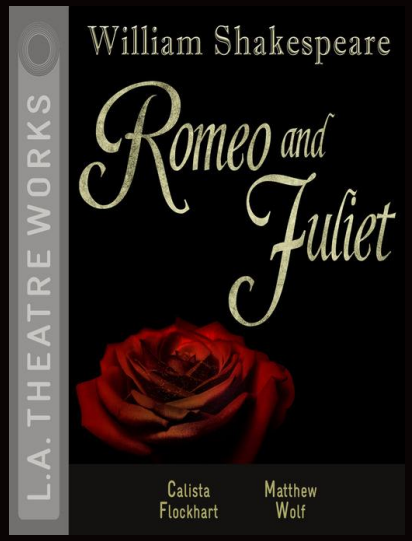
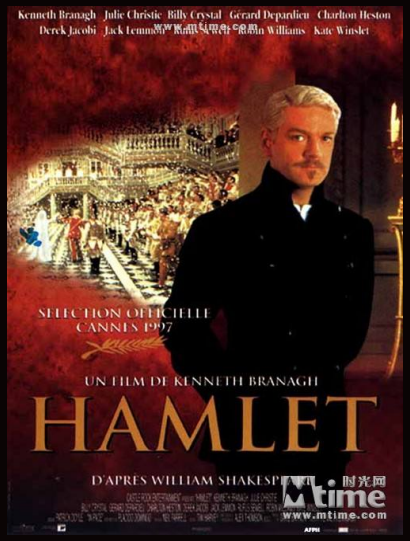
词汇密度

莎士比亚与曹雪芹文学素养对比

史志茹
产业经济学
经济管理学院
2018年3月23日


L.A. THEATRE WORKS
William Shakespeare
Romeo and Juliet
Calista Flockhart
Matthew Wolf


Kenneth Branagh Julie Christie Billy Crystal Gérard Depardieu Charlton Heston
Derek Jacobi Jack Lemmon Kate Winslet
SELECTION OFFICIELLE
CANNES 1997
UN FILM DE KENNETH BRANAGH
HAMLET
mtime
www.mtime.com


紅樓夢

莎士比亚全集

红楼梦

莎士比亚简介

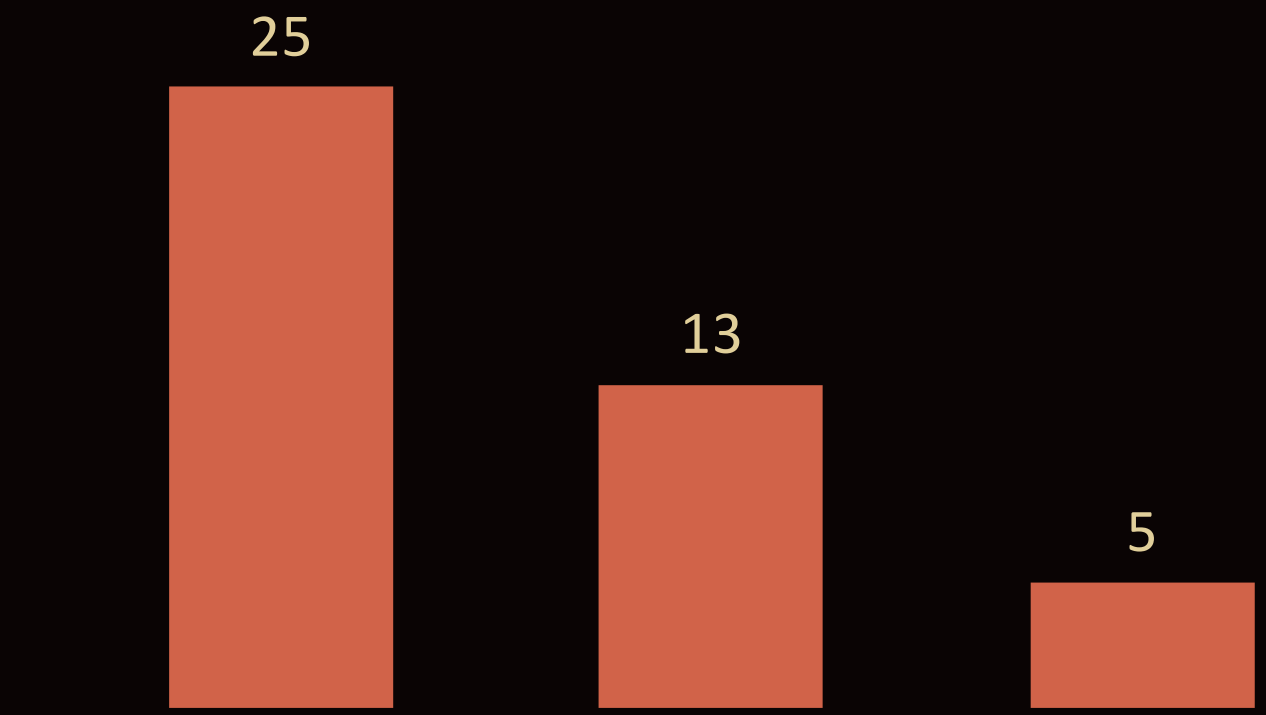
莎士比亚不仅是一位举世闻名的文学大师，更是一位出类拔萃的语言大师；就个人而言，他对英语语言的影响和贡献无人可比。

——周海中先生《莎士比亚文学作品的语言特色》

莎士比亚流传下来的作品包括39部戏剧、154首十四行诗、两首长叙事诗。他的戏剧有各种主要语言的译本，且表演次数远远超过其他所有戏剧家的作品。

代表作品 《罗密欧与朱丽叶》《哈姆雷特》《李尔王》《麦克白》《奥赛罗》《威尼斯商人》《驯悍记》等

莎士比亚三个时期作品数



莎士比亚全集TTR计算

```
$cat SHAKESPEARE/*>SHAKESPEAREALL.txt
$cat SHAKESPEAREALL.txt  |tr -cs "[:alnum:]"  "\n"|tr  [:upper:]
[:lower:]>SHAKESPEAREALL.pure.txt
$wc SHAKESPEAREALL.pure.txt
$wc SHAKESPEAREALL.pure.txt
$sort SHAKESPEAREALL.pure.txt|uniq|wc

$ wc SHAKESPEAREALL.pure.txt
  960368  960367 4862682 SHAKESPEAREALL.pure.txt
$ sort SHAKESPEAREALL.pure.txt|uniq|wc
   24426   24425  196938
```

TTR=24425/960367=2.54%

中文分词Python代码

```
9
10 import jieba
11 import os
12 import  codecs
13
14 dir_name = r'C:\Users\Administrator\Desktop'
15 file_name = 'HLM1.txt'
16
17 file1 = open(os.path.join(dir_name, file_name), 'r')
18 line_list = []
19 jieba_list = []
20 for line in file1:
21     line_list.append(line.decode('utf-8'))
22     jieba_list.append([temp for temp in list(jieba.cut(line_list[-1], cut_all = False)) if temp not in
23                        [u', ', u'。', u'、', u'; ', u': ', u'! ', u'?', u'……', u'—']])
24 file1.close()
25
26 fobj = codecs.open('jiebafenci_HLM.txt', 'w', 'utf-8')
27 kk = ([[(fobj.write(temp), fobj.write('\n')) for temp in item] for item in jieba_list])
28 fobj.close()
```

莎士比亚三个时期词语密度对比代码

```
$cat zaoqi1590-1600/*>zaoqi1590-1600ALL.txt
$cat zhongqi1601-1607/*>zhongqi1601-1607ALL.txt
$cat wanqi1608-1612/*>wanqi1608-1612ALL.txt
$cat zaoqi1590-1600ALL.txt  |tr -cs "[:alnum:]"  "\n"|tr  [:upper:]
[:lower:]>zaoqi1590-1600ALL.pure.txt
$cat zhongqi1601-1607ALL.txt  |tr -cs "[:alnum:]"  "\n"|tr
[:upper:]  [:lower:]>zhongqi1601-1607ALL.pure.txt
$cat wanqi1608-1612ALL.txt |tr -cs "[:alnum:]" "\n"|tr [:upper:]
[:lower:]>wanqi1608-1612ALL.pure.txt
$sort zaoqi1590-1600ALL.pure.txt|uniq|wc
$sort zhongqi1601-1607ALL.pure.txt|uniq|wc
$sort wanqi1608-1612ALL.pure.txt|uniq|wc
$wc zaoqi1590-1600ALL.pure.txt
$wc zhongqi1601-1607ALL.pure.txt
$wc wanqi1608-1612ALL.pure.txt
```

莎士比亚三个时期的TTR

```
$ wc zaoqi1590-1600ALL.pure.txt
  534054  534053 2708633 zaoqi1590-
1600ALL.pure.txt
$ wc zhongqi1601-1607ALL.pure.txt
  324076  324075 1638090 zhongqi1601-
1607ALL.pure.txt
$ wc wanqi1608-1612ALL.pure.txt
  102240  102239  515961 wanqi1608-
1612ALL.pure.txt
$ sort zaoqi1590-1600ALL.pure.txt|uniq|wc
   18451   18450  146022
$ sort zhongqi1601-1607ALL.pure.txt|uniq|wc
   15224   15223  119380
$ sort wanqi1608-1612ALL.pure.txt|uniq|wc
    8037    8036   60274
```

TTR(zaoqi)= 18450 /534053=3.45%

TTR(zhongqi)= 15223 / 324075 =4.70%

TTR(zaoqi)= 8036 / 102239 =7.86%

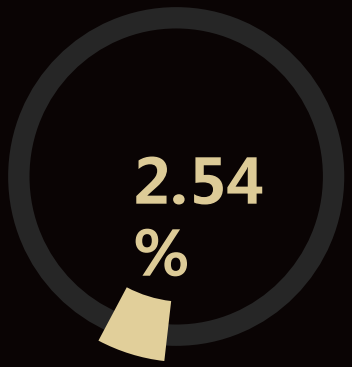
红楼梦TTR计算

```
$wc jiebafenci_HLM.txt
$sort jiebafenci_HLM.txt  |uniq|wc
$wc jiebafenci_HLM.txt
```

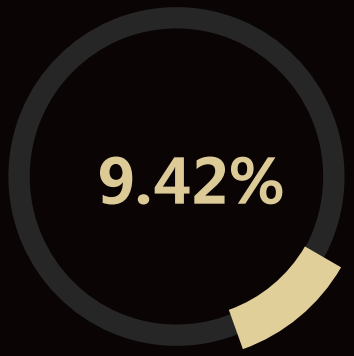
```
$ wc jiebafenci_HLM.txt
  504213  509594 2759616 jiebafenci_HLM.txt
$   sort jiebafenci_HLM.txt  |uniq|wc
   43335   48021  330929
```

TTR=48021/509594=9.42%

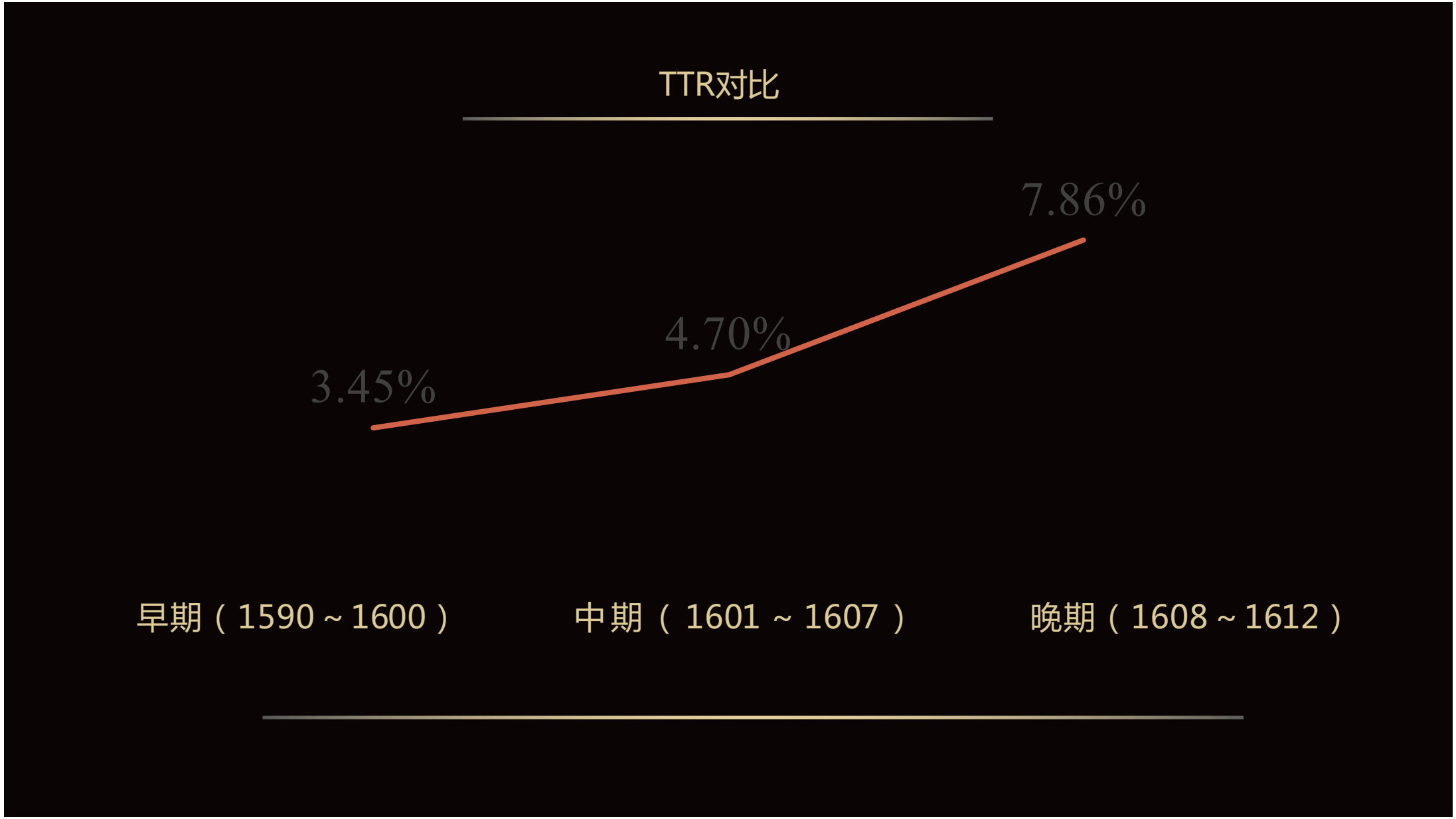
TTR对比



莎士比亚全集



红楼梦


TTR对比
7.86%
4.70%
3.45%
早期（1590～1600）
中期（1601～1607）
晚期（1608～1612）

谢谢观看