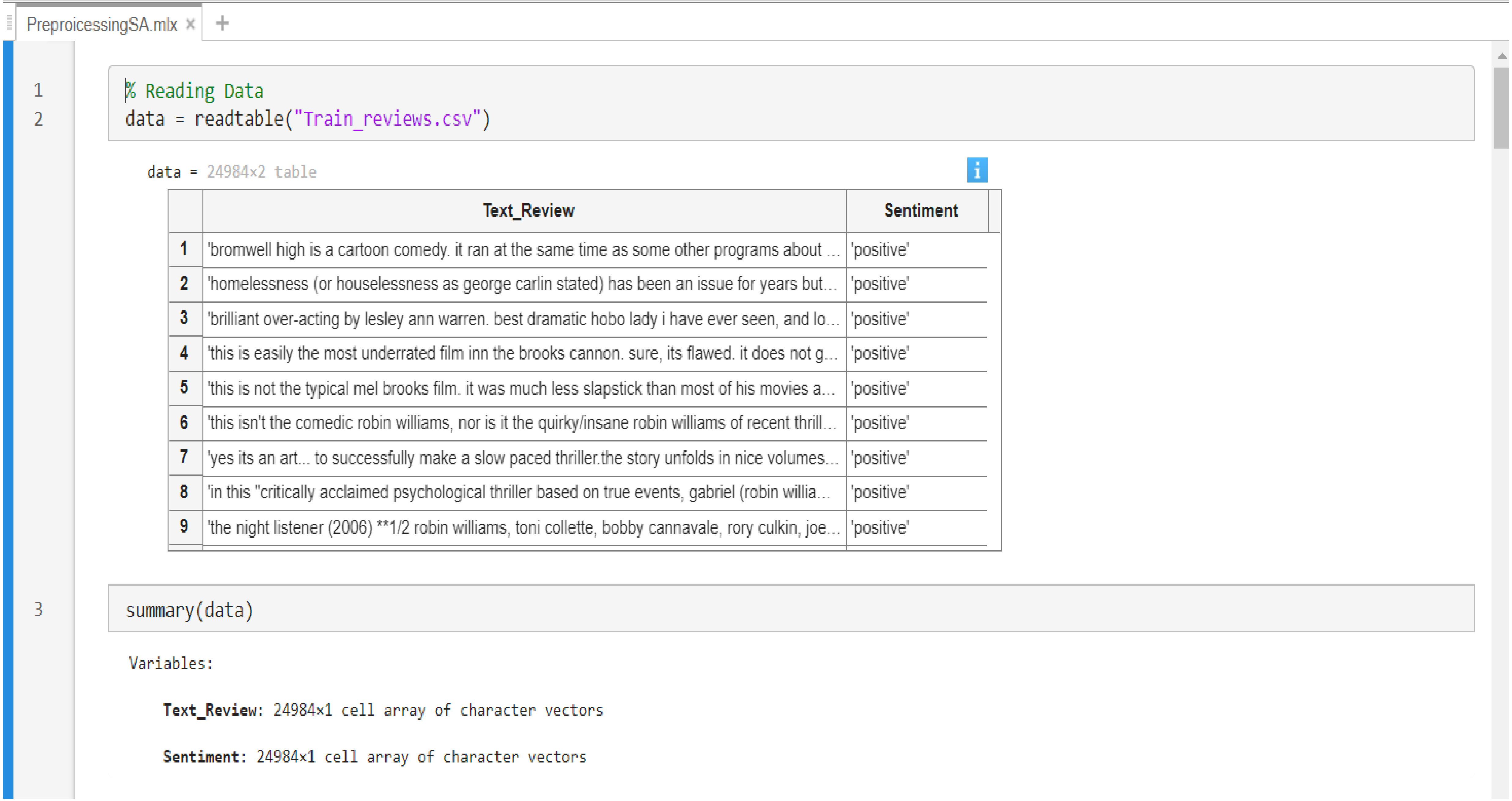


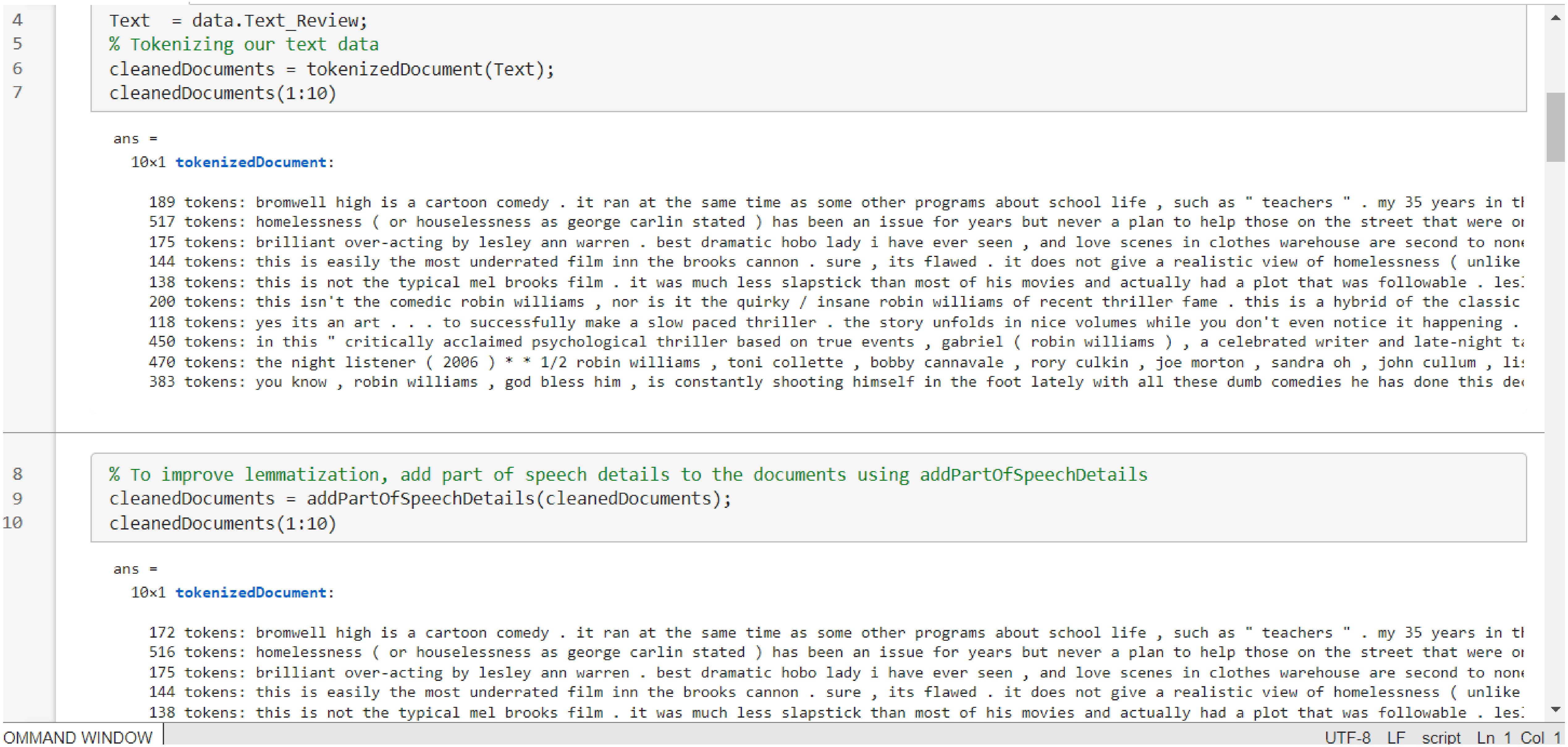
```
% Reading Data
data = readtable("Train_reviews.csv")
```

```
data = 24984×2 table
```

	Text_Review	Sentiment
1	'bromwell high is a cartoon comedy. it ran at the same time as some other programs about ...	'positive'
2	'homelessness (or houselessness as george carlin stated) has been an issue for years but...	'positive'
3	'brilliant over-acting by lesley ann warren. best dramatic hobo lady i have ever seen, and lo...	'positive'
4	'this is easily the most underrated film inn the brooks cannon. sure, its flawed. it does not g...	'positive'
5	'this is not the typical mel brooks film. it was much less slapstick than most of his movies a...	'positive'
6	'this isn't the comedic robin williams, nor is it the quirky/insane robin williams of recent thrill...	'positive'
7	'yes its an art... to successfully make a slow paced thriller.the story unfolds in nice volumes...	'positive'
8	'in this "critically acclaimed psychological thriller based on true events, gabriel (robin willia...	'positive'
9	'the night listener (2006) **1/2 robin williams, toni collette, bobby cannavale, rory culkin, joe...	'positive'

```
summary(data)
```

```
Variables:

    Text_Review: 24984×1 cell array of character vectors

    Sentiment: 24984×1 cell array of character vectors
```

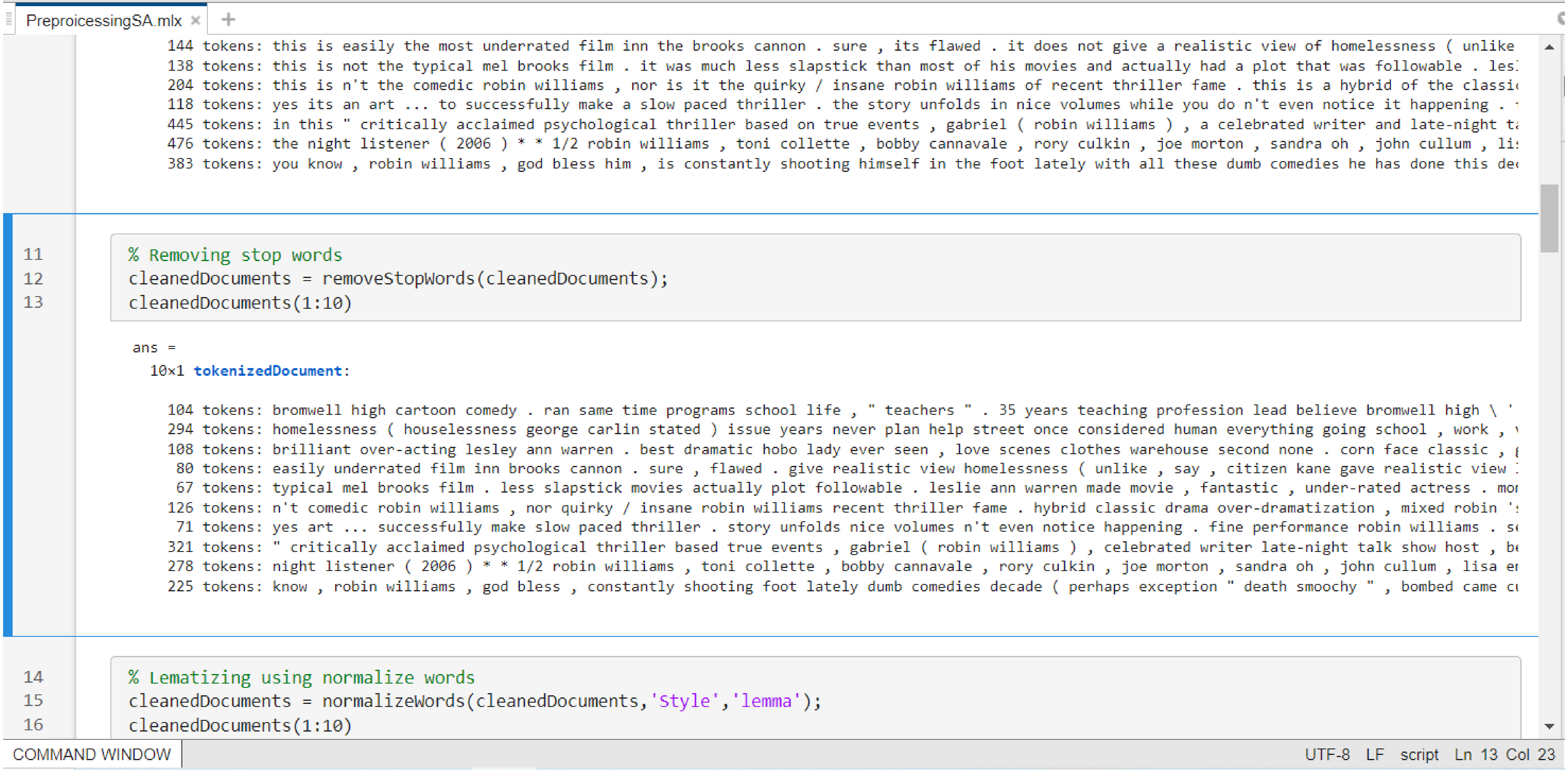


```
Text  = data.Text_Review;
% Tokenizing our text data
cleanedDocuments = tokenizedDocument(Text);
cleanedDocuments(1:10)
```

```
ans =
  10×1 tokenizedDocument:

    189 tokens: bromwell high is a cartoon comedy . it ran at the same time as some other programs about school life , such as " teachers " . my 35 years in th
    517 tokens: homelessness ( or houselessness as george carlin stated ) has been an issue for years but never a plan to help those on the street that were o
    175 tokens: brilliant over-acting by lesley ann warren . best dramatic hobo lady i have ever seen , and love scenes in clothes warehouse are second to none
    144 tokens: this is easily the most underrated film inn the brooks cannon . sure , its flawed . it does not give a realistic view of homelessness ( unlike
    138 tokens: this is not the typical mel brooks film . it was much less slapstick than most of his movies and actually had a plot that was followable . les
    200 tokens: this isn't the comedic robin williams , nor is it the quirky / insane robin williams of recent thriller fame . this is a hybrid of the classic
    118 tokens: yes its an art . . . to successfully make a slow paced thriller . the story unfolds in nice volumes while you don't even notice it happening .
    450 tokens: in this " critically acclaimed psychological thriller based on true events , gabriel ( robin williams ) , a celebrated writer and late-night ta
    470 tokens: the night listener ( 2006 ) * * 1/2 robin williams , toni collette , bobby cannavale , rory culkin , joe morton , sandra oh , john cullum , lis
    383 tokens: you know , robin williams , god bless him , is constantly shooting himself in the foot lately with all these dumb comedies he has done this de
```

```
% To improve lemmatization, add part of speech details to the documents using addPartOfSpeechDetails
cleanedDocuments = addPartOfSpeechDetails(cleanedDocuments);
cleanedDocuments(1:10)
```

```
ans =
  10×1 tokenizedDocument:

    172 tokens: bromwell high is a cartoon comedy . it ran at the same time as some other programs about school life , such as " teachers " . my 35 years in th
    516 tokens: homelessness ( or houselessness as george carlin stated ) has been an issue for years but never a plan to help those on the street that were o
    175 tokens: brilliant over-acting by lesley ann warren . best dramatic hobo lady i have ever seen , and love scenes in clothes warehouse are second to none
    144 tokens: this is easily the most underrated film inn the brooks cannon . sure , its flawed . it does not give a realistic view of homelessness ( unlike
    138 tokens: this is not the typical mel brooks film . it was much less slapstick than most of his movies and actually had a plot that was followable . les
```

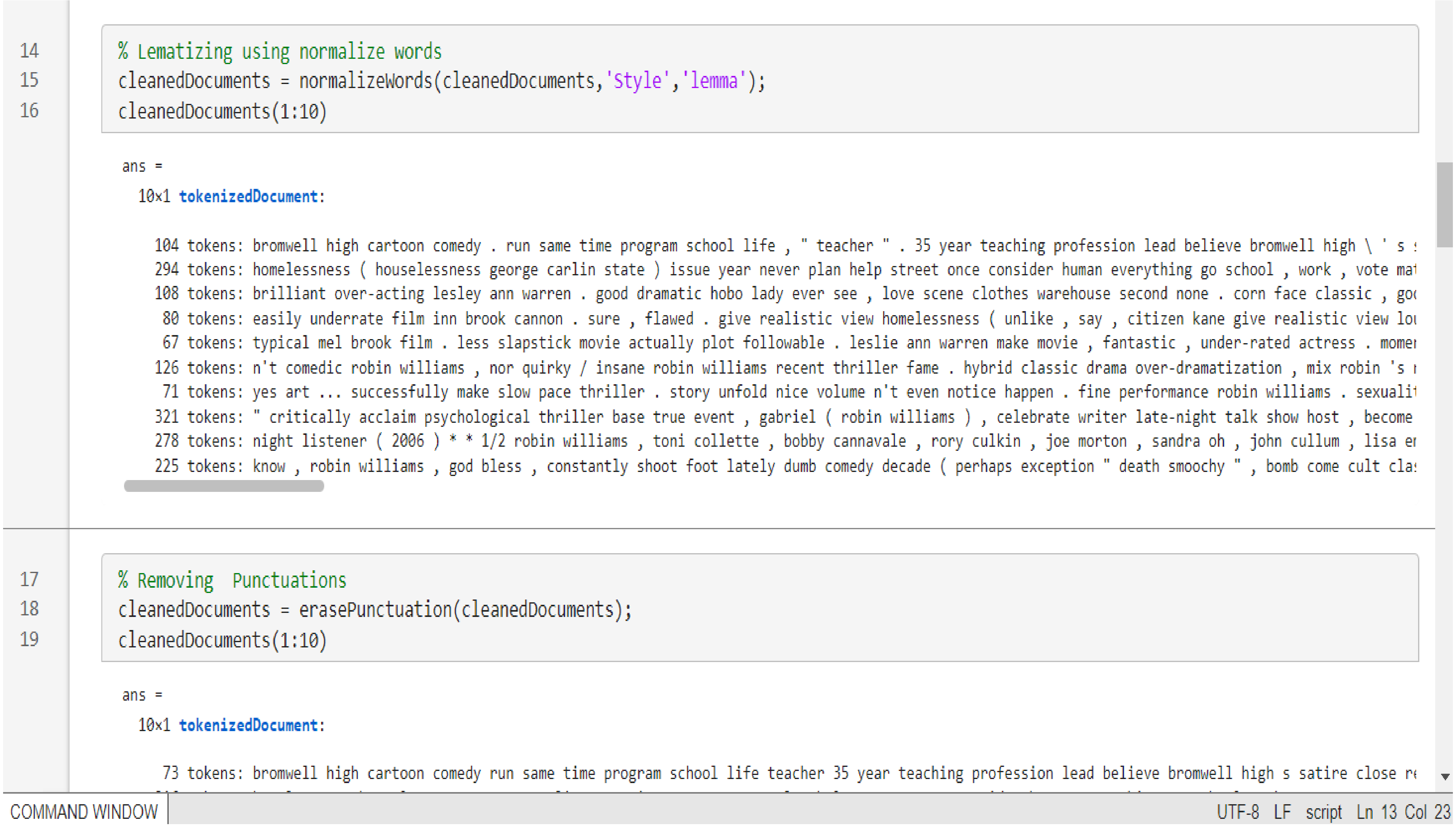


```
    144 tokens: this is easily the most underrated film inn the brooks cannon . sure , its flawed . it does not give a realistic view of homelessness ( unlike
    138 tokens: this is not the typical mel brooks film . it was much less slapstick than most of his movies and actually had a plot that was followable . les
    204 tokens: this is n't the comedic robin williams , nor is it the quirky / insane robin williams of recent thriller fame . this is a hybrid of the classi
    118 tokens: yes its an art ... to successfully make a slow paced thriller . the story unfolds in nice volumes while you do n't even notice it happening .
    445 tokens: in this " critically acclaimed psychological thriller based on true events , gabriel ( robin williams ) , a celebrated writer and late-night t
    476 tokens: the night listener ( 2006 ) * * 1/2 robin williams , toni collette , bobby cannavale , rory culkin , joe morton , sandra oh , john cullum , li
    383 tokens: you know , robin williams , god bless him , is constantly shooting himself in the foot lately with all these dumb comedies he has done this de
```

```
% Removing stop words
cleanedDocuments = removeStopWords(cleanedDocuments);
cleanedDocuments(1:10)
```

```
ans = 
  10×1 tokenizedDocument:

    104 tokens: bromwell high cartoon comedy . ran same time programs school life , " teachers " . 35 years teaching profession lead believe bromwell high \ '
    294 tokens: homelessness ( houselessness george carlin stated ) issue years never plan help street once considered human everything going school , work ,
    108 tokens: brilliant over-acting lesley ann warren . best dramatic hobo lady ever seen , love scenes clothes warehouse second none . corn face classic ,
     80 tokens: easily underrated film inn brooks cannon . sure , flawed . give realistic view homelessness ( unlike , say , citizen kane gave realistic view
     67 tokens: typical mel brooks film . less slapstick movies actually plot followable . leslie ann warren made movie , fantastic , under-rated actress . mo
    126 tokens: n't comedic robin williams , nor quirky / insane robin williams recent thriller fame . hybrid classic drama over-dramatization , mixed robin '
     71 tokens: yes art ... successfully make slow paced thriller . story unfolds nice volumes n't even notice happening . fine performance robin williams . s
    321 tokens: " critically acclaimed psychological thriller based true events , gabriel ( robin williams ) , celebrated writer late-night talk show host , b
    278 tokens: night listener ( 2006 ) * * 1/2 robin williams , toni collette , bobby cannavale , rory culkin , joe morton , sandra oh , john cullum , lisa e
    225 tokens: know , robin williams , god bless , constantly shooting foot lately dumb comedies decade ( perhaps exception " death smoochy " , bombed came c
```

```
% Lematizing using normalize words
cleanedDocuments = normalizeWords(cleanedDocuments,'Style','lemma');
cleanedDocuments(1:10)
```

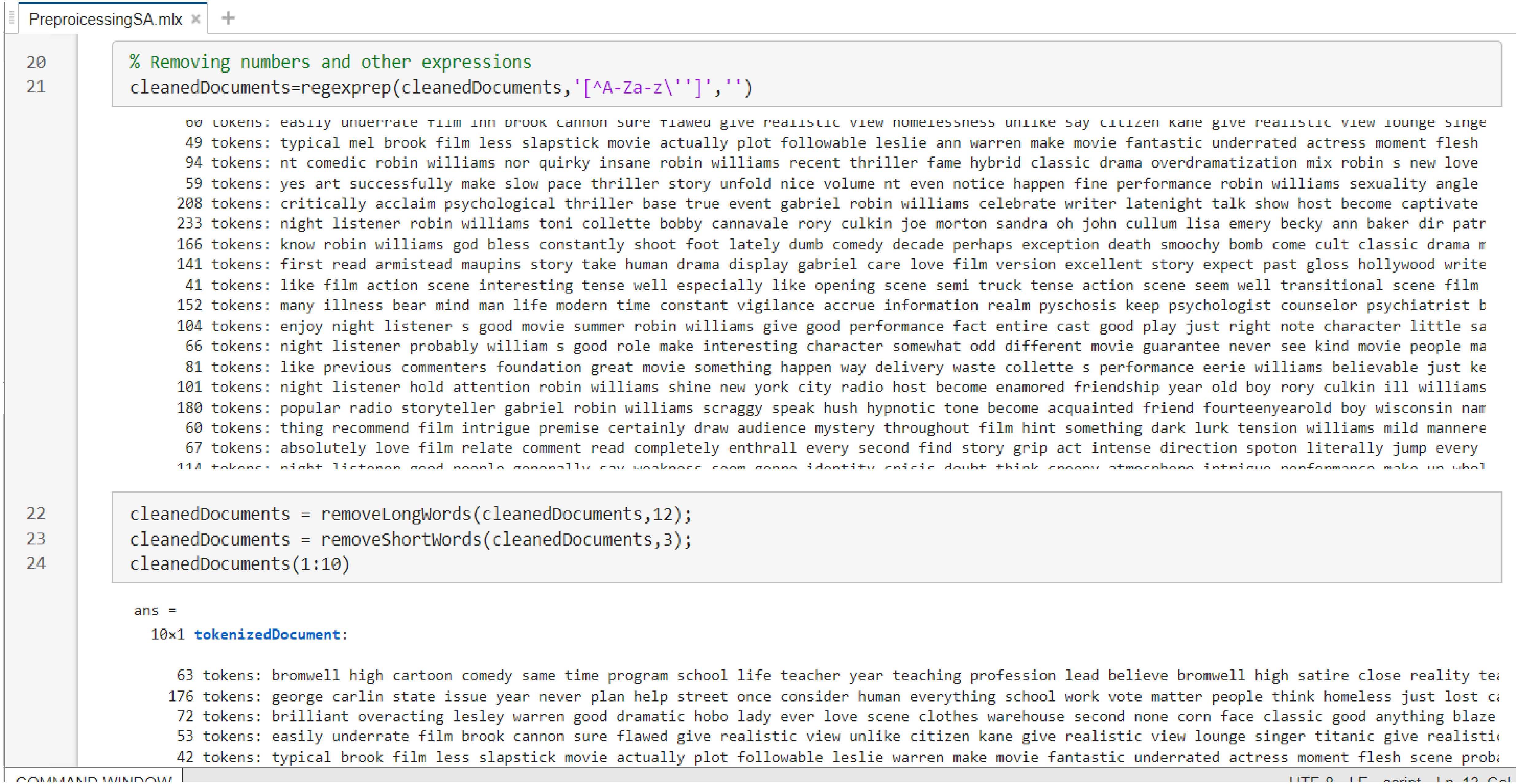


```
% Lematizing using normalize words
cleanedDocuments = normalizeWords(cleanedDocuments,'Style','lemma');
cleanedDocuments(1:10)
```

```
ans = 
  10×1 tokenizedDocument:

    104 tokens: bromwell high cartoon comedy . run same time program school life , " teacher " . 35 year teaching profession lead believe bromwell high \ ' s :
    294 tokens: homelessness ( houselessness george carlin state ) issue year never plan help street once consider human everything go school , work , vote mat
    108 tokens: brilliant over-acting lesley ann warren . good dramatic hobo lady ever see , love scene clothes warehouse second none . corn face classic , goo
     80 tokens: easily underrate film inn brook cannon . sure , flawed . give realistic view homelessness ( unlike , say , citizen kane give realistic view lou
     67 tokens: typical mel brook film . less slapstick movie actually plot followable . leslie ann warren make movie , fantastic , under-rated actress . momen
    126 tokens: n't comedic robin williams , nor quirky / insane robin williams recent thriller fame . hybrid classic drama over-dramatization , mix robin 's m
     71 tokens: yes art ... successfully make slow pace thriller . story unfold nice volume n't even notice happen . fine performance robin williams . sexualit
    321 tokens: " critically acclaim psychological thriller base true event , gabriel ( robin williams ) , celebrate writer late-night talk show host , become
    278 tokens: night listener ( 2006 ) * * 1/2 robin williams , toni collette , bobby cannavale , rory culkin , joe morton , sandra oh , john cullum , lisa er
    225 tokens: know , robin williams , god bless , constantly shoot foot lately dumb comedy decade ( perhaps exception " death smoochy " , bomb come cult clas
```

```
% Removing  Punctuations
cleanedDocuments = erasePunctuation(cleanedDocuments);
cleanedDocuments(1:10)
```

```
ans = 
  10×1 tokenizedDocument:

     73 tokens: bromwell high cartoon comedy run same time program school life teacher 35 year teaching profession lead believe bromwell high s satire close re
```



```
% Removing numbers and other expressions
cleanedDocuments=regexprep(cleanedDocuments,'[^A-Za-z\'']','')
```

```
     60 tokens: easily underrate film inn brook cannon sure flawed give realistic view homelessness unlike say citizen kane give realistic view lounge singe
     49 tokens: typical mel brook film less slapstick movie actually plot followable leslie ann warren make movie fantastic underrated actress moment flesh
     94 tokens: nt comedic robin williams nor quirky insane robin williams recent thriller fame hybrid classic drama overdramatization mix robin s new love
     59 tokens: yes art successfully make slow pace thriller story unfold nice volume nt even notice happen fine performance robin williams sexuality angle
    208 tokens: critically acclaim psychological thriller base true event gabriel robin williams celebrate writer latenight talk show host become captivate
    233 tokens: night listener robin williams toni collette bobby cannavale rory culkin joe morton sandra oh john cullum lisa emery becky ann baker dir patr
    166 tokens: know robin williams god bless constantly shoot foot lately dumb comedy decade perhaps exception death smoochy bomb come cult classic drama m
    141 tokens: first read armistead maupins story take human drama display gabriel care love film version excellent story expect past gloss hollywood write
     41 tokens: like film action scene interesting tense well especially like opening scene semi truck tense action scene seem well transitional scene film
    152 tokens: many illness bear mind man life modern time constant vigilance accrue information realm pyschosis keep psychologist counselor psychiatrist b
    104 tokens: enjoy night listener s good movie summer robin williams give good performance fact entire cast good play just right note character little sa
     66 tokens: night listener probably william s good role make interesting character somewhat odd different movie guarantee never see kind movie people ma
     81 tokens: like previous commenters foundation great movie something happen way delivery waste collette s performance eerie williams believable just ke
    101 tokens: night listener hold attention robin williams shine new york city radio host become enamored friendship year old boy rory culkin ill williams
    180 tokens: popular radio storyteller gabriel robin williams scraggy speak hush hypnotic tone become acquainted friend fourteenyearold boy wisconsin nam
     60 tokens: thing recommend film intrigue premise certainly draw audience mystery throughout film hint something dark lurk tension williams mild mannere
     67 tokens: absolutely love film relate comment read completely enthrall every second find story grip act intense direction spoton literally jump every
    114 tokens: night listener good people generally say weakness seem genre identity crisis doubt think creepy atmosphere intrigue performance make up whol
```

```
cleanedDocuments = removeLongWords(cleanedDocuments,12);
cleanedDocuments = removeShortWords(cleanedDocuments,3);
cleanedDocuments(1:10)
```

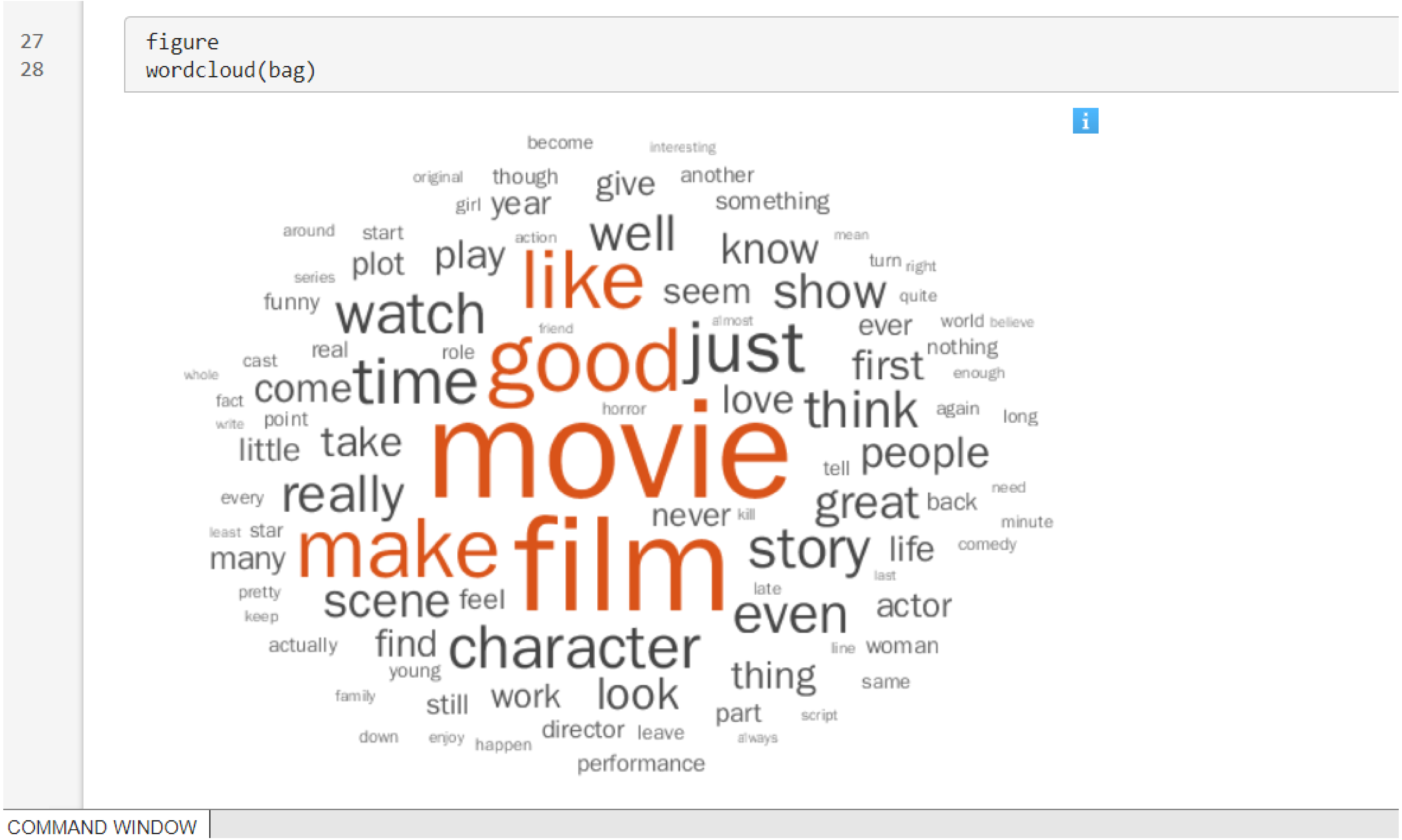
```
ans = 
  10×1 tokenizedDocument:

    63 tokens: bromwell high cartoon comedy same time program school life teacher year teaching profession lead believe bromwell high satire close reality tea
   176 tokens: george carlin state issue year never plan help street once consider human everything school work vote matter people think homeless just lost ca
    72 tokens: brilliant overacting lesley warren good dramatic hobo lady ever love scene clothes warehouse second none corn face classic good anything blaze
    53 tokens: easily underrate film brook cannon sure flawed give realistic view unlike citizen kane give realistic view lounge singer titanic give realisti
    42 tokens: typical brook film less slapstick movie actually plot followable leslie warren make movie fantastic underrated actress moment flesh scene proba
   316 tokens: back remember year clinton dan clone research unfortunate death princess diana marlin win world series woman give birth septuplets big year
   332 tokens: james cameron s titanic essentially romantic adventure visual grandeur magnificence timeless tragic love story set against background major
```

```
cleanedDocuments = removeLongWords(cleanedDocuments,12);
cleanedDocuments = removeShortWords(cleanedDocuments,3);
cleanedDocuments(1:10)
```

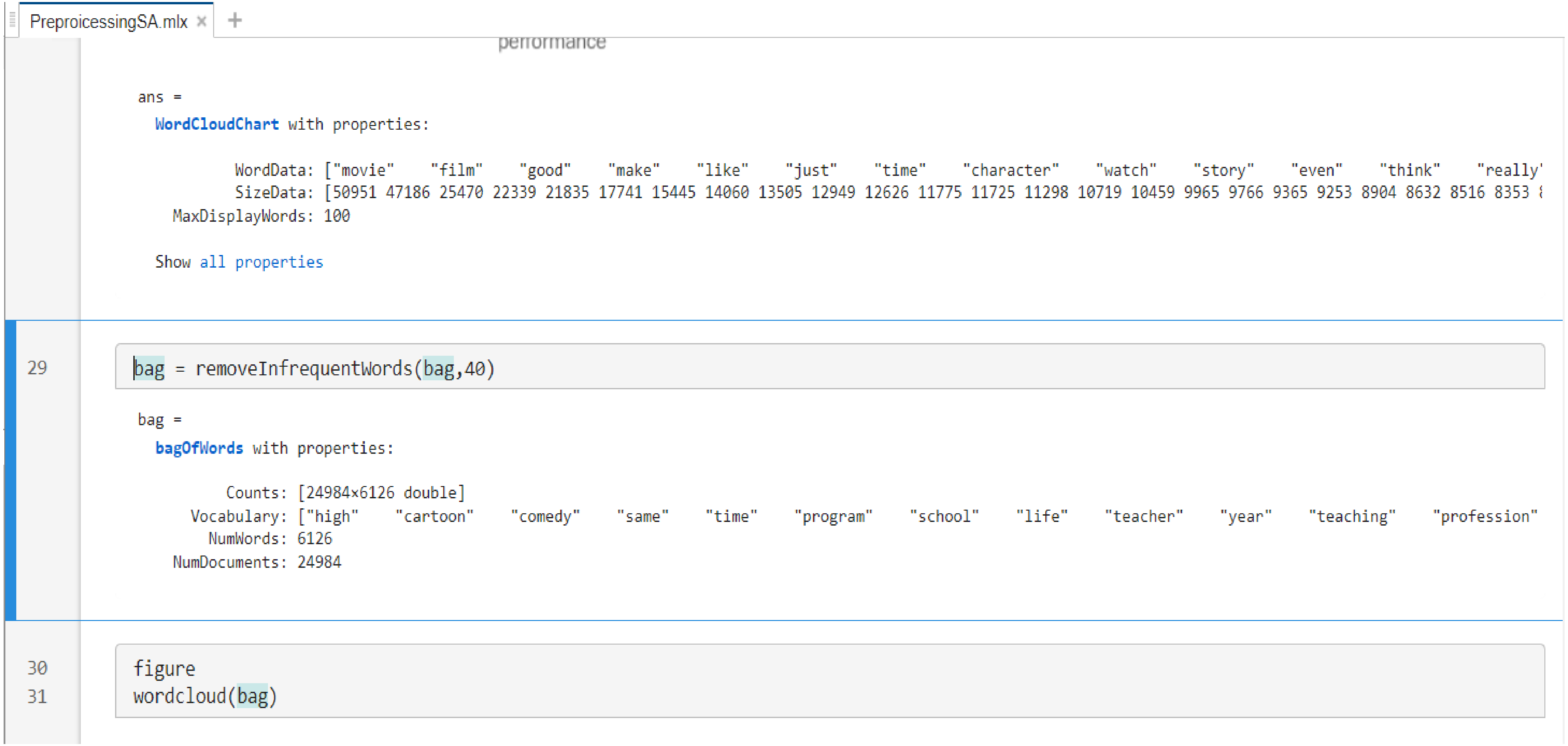
```
ans = 
  10×1 tokenizedDocument:

     63 tokens: bromwell high cartoon comedy same time program school life teacher year teaching profession lead believe bromwell high satire close reality te
    176 tokens: george carlin state issue year never plan help street once consider human everything school work vote matter people think homeless just lost c
     72 tokens: brilliant overacting lesley warren good dramatic hobo lady ever love scene clothes warehouse second none corn face classic good anything blaze
     53 tokens: easily underrate film brook cannon sure flawed give realistic view unlike citizen kane give realistic view lounge singer titanic give realisti
     42 tokens: typical brook film less slapstick movie actually plot followable leslie warren make movie fantastic underrated actress moment flesh scene prob
     77 tokens: comedic robin williams quirky insane robin williams recent thriller fame hybrid classic drama robin love thriller thriller mystery suspense ve
     52 tokens: make slow pace thriller story unfold nice volume even notice happen fine performance robin williams sexuality angle film seem unnecessary prob
    170 tokens: critically acclaim thriller base true event gabriel robin williams celebrate writer latenight talk show host become captivate harrowing story
    199 tokens: night listener robin williams toni collette bobby cannavale rory culkin morton sandra john cullum lisa emery becky baker patrick stettner susp
    152 tokens: know robin williams bless constantly shoot foot lately dumb comedy decade perhaps exception death smoochy bomb come cult classic drama make la
```

```
% Bag of Words
 bag = bagOfWords(cleanedDocuments)
```

```
bag = 
  bagOfWords with properties:

          Counts: [24984×65099 double]
      Vocabulary: ["bromwell"    "high"    "cartoon"    "comedy"    "same"    "time"    "program"    "school"    "life"    "teacher"    "year"    "teaching"
        NumWords: 65099
    NumDocuments: 24984
```



```
figure
wordcloud(bag)
```

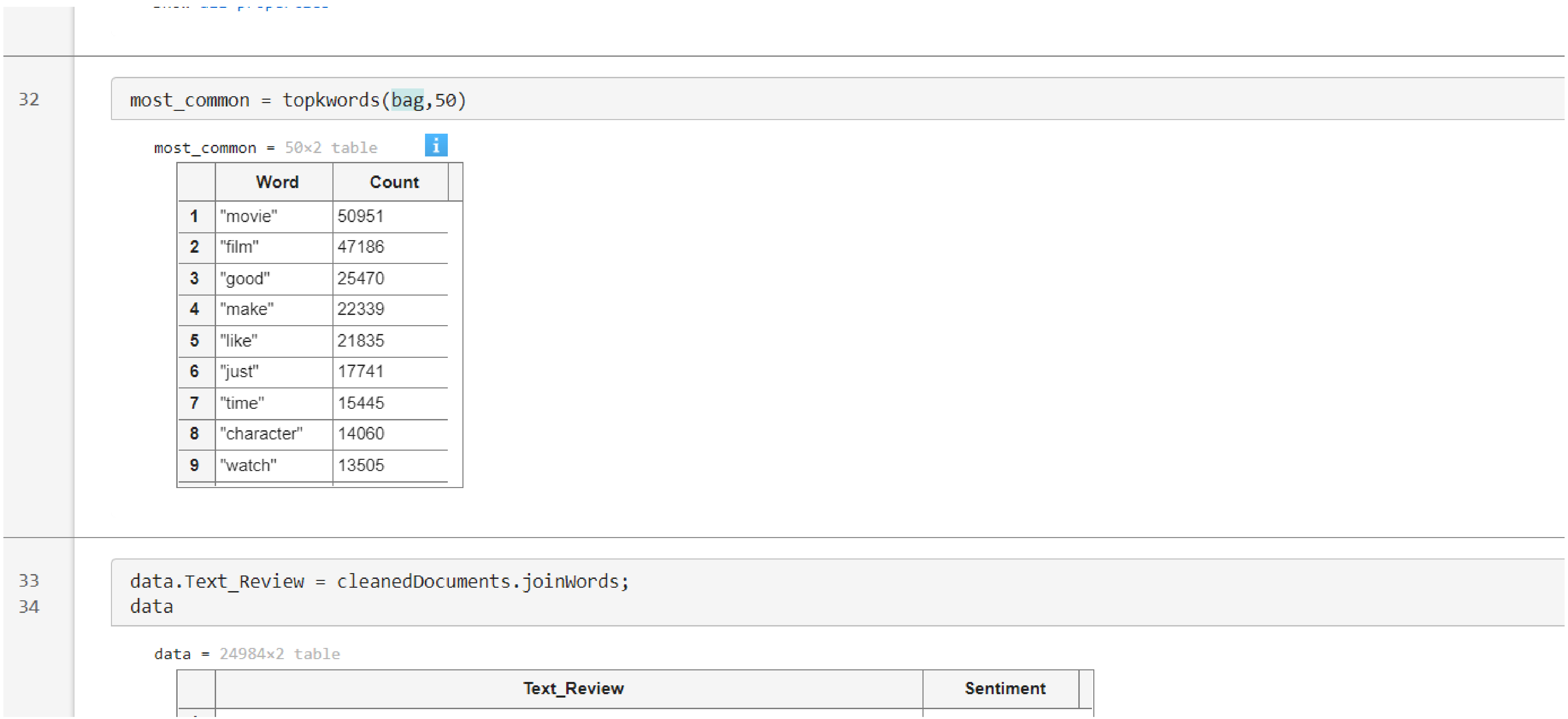


```
ans = 
  WordCloudChart with properties:

           WordData: ["movie"    "film"    "good"    "make"    "like"    "just"    "time"    "character"    "watch"    "story"    "even"    "think"    "really'
           SizeData: [50951 47186 25470 22339 21835 17741 15445 14060 13505 12949 12626 11775 11725 11298 10719 10459 9965 9766 9365 9253 8904 8632 8516 8353
    MaxDisplayWords: 100

  Show all properties
```

```
bag = removeInfrequentWords(bag,40)
```

```
bag = 
  bagOfWords with properties:

          Counts: [24984×6126 double]
      Vocabulary: ["high"    "cartoon"    "comedy"    "same"    "time"    "program"    "school"    "life"    "teacher"    "year"    "teaching"    "profession"
        NumWords: 6126
    NumDocuments: 24984
```

```
figure
wordcloud(bag)
```



Show all properties

```
most_common = topkwords(bag,50)
```

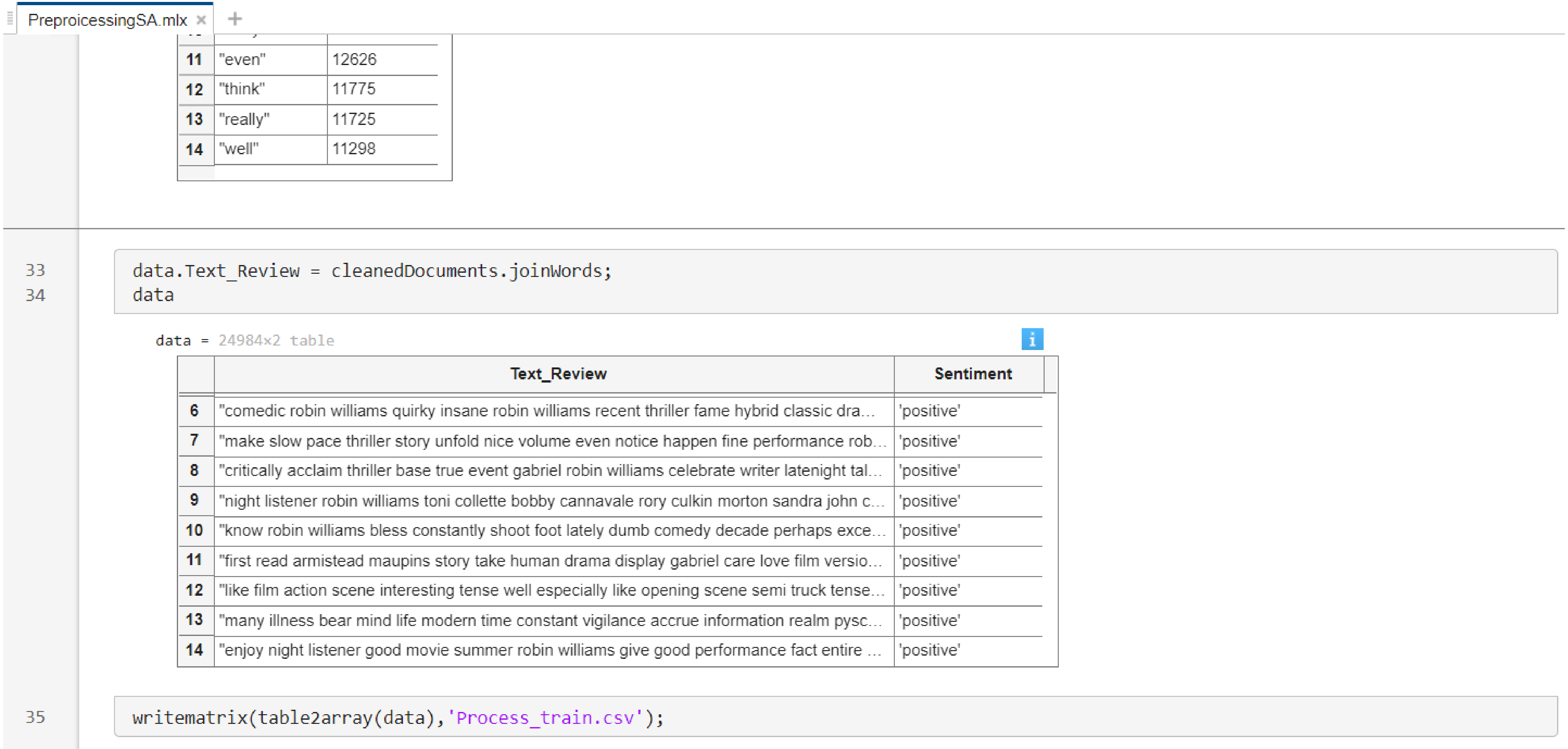
most_common = 50×2 table

	Word	Count
1	"movie"	50951
2	"film"	47186
3	"good"	25470
4	"make"	22339
5	"like"	21835
6	"just"	17741
7	"time"	15445
8	"character"	14060
9	"watch"	13505

```
data.Text_Review = cleanedDocuments.joinWords;
data
```

data = 24984×2 table

	Text_Review	Sentiment



11	"even"	12626
12	"think"	11775
13	"really"	11725
14	"well"	11298

```
data.Text_Review = cleanedDocuments.joinWords;
data
```

```
data = 24984×2 table
```

	Text_Review	Sentiment
6	"comedic robin williams quirky insane robin williams recent thriller fame hybrid classic dra…	'positive'
7	"make slow pace thriller story unfold nice volume even notice happen fine performance rob…	'positive'
8	"critically acclaim thriller base true event gabriel robin williams celebrate writer latenight tal…	'positive'
9	"night listener robin williams toni collette bobby cannavale rory culkin morton sandra john c…	'positive'
10	"know robin williams bless constantly shoot foot lately dumb comedy decade perhaps exce…	'positive'
11	"first read armistead maupins story take human drama display gabriel care love film versio…	'positive'
12	"like film action scene interesting tense well especially like opening scene semi truck tense…	'positive'
13	"many illness bear mind life modern time constant vigilance accrue information realm pysc…	'positive'
14	"enjoy night listener good movie summer robin williams give good performance fact entire …	'positive'

```
writematrix(table2array(data),'Process_train.csv');
```