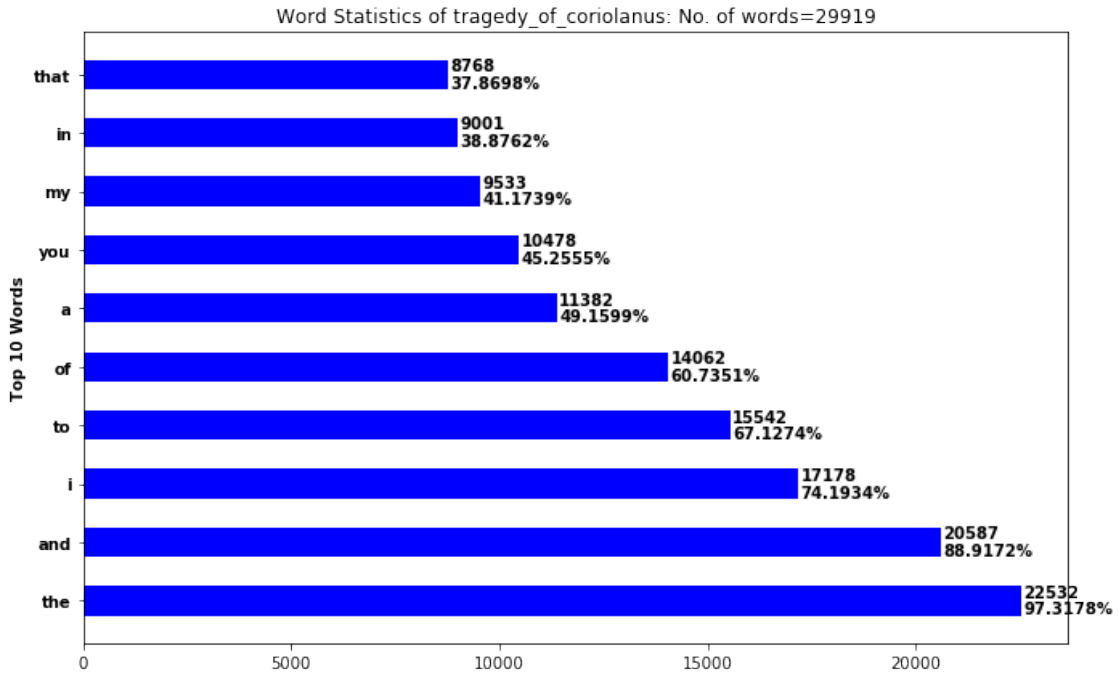

```
------------------------------30------------------------------
[('the', 20498), ('and', 19107), ('i', 15822), ('to', 14243), ('of', 13105), ('a', 10529), ('y
No. of words in tragedy_julius_caesar is 21171
No. of different words in tragedy_julius_caesar is 2787
```



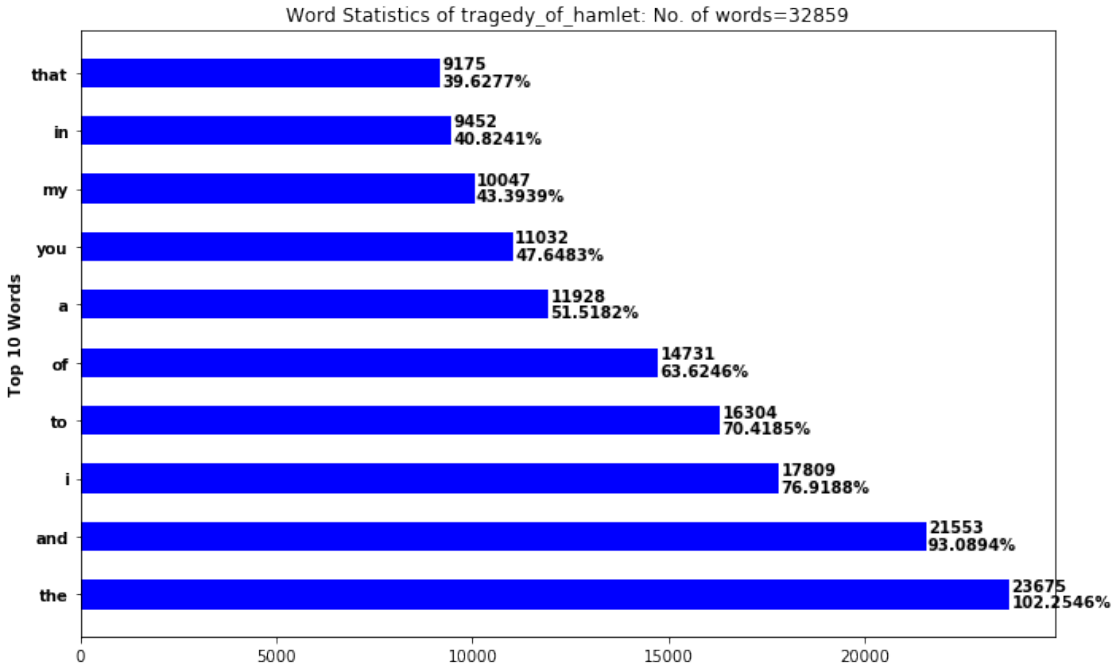

```
------------------------------31------------------------------
[('the', 21408), ('and', 19844), ('i', 16551), ('to', 14819), ('of', 13580), ('a', 10944), ('y
No. of words in tragedy_king_lear is 28452
No. of different words in tragedy_king_lear is 3998
