Loading data

In [22]:

```
def split_in_sets(data):
    essay_sets = []
    min_scores = []
    max_scores = []
    for s in range(1,9):
        essay_set = data[data["essay_set"] == s]
        essay_set.dropna(axis=1, inplace=True)
        n, d = essay_set.shape
        set_scores = essay_set["domain1_score"]
        print ("Set", s, ": Essays = ", n , "\t Attributes = ", d)
        min_scores.append(set_scores.min())
        max_scores.append(set_scores.max())
        essay_sets.append(essay_set)
    return (essay_sets, min_scores, max_scores)
```

In [23]:
```
dataset_path = "./asap-aes/training_set_rel3.tsv"

import os
import pandas as pd

data = pd.read_csv(dataset_path, sep="\t", encoding="ISO-8859-1")
essay_sets, min_scores, max_scores = split_in_sets(data)
set1, set2, set3, set4, set5, set6, set7, set8 = tuple(essay_sets)
data.dropna(axis=1, inplace=True)

data.drop(columns=["rater1_domain1", "rater2_domain1"], inplace=Tru
data.head()
```

```
Set 1 : Essays =  1783   Attributes =  6
Set 2 : Essays =  1800   Attributes =  9
Set 3 : Essays =  1726   Attributes =  6
Set 4 : Essays =  1770   Attributes =  6
Set 5 : Essays =  1805   Attributes =  6
Set 6 : Essays =  1800   Attributes =  6
Set 7 : Essays =  1569   Attributes =  14
Set 8 : Essays =  723    Attributes =  18
```

Out[23]:

	essay_id	essay_set	essay	domain1_score
0	1	1	Dear local newspaper, I think effects computer...	8
1	2	1	Dear @CAPS1 @CAPS2, I believe that using compu...	9
2	3	1	Dear, @CAPS1 @CAPS2 @CAPS3 More and more peopl...	7
3	4	1	Dear Local Newspaper, @CAPS1 I have found that...	10
4	5	1	Dear @LOCATION1, I know having computers has a...	8

In [24]:
```
print("Minimum Scores: ", min_scores)
print("Maximum Scores: ", max_scores)
```

```
Minimum Scores:  [2, 1, 0, 0, 0, 0, 2, 10]
Maximum Scores:  [12, 6, 3, 3, 4, 4, 24, 60]
```

In [25]:
```
essay_id_key = "essay_id"
essay_set_key = "essay_set"
essay_key = "essay"
domain1_score_key = "domain1"
```

In [55]:

```
#Feature keys
char_count_key = "char_count"
word_count_key = "word_count"
diff_words_key = "diff_words"
diff_words_count_key = "diff_words_count"
word_count_root_key = "word_count_root"
sen_count_key = "sen_count"
avg_word_len_key = "avg_word_len"
avg_sen_len_key = "avg_sen_len"
l5_word_count_key = "l5_word_count"
l6_word_count_key = "l6_word_count"
l7_word_count_key = "l7_word_count"
l8_word_count_key = "l8_word_count"
```

In [56]:

```
import numpy as np
import nltk
import re
from nltk.corpus import stopwords

def sentence_to_word_list(sentence, remove_stopwords):
    # Remove non letter from sentenece and stop words
    sen_char_count = 0
    sen_word_count = 0
    l5_sen_word_count = 0
    l6_sen_word_count = 0
    l7_sen_word_count = 0
    l8_sen_word_count = 0
    sen_diff_words = set()


    sentence = re.sub("[^a-zA-Z]", " ", sentence)
    stops = set(stopwords.words("english"))
    all_words = sentence.lower().split()
    kept_words = []

    for word in all_words:
        sen_char_count += len(word)
        sen_word_count += 1
        word_len = len(word)
        if word_len > 5:
            l5_sen_word_count += 1
        if word_len > 6:
            l6_sen_word_count += 1
        if word_len > 7:
            l7_sen_word_count += 1
        if word_len > 8:
            l8_sen_word_count += 1

        sen_diff_words.add(word)

        if remove_stopwords and word not in stops:
            kept_words.append(word)
        else:
```

```
        else:
            kept_words.append(word)

    features = {
         char_count_key: sen_char_count,
         word_count_key: sen_word_count,
         l5_word_count_key: l5_sen_word_count,
         l6_word_count_key: l6_sen_word_count,
         l7_word_count_key: l7_sen_word_count,
         l8_word_count_key: l8_sen_word_count,
         diff_words_key: sen_diff_words
    }

    return (kept_words, features)

def essay_to_sentences(essay, remove_stopwords = False):
    # Convert essay into sentence
    tokenizer = nltk.data.load('tokenizers/punkt/english.pickle')
    sentences = tokenizer.tokenize(essay.strip())
    split_sentences = []

    char_count = 0
    word_count = 0
    diff_words = set()
    word_count_root = 0
    sen_count = 0
    avg_word_len = 0
    avg_sen_len = 0
    l5_word_count = 0
    l6_word_count = 0
    l7_word_count = 0
    l8_word_count = 0

    for sentence in sentences:
        if len(sentence) > 0:

            kept_words, features = sentence_to_word_list(sentence,
            split_sentences.append(kept_words)

            sen_count +=1
            char_count += features[char_count_key]
            word_count += features[word_count_key]
            l5_word_count += features[l5_word_count_key]
            l6_word_count += features[l6_word_count_key]
            l7_word_count += features[l7_word_count_key]
            l8_word_count += features[l8_word_count_key]
            diff_words = diff_words|features[diff_words_key]

    word_count_root = word_count ** (1/4)
    avg_word_len = char_count / word_count
    avg_sen_len = word_count / sen_count

    features = {
        char_count_key: char_count,
        word_count_key: word_count,
        word_count_key: word_count,
        diff_words_count_key: len(diff_words),
        word_count_root_key: word_count_root,
        sen_count_key: sen_count,
        avg_word_len_key: avg_word_len,
        avg_sen_len_key: avg_sen_len,
        l5_word_count_key: l5_word_count,
        l6_word_count_key: l6_word_count,
        l7_word_count_key: l7_word_count,
        l8_word_count_key: l8_word_count

    }

    return (split_sentences, features)
```

In [63]:

```
import pprint
pp = pprint.PrettyPrinter(indent=4)

#Featrues
first_essay = data.iloc[0][essay_key]
print(first_essay)
split_sentences, features = essay_to_sentences(first_essay)
# print(split_sentences)
print("\n\nFeatures: ")
pp.pprint(features)
```

```
Dear local newspaper, I think effects computers have on people are
great learning skills/affects because they give us time to chat wi
th friends/new people, helps us learn about the globe(astronomy) a
nd keeps us out of troble! Thing about! Dont you think so? How wou
ld you feel if your teenager is always on the phone with friends!
Do you ever time to chat with your friends or buisness partner abo
ut things. Well now - there's a new way to chat the computer, thei
rs plenty of sites on the internet to do so: @ORGANIZATION1, @ORGA
NIZATION2, @CAPS1, facebook, myspace ect. Just think now while you
r setting up meeting with your boss on the computer, your teenager
is having fun on the phone not rushing to get off cause you want t
o use it. How did you learn about other countrys/states outside of
yours? Well I have by computer/internet, it's a new way to learn a
bout what going on in our time! You might think your child spends
a lot of time on the computer, but ask them so question about the
economy, sea floor spreading or even about the @DATE1's you'll be
surprise at how much he/she knows. Believe it or not the computer
is much interesting then in class all day reading out of books. If
your child is home on your computer or at a local library, it's be
tter than being out with friends being fresh, or being perpressure
d to doing something they know isnt right. You might not know wher
e your child is, @CAPS2 forbidde in a hospital bed because of a dr
ive-by. Rather than your child on the computer learning, chatting
or just playing games, safe and sound in your home or community pl
or just playing games, safe and sound in your home or community pl
ace. Now I hope you have reached a point to understand and agree w
ith me, because computers can have great effects on you or child b
ecause it gives us time to chat with friends/new people, helps us
learn about the globe and believe or not keeps us out of troble. T
hank you for listening.


Features:
{   'avg_sen_len': 21.875,
    'avg_word_len': 4.222857142857142,
    'char_count': 1478,

    'diff_words_count': 164,
    'l5_word_count': 74,
    'l6_word_count': 59,
    'l7_word_count': 34,
    'l8_word_count': 13,
    'sen_count': 16,
    'word_count': 350,
    'word_count_root': 4.3253077270721105}
```

In [ ]:

In [ ]: