PROCESS NOTES

WILL BE DONE

DONE

- data is devided into two parts as train(%80) and test(%20). Test data will stay as untouched.

The set is data.csv under data folder. Each sentence is splitted into many phrases with individual SentenceId and PhraseId's by using Standord parser and related sentiment score from 0 to 5 is assigned under Sentiment column (0 is bad 5 is good).

In [19]:

```
import numpy as np
import pandas as pd
df=pd.read_csv('data/data.csv',sep="\t")
df.head()
```

Out[19]:

	PhraseId	SentenceId	Phrase	Sentiment
0	1	1	A series of escapades demonstrating the adage ...	1
1	2	1	A series of escapades demonstrating the adage ...	2
2	3	1	A series	2
3	4	1	A	2
4	5	1	series	2

Get total number of phrases and sentences.

In [28]:

```
print("number of total phrases=",len(df.index))

#create mask for sentences
msk=[]
tmp=0

for x in df.SentenceId:
    if x !=tmp:
        msk.append(True)
        tmp=x
    else:
        msk.append(False)

mskSentence=pd.Series(msk)
print("number of total sentences=",sum(mskSentence))
print("the set of just sentences as follows:")
df[mskSentence].head()
```

```
number of total phrases= 156060
number of total sentences= 8529
the set of just sentences as follows:
```

Out[28]:

	PhraseId	SentenceId	Phrase	Sentiment
0	1	1	A series of escapades demonstrating the adage ...	1
63	64	2	This quiet , introspective and entertaining in...	4
81	82	3	Even fans of Ismail Merchant 's work , I suspe...	1
116	117	4	A positively thrilling combination of ethnogra...	3
156	157	5	Aggressive self-glorification and a manipulati...	1

Devide data into two diffent set as train(%80 of data) and test(%20 of data). Save them under data folder as train.csv and test.csv respectively.

In [29]:

```
#create files using random mask
mskDevide=np.random.rand(len(df))<0.8
train = df[mskDevide]
test = df[~mskDevide]

#reset indexing
train=train.reset_index(drop=True)
test=train.reset_index(drop=True)

#save files
train.to_csv("data/train.csv", sep='\t')
test.to_csv("data/test.csv", sep='\t')

print("train size=",sum(mskDevide))
print("test size=",len(df.index)-sum(mskDevide))
print("train set is seen as follows:")
train.head()
```

```
train size= 125049
test size= 31011
train set is seen as follows
```

Out[29]:

	PhraseId	SentenceId	Phrase	Sentiment
0	1	1	A series of escapades demonstrating the adage ...	1
1	2	1	A series of escapades demonstrating the adage ...	2
2	4	1	A	2
3	5	1	series	2
4	6	1	of escapades demonstrating the adage that what...	2

Get basic statistics about train and test set.

In [37]:

```
print("train set:")
print(train.Sentiment.describe())
print()
print("test set:")
print(test.Sentiment.describe())
```

```
train set:
count    125049.000000
mean          2.062480
std           0.893042
min           0.000000
25%           2.000000
50%           2.000000
75%           3.000000
max           4.000000
Name: Sentiment, dtype: float64

test set:
count    125049.000000
mean          2.062480
std           0.893042
min           0.000000
25%           2.000000
50%           2.000000
75%           3.000000
max           4.000000
Name: Sentiment, dtype: float64
```

In [38]:

```
train.head()
```

Out[38]:

	PhraseId	SentenceId	Phrase	Sentiment
0	1	1	A series of escapades demonstrating the adage ...	1
1	2	1	A series of escapades demonstrating the adage ...	2
2	4	1	A	2
3	5	1	series	2
4	6	1	of escapades demonstrating the adage that what...	2