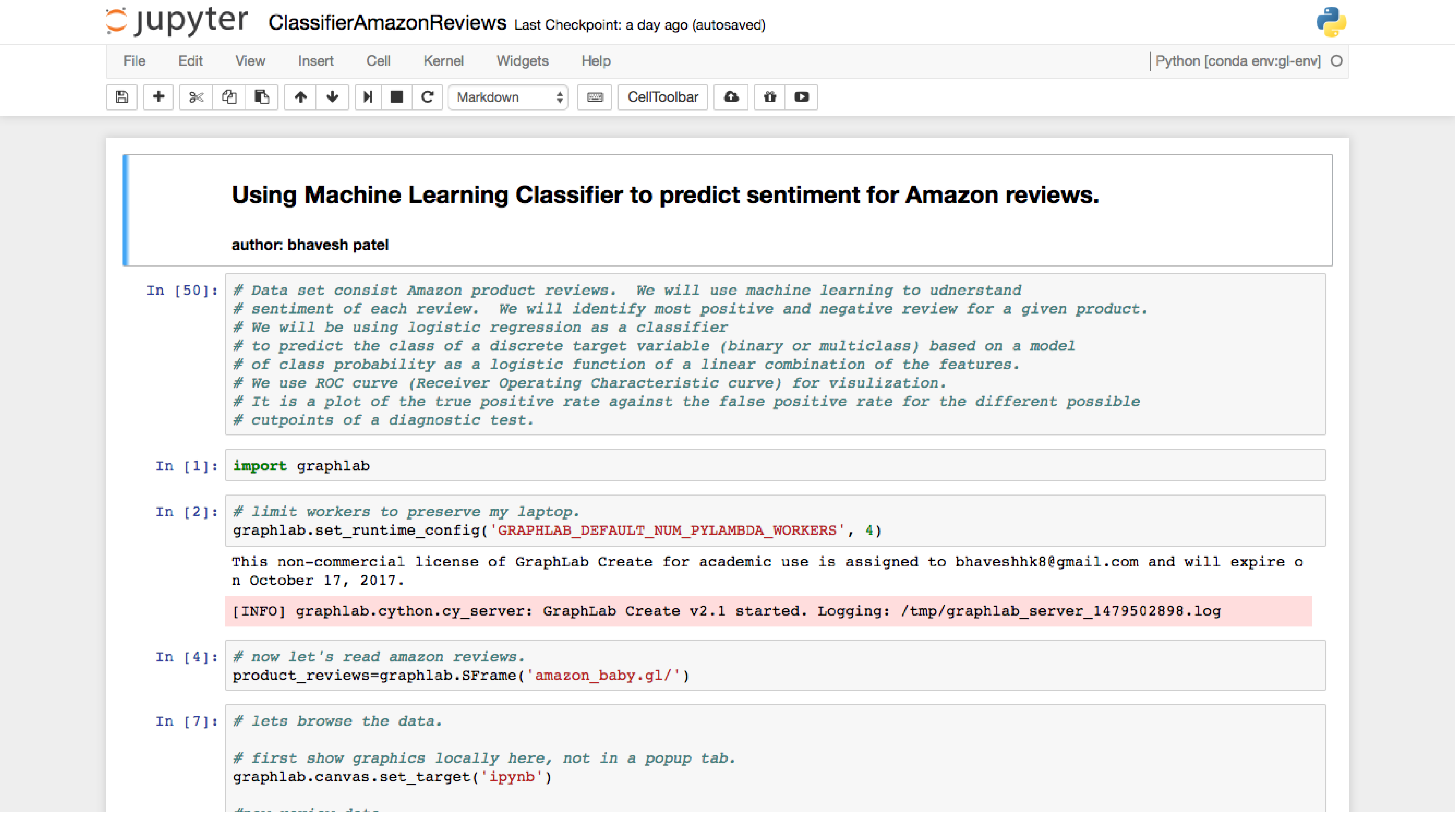

jupyter
ClassifierAmazonReviews Last Checkpoint: a day ago (autosaved)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Python [conda env:gl-env]
Markdown
CellToolbar

Using Machine Learning Classifier to predict sentiment for Amazon reviews.

author: bhavesh patel

In [50]:
```
# Data set consist Amazon product reviews.  We will use machine learning to udnerstand
# sentiment of each review.  We will identify most positive and negative review for a given product.
# We will be using logistic regression as a classifier
# to predict the class of a discrete target variable (binary or multiclass) based on a model
# of class probability as a logistic function of a linear combination of the features.
# We use ROC curve (Receiver Operating Characteristic curve) for visulization.
# It is a plot of the true positive rate against the false positive rate for the different possible
# cutpoints of a diagnostic test.
```

In [1]:
```
import graphlab
```

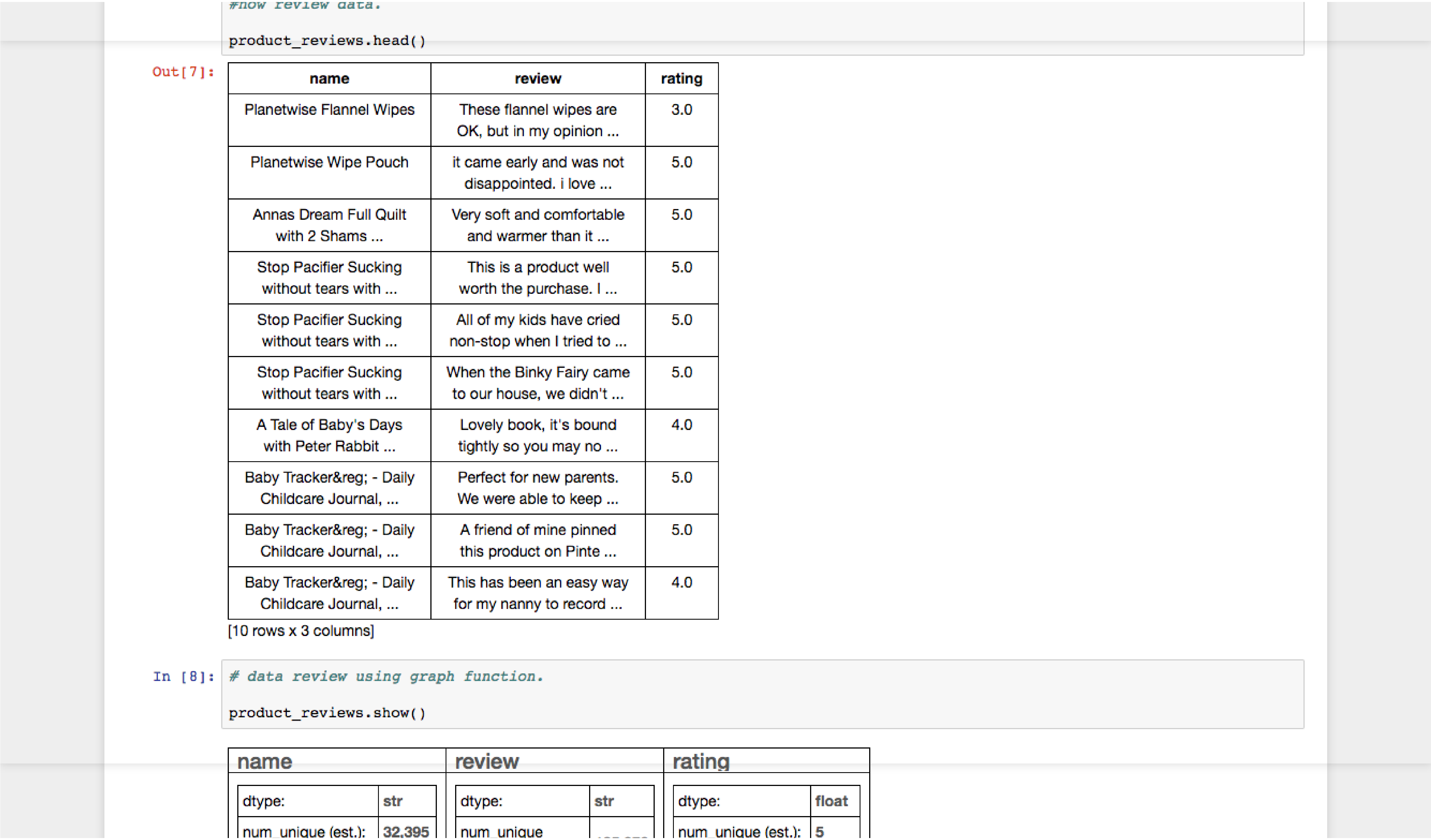
In [2]:
```
# limit workers to preserve my laptop.
graphlab.set_runtime_config('GRAPHLAB_DEFAULT_NUM_PYLAMBDA_WORKERS', 4)
```

```
This non-commercial license of GraphLab Create for academic use is assigned to bhaveshhk8@gmail.com and will expire o
n October 17, 2017.
```

```
[INFO] graphlab.cython.cy_server: GraphLab Create v2.1 started. Logging: /tmp/graphlab_server_1479502898.log
```

In [4]:
```
# now let's read amazon reviews.
product_reviews=graphlab.SFrame('amazon_baby.gl/')
```

In [7]:
```
# lets browse the data.

# first show graphics locally here, not in a popup tab.
graphlab.canvas.set_target('ipynb')

#now review data
```



```
#now review data.

product_reviews.head()
```

Out[7]:

name	review	rating
Planetwise Flannel Wipes	These flannel wipes are OK, but in my opinion ...	3.0
Planetwise Wipe Pouch	it came early and was not disappointed. i love ...	5.0
Annas Dream Full Quilt with 2 Shams ...	Very soft and comfortable and warmer than it ...	5.0
Stop Pacifier Sucking without tears with ...	This is a product well worth the purchase. I ...	5.0
Stop Pacifier Sucking without tears with ...	All of my kids have cried non-stop when I tried to ...	5.0
Stop Pacifier Sucking without tears with ...	When the Binky Fairy came to our house, we didn't ...	5.0
A Tale of Baby's Days with Peter Rabbit ...	Lovely book, it's bound tightly so you may no ...	4.0
Baby Tracker&reg; - Daily Childcare Journal, ...	Perfect for new parents. We were able to keep ...	5.0
Baby Tracker&reg; - Daily Childcare Journal, ...	A friend of mine pinned this product on Pinte ...	5.0
Baby Tracker&reg; - Daily Childcare Journal, ...	This has been an easy way for my nanny to record ...	4.0

[10 rows x 3 columns]

In [8]:

```
# data review using graph function.

product_reviews.show()
```

name		review		rating	
dtype:	str	dtype:	str	dtype:	float
num_unique (est.):	32,395	num_unique		num_unique (est.):	5

num_undefined:	284

frequent items:

Vulli Sophie the ...
Simple Wishes ...
Infant Optics ...
Baby Einstein Take ...
Cloud b Twilight ...
Fisher-Price ...
Fisher-Price ...
Graco Nautilus ...
Leachco Snoogle ...
Regalo Easy Step ...
Baby Trend Diaper ...
Skip Hop Zoo Pack ...

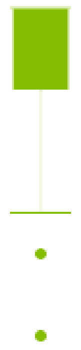
(est.):	185,979
num_undefined:	0

frequent items:

''

num_undefined:	0
min:	1
max:	5
median:	5
mean:	4.12
std:	1.285

distribution of values:

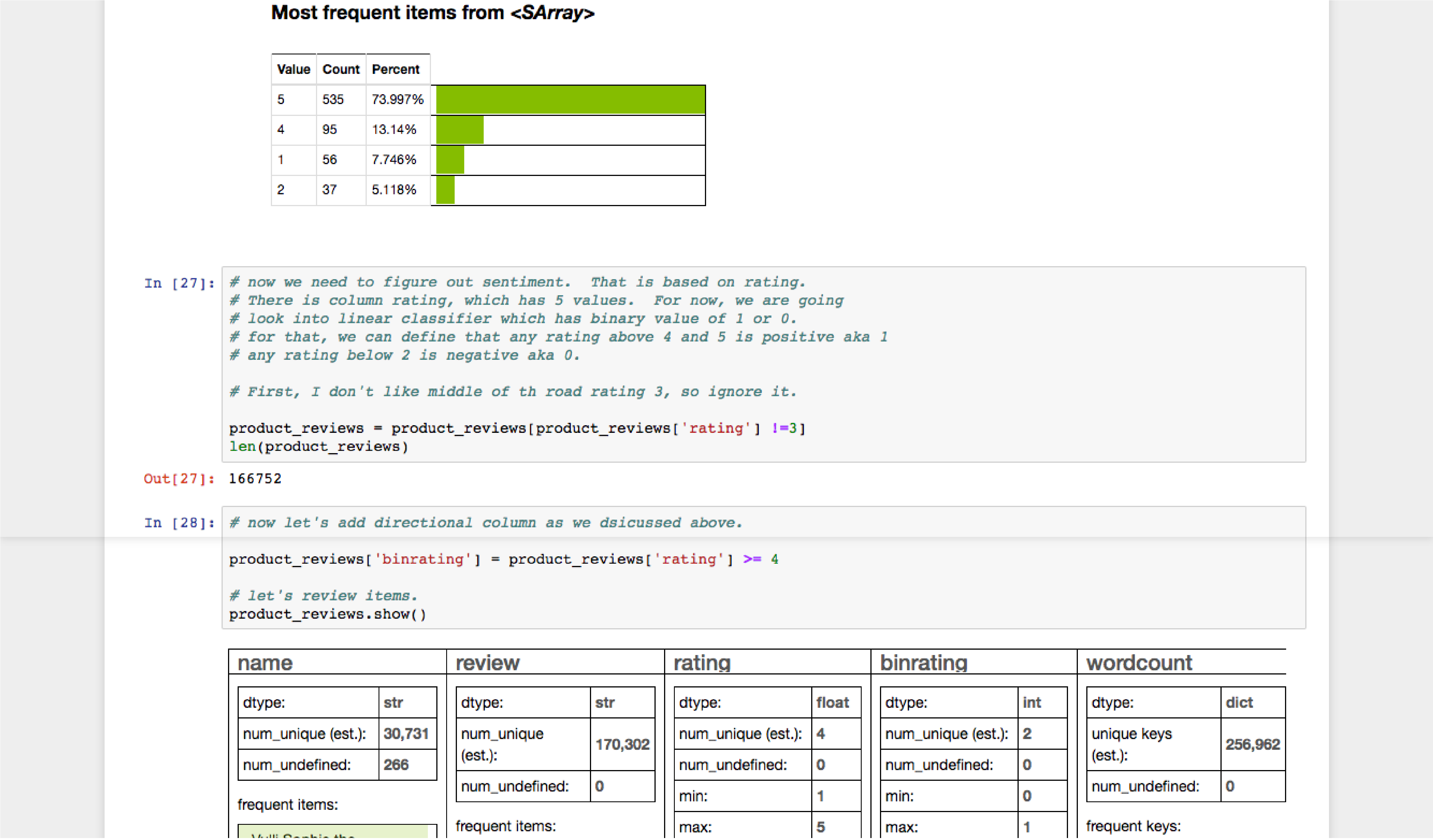


```
In [24]: # remeber the defination of accuracy, which is defined as number of correct gueses over total data set records.
         # Let's add word count to the data set.

         product_reviews['wordcount'] = graphlab.text_analytics.count_words(product_reviews['review'])
```

```
In [25]: # Vulli Shopie (it is a giraffer toy for baby teething) has the most data, so we will use this for futher analysis.

         # let's get all reviews for that.
         vs_reviews = product_reviews[product_reviews['name']=='Vulli Sophie the Giraffe Teether']

         # how many reviews for this product?
         len(vs_reviews)
```

```
Out[25]: 723
```

```
In [26]: vs_reviews['rating'].show(view='Categorical')
```



Most frequent items from *<SArray>*

Value	Count	Percent	
5	535	73.997%	🟩🟩🟩🟩🟩🟩🟩🟩🟩🟩
4	95	13.14%	🟩🟩
1	56	7.746%	🟩
2	37	5.118%	🟩

In [27]:
```
# now we need to figure out sentiment.  That is based on rating.
# There is column rating, which has 5 values.  For now, we are going
# look into linear classifier which has binary value of 1 or 0.
# for that, we can define that any rating above 4 and 5 is positive aka 1
# any rating below 2 is negative aka 0.

# First, I don't like middle of th road rating 3, so ignore it.

product_reviews = product_reviews[product_reviews['rating'] !=3]
len(product_reviews)
```

Out[27]:
```
166752
```

In [28]:
```
# now let's add directional column as we dsicussed above.

product_reviews['binrating'] = product_reviews['rating'] >= 4

# let's review items.
product_reviews.show()
```

name

dtype:	str
num_unique (est.):	30,731
num_undefined:	266

frequent items:

Vulli Sophie the

review

dtype:	str
num_unique (est.):	170,302
num_undefined:	0

frequent items:

rating

dtype:	float
num_unique (est.):	4
num_undefined:	0
min:	1
max:	5

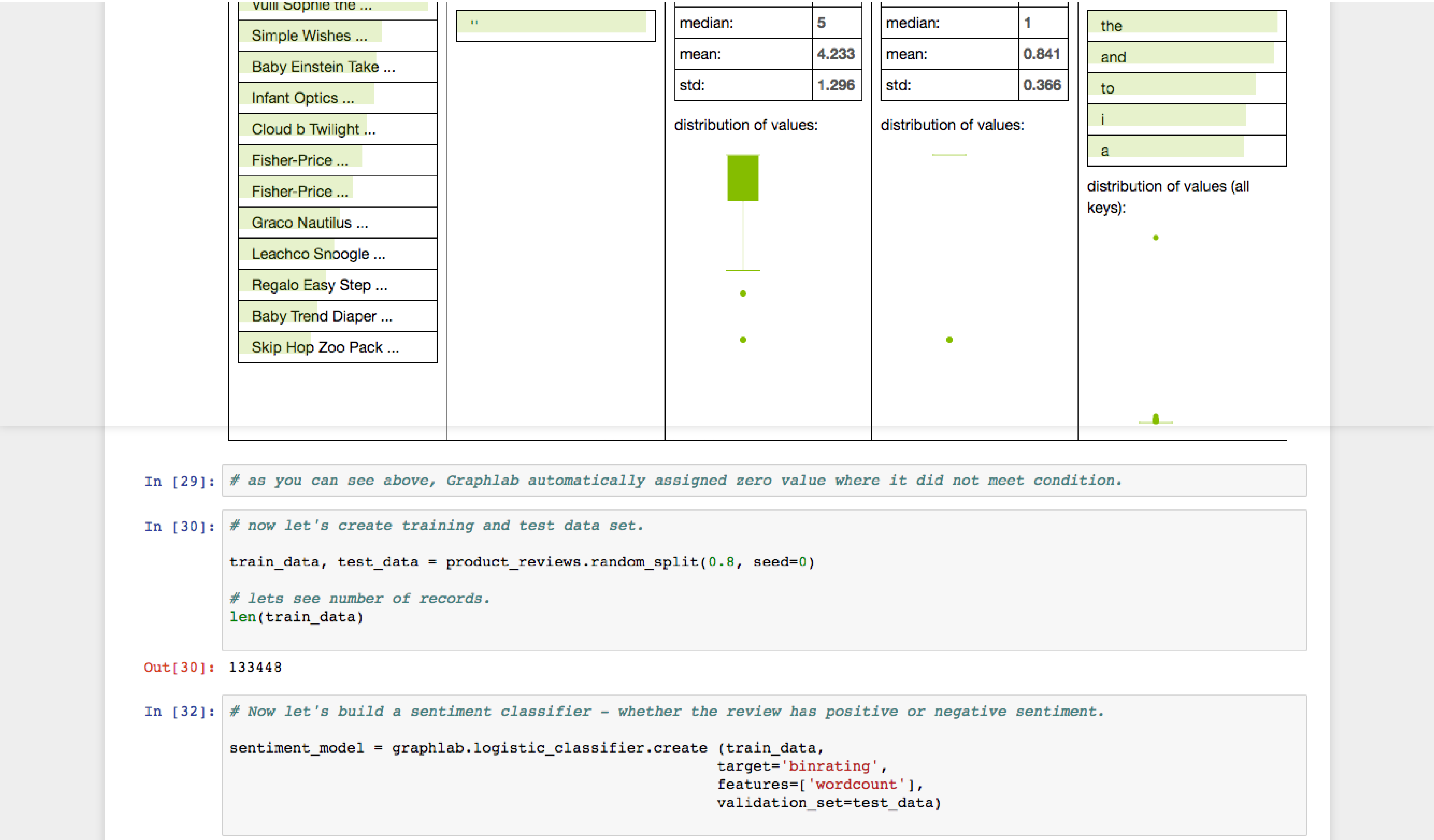
binrating

dtype:	int
num_unique (est.):	2
num_undefined:	0
min:	0
max:	1

wordcount

dtype:	dict
unique keys (est.):	256,962
num_undefined:	0

frequent keys:



Vulli Sophie the ...

Simple Wishes ...

Baby Einstein Take ...

Infant Optics ...

Cloud b Twilight ...

Fisher-Price ...

Fisher-Price ...

Graco Nautilus ...

Leachco Snoogle ...

Regalo Easy Step ...

Baby Trend Diaper ...

Skip Hop Zoo Pack ...

''

median:	5
mean:	4.233
std:	1.296

distribution of values:

median:	1
mean:	0.841
std:	0.366

distribution of values:

the

and

to

i

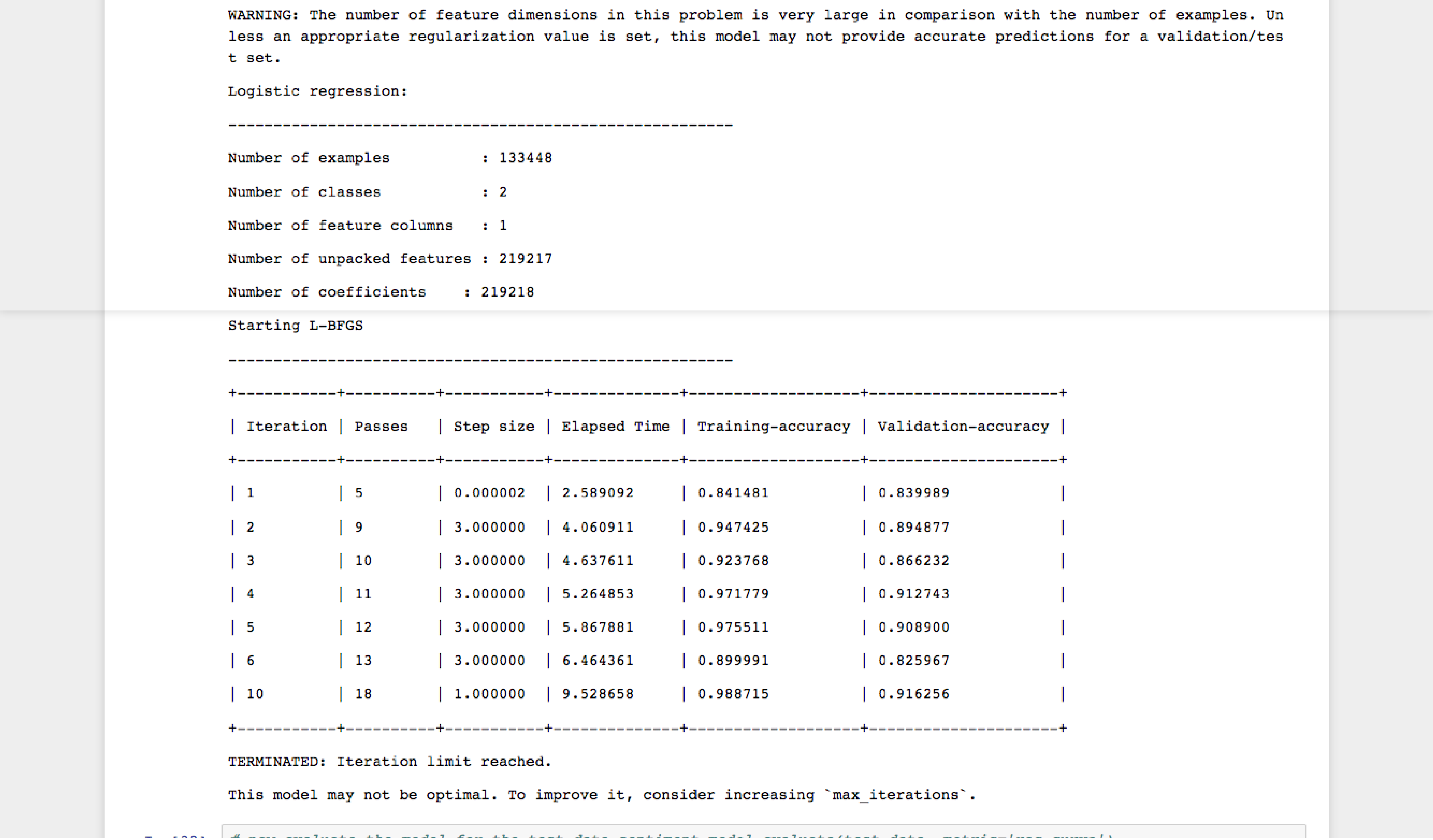
a

distribution of values (all keys):

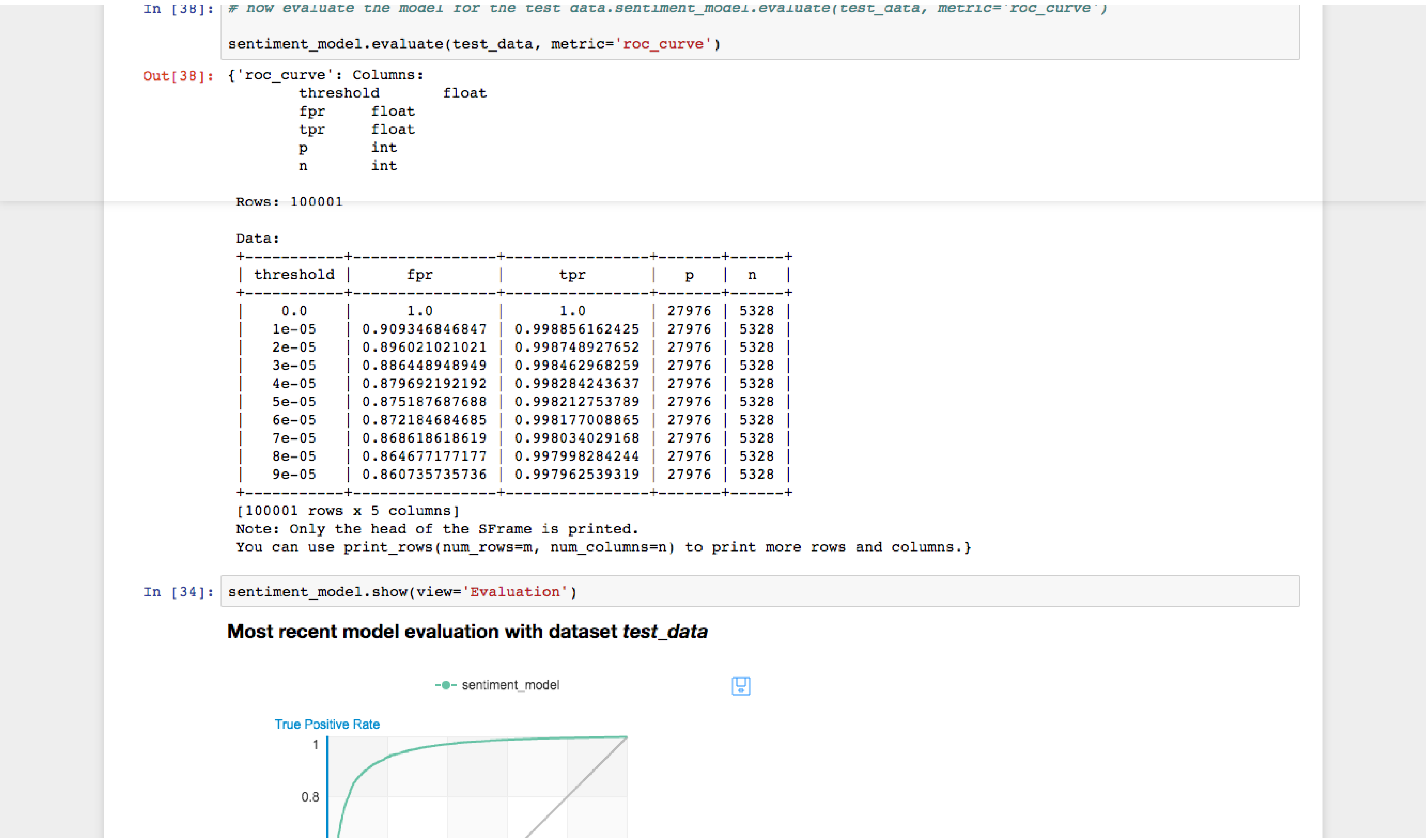
```
In [29]: # as you can see above, Graphlab automatically assigned zero value where it did not meet condition.
```

```
In [30]: # now let's create training and test data set.

         train_data, test_data = product_reviews.random_split(0.8, seed=0)

         # lets see number of records.
         len(train_data)

Out[30]: 133448
```

```
In [32]: # Now let's build a sentiment classifier - whether the review has positive or negative sentiment.

         sentiment_model = graphlab.logistic_classifier.create (train_data,
                                                                 target='binrating',
                                                                 features=['wordcount'],
                                                                 validation_set=test_data)
```



```
WARNING: The number of feature dimensions in this problem is very large in comparison with the number of examples. Un
less an appropriate regularization value is set, this model may not provide accurate predictions for a validation/tes
t set.

Logistic regression:

--------------------------------------------------------

Number of examples          : 133448

Number of classes           : 2

Number of feature columns   : 1

Number of unpacked features : 219217

Number of coefficients    : 219218

Starting L-BFGS

--------------------------------------------------------

+-----------+----------+-----------+--------------+-------------------+---------------------+

| Iteration | Passes   | Step size | Elapsed Time | Training-accuracy | Validation-accuracy |

+-----------+----------+-----------+--------------+-------------------+---------------------+

| 1         | 5        | 0.000002  | 2.589092     | 0.841481          | 0.839989            |

| 2         | 9        | 3.000000  | 4.060911     | 0.947425          | 0.894877            |

| 3         | 10       | 3.000000  | 4.637611     | 0.923768          | 0.866232            |

| 4         | 11       | 3.000000  | 5.264853     | 0.971779          | 0.912743            |

| 5         | 12       | 3.000000  | 5.867881     | 0.975511          | 0.908900            |

| 6         | 13       | 3.000000  | 6.464361     | 0.899991          | 0.825967            |

| 10        | 18       | 1.000000  | 9.528658     | 0.988715          | 0.916256            |

+-----------+----------+-----------+--------------+-------------------+---------------------+

TERMINATED: Iteration limit reached.

This model may not be optimal. To improve it, consider increasing `max_iterations`.
```

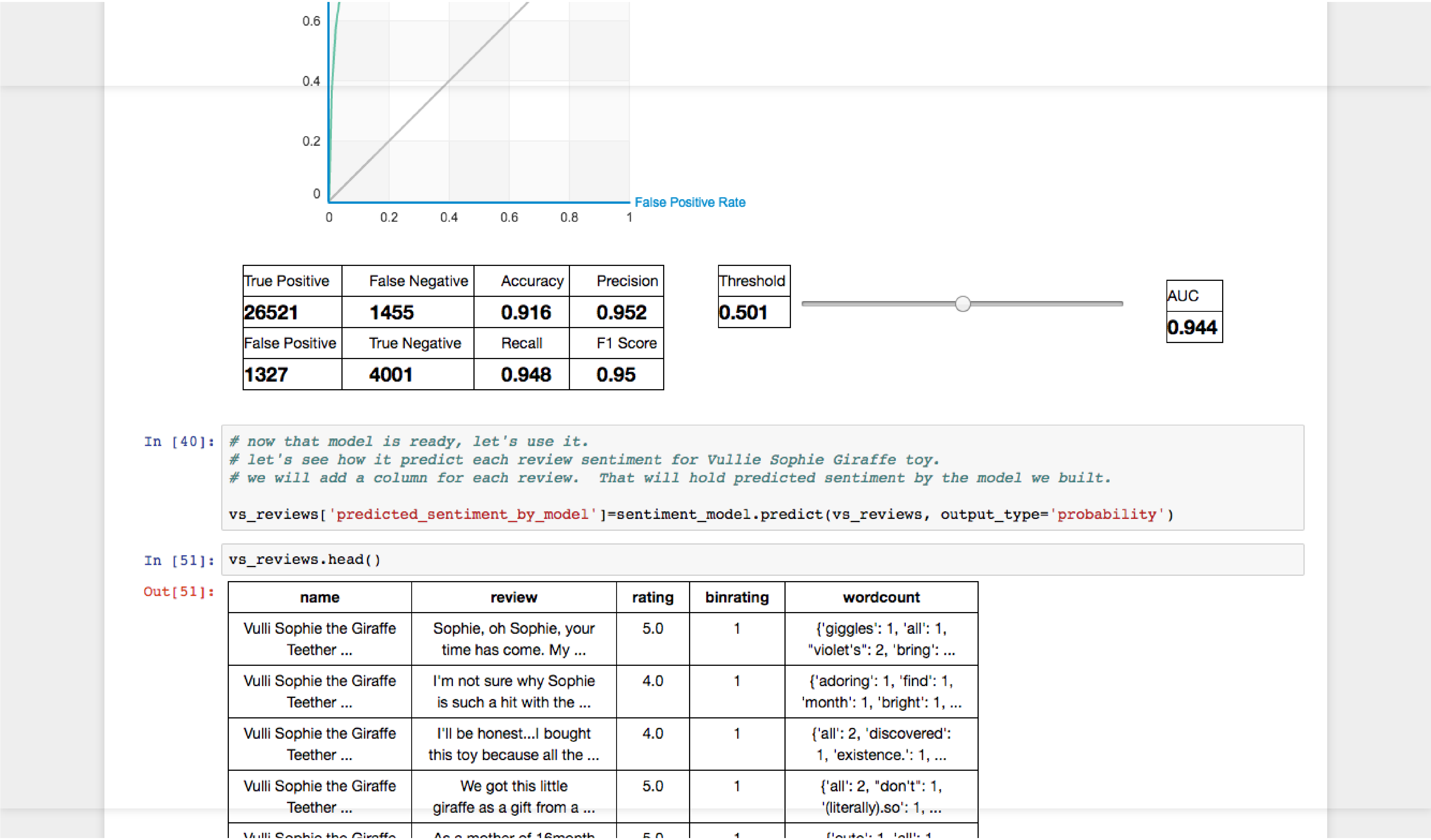


```
In [38]: # now evaluate the model for the test data.sentiment_model.evaluate(test_data, metric='roc_curve')

         sentiment_model.evaluate(test_data, metric='roc_curve')
```

```
Out[38]: {'roc_curve': Columns:
                 threshold       float
                 fpr     float
                 tpr     float
                 p       int
                 n       int

          Rows: 100001

          Data:
          +