```



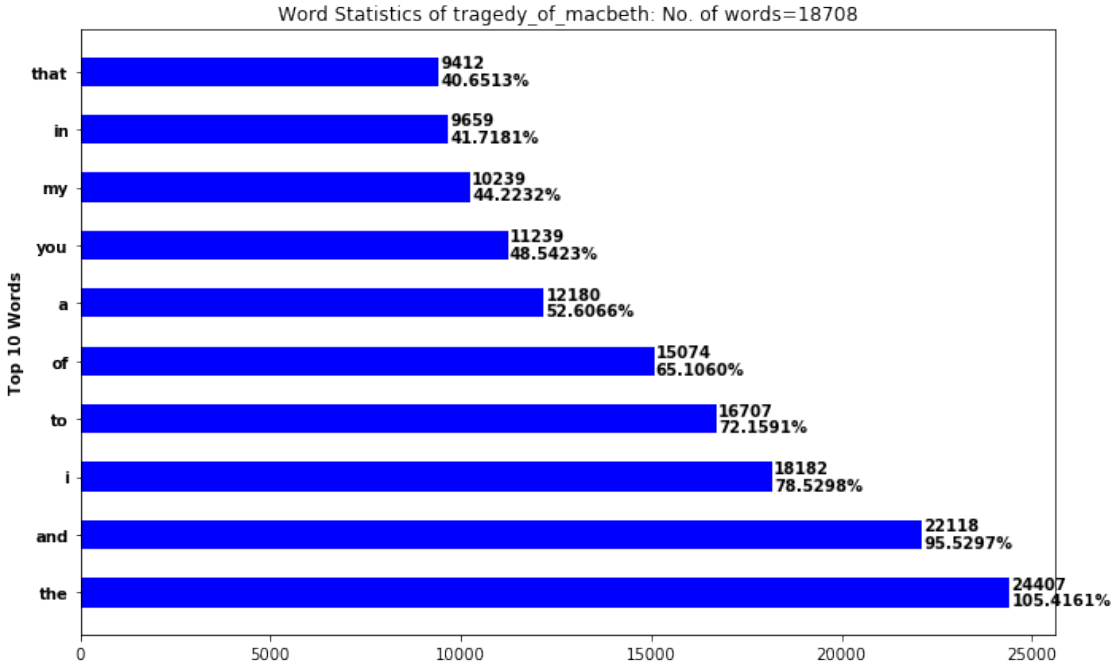
```
------------------------------32------------------------------
[('the', 22532), ('and', 20587), ('i', 17178), ('to', 15542), ('of', 14062), ('a', 11382), ('y
No. of words in tragedy_of_coriolanus is 29919
No. of different words in tragedy_of_coriolanus is 3872
```



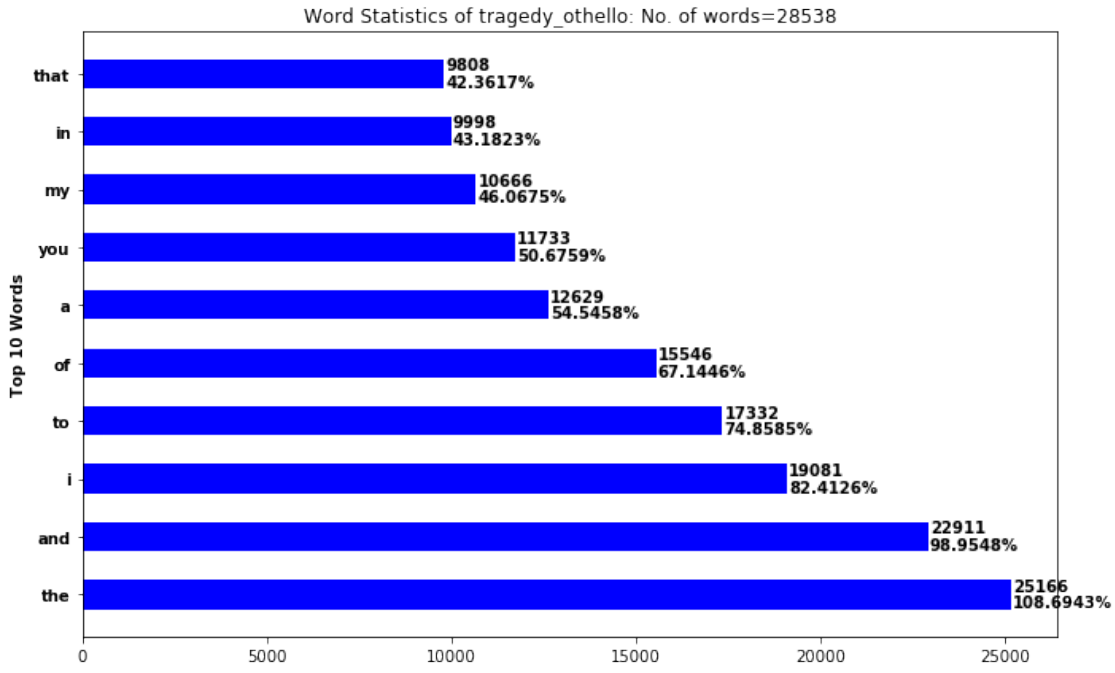

```
------------------------------33------------------------------
[('the', 23675), ('and', 21553), ('i', 17809), ('to', 16304), ('of', 14731), ('a', 11928), ('y
No. of words in tragedy_of_hamlet is 32859
No. of different words in tragedy_of_hamlet is 4541
```



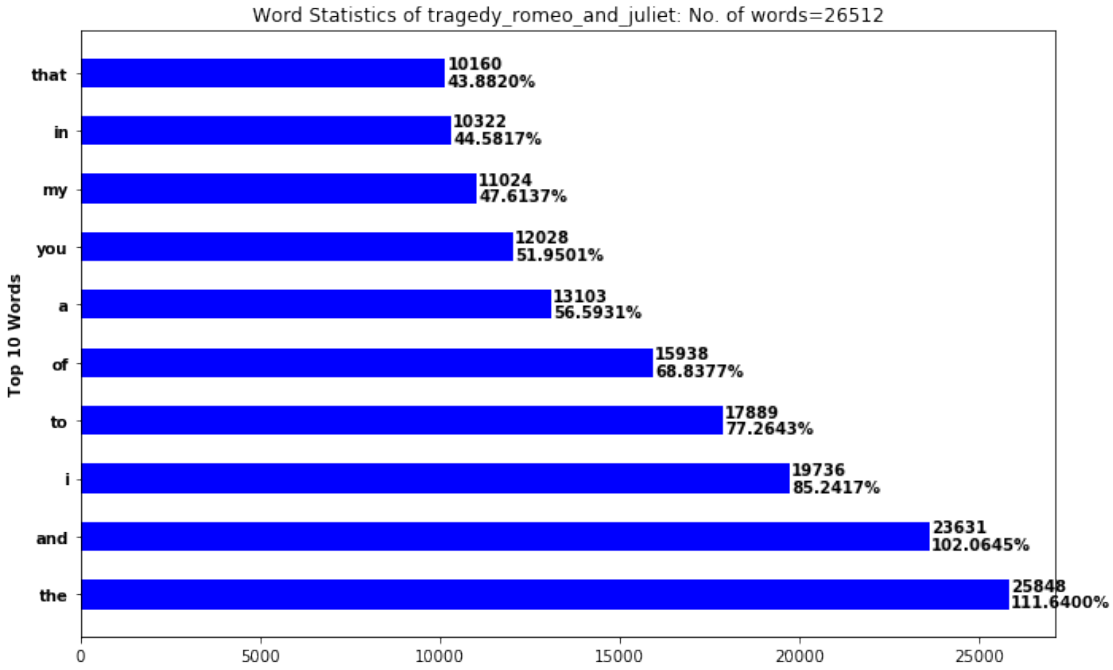

```
--------------------------------34--------------------------------
[('the', 24407), ('and', 22118), ('i', 18182), ('to', 16707), ('of', 15074), ('a', 12180), ('y
No. of words in tragedy_of_macbeth is 18708
No. of different words in tragedy_of_macbeth is 3201
```



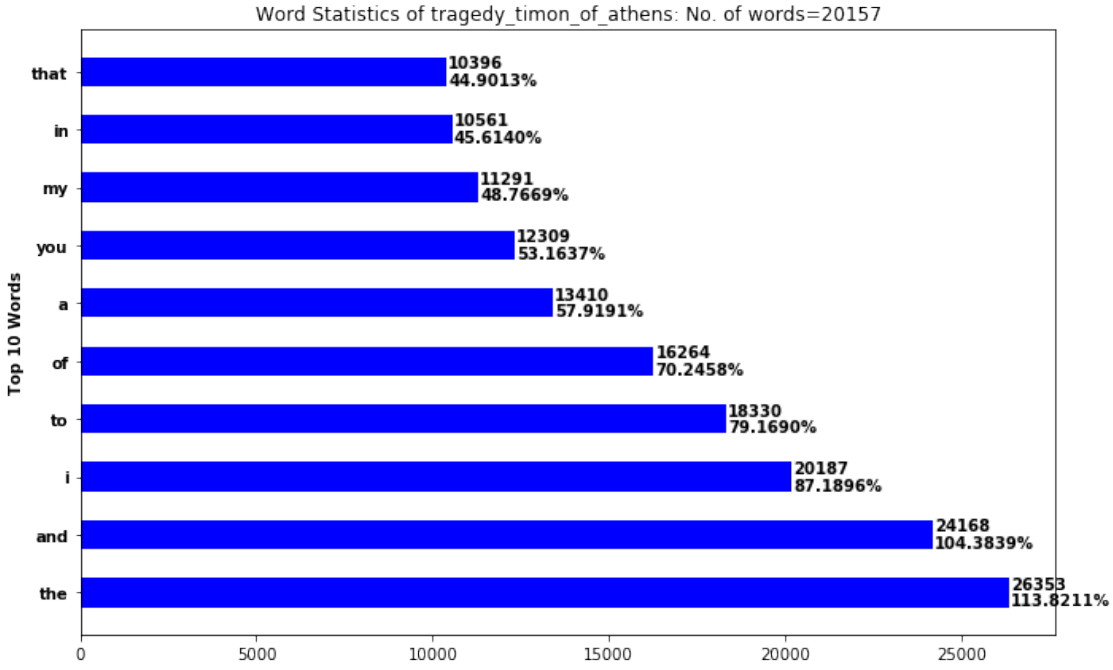
```
------------------------------35------------------------------
[('the', 25166), ('and', 22911), ('i', 19081), ('to', 17332), ('of', 15546), ('a', 12629), ('y
No. of words in tragedy_othello is 28538
No. of different words in tragedy_othello is 3650
```



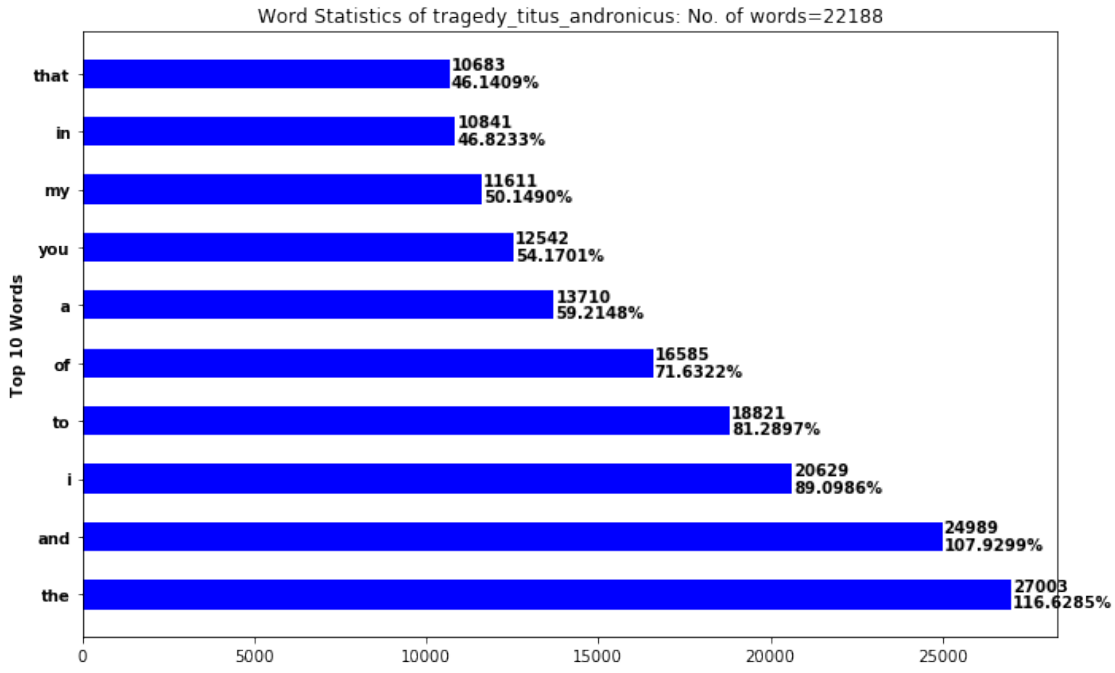

```
------------------------------36------------------------------
[('the', 25848), ('and', 23631), ('i', 19736), ('to', 17889), ('of', 15938), ('a', 13103), ('y
No. of words in tragedy_romeo_and_juliet is 26512
No. of different words in tragedy_romeo_and_juliet is 3541
```



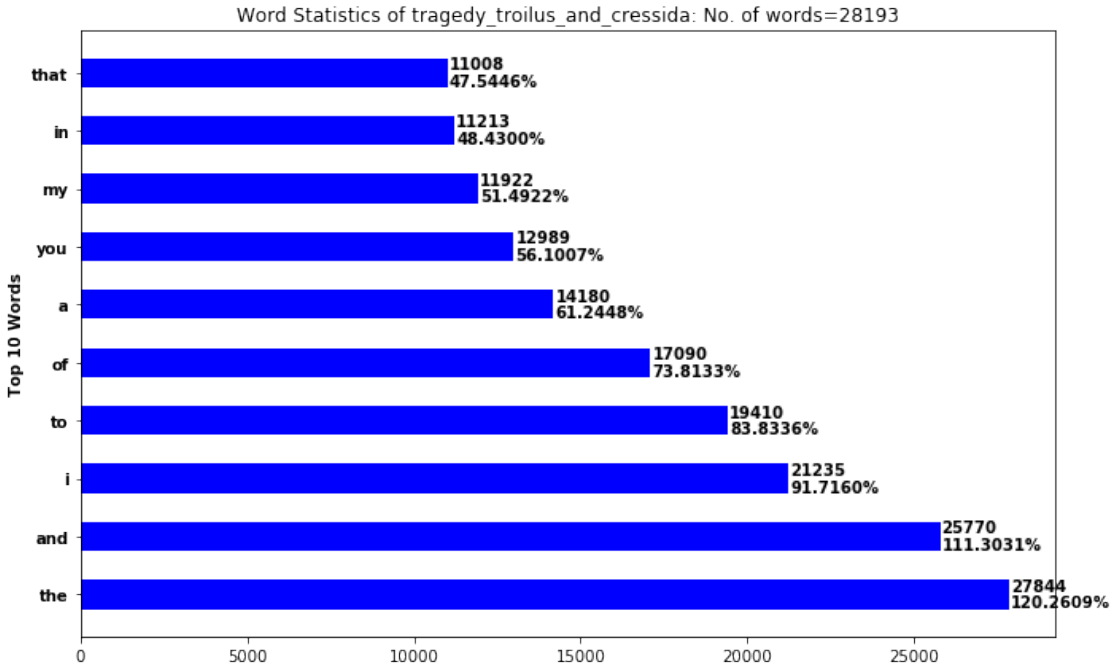

```
--------------------------------37--------------------------------
[('the', 26353), ('and', 24168), ('i', 20187), ('to', 18330), ('of', 16264), ('a', 13410), ('y
No. of words in tragedy_timon_of_athens is 20157
No. of different words in tragedy_timon_of_athens is 3182
```



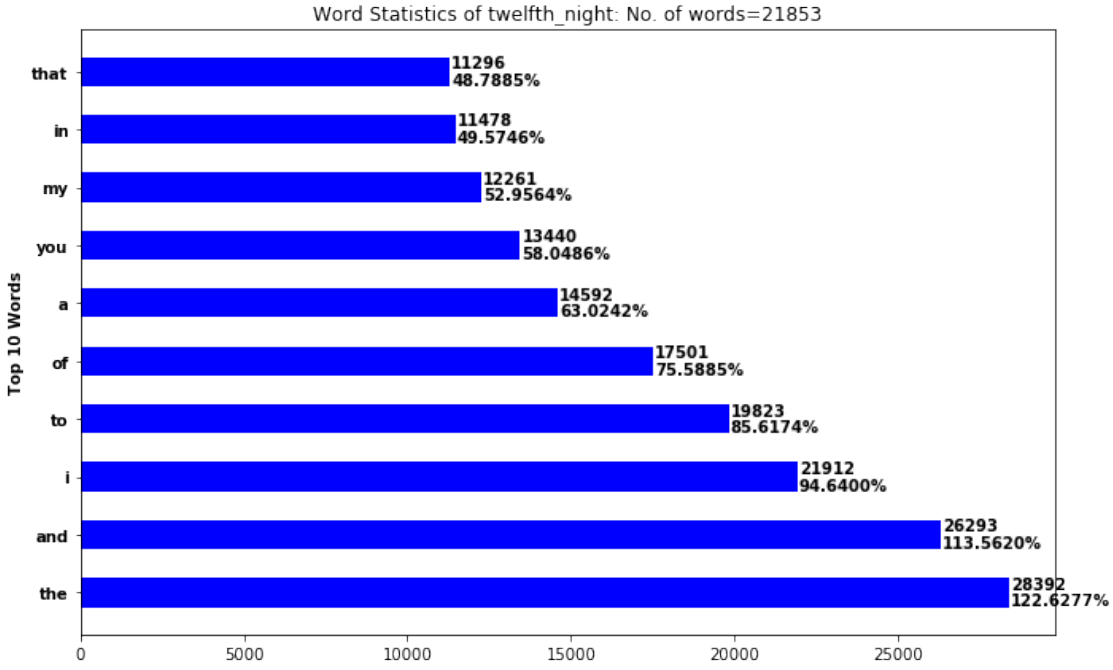
```
------------------------------38------------------------------
[('the', 27003), ('and', 24989), ('i', 20629), ('to', 18821), ('of', 16585), ('a', 13710), ('y
No. of words in tragedy_titus_andronicus is 22188
No. of different words in tragedy_titus_andronicus is 3298
```



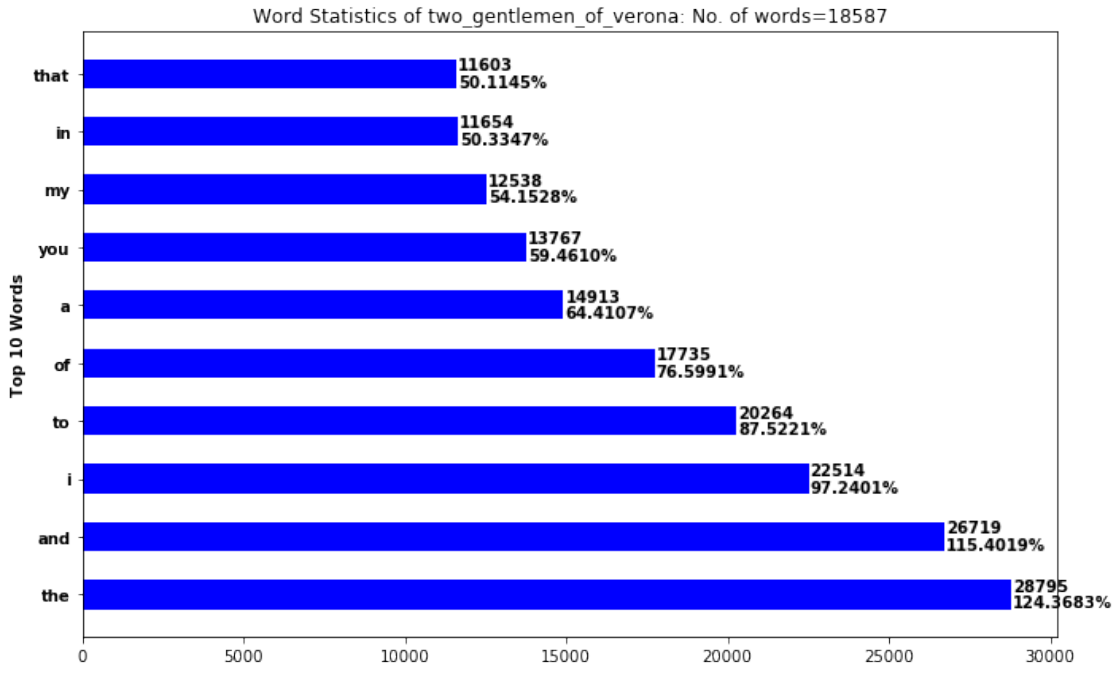

```
------------------------------39------------------------------
[('the', 27844), ('and', 25770), ('i', 21235), ('to', 19410), ('of', 17090), ('a', 14180), ('y
No. of words in tragedy_troilus_and_cressida is 28193
No. of different words in tragedy_troilus_and_cressida is 4114
```



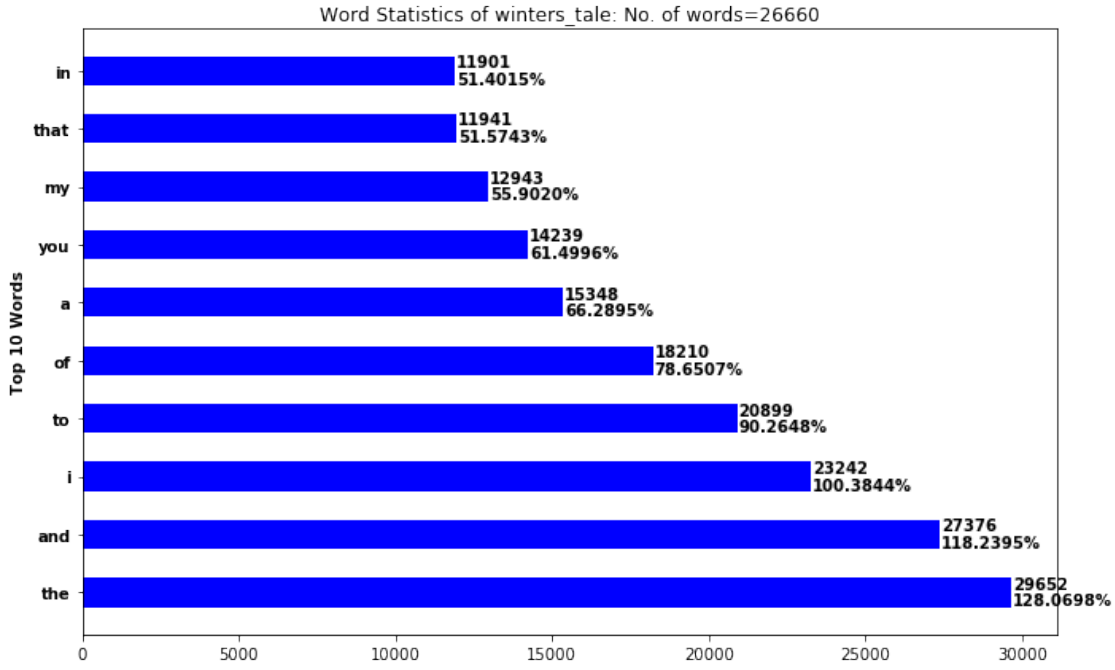

```
--------------------------------40--------------------------------
[('the', 28392), ('and', 26293), ('i', 21912), ('to', 19823), ('of', 17501), ('a', 14592), ('y
No. of words in twelfth_night is 21853
No. of different words in twelfth_night is 3022
```



```
------------------------------41------------------------------
[('the', 28795), ('and', 26719), ('i', 22514), ('to', 20264), ('of', 17735), ('a', 14913), ('y
No. of words in two_gentlemen_of_verona is 18587
No. of different words in two_gentlemen_of_verona is 2638
```




```
------------------------------42------------------------------
[('the', 29652), ('and', 27376), ('i', 23242), ('to', 20899), ('of', 18210), ('a', 15348), ('y
No. of words in winters_tale is 26660
No. of different words in winters_tale is 3712
```




```
The total number of words Shakespeare used in the complete work:  956852

The total number of different words Shakespeare used in the complete work:  23927
```

In [ ]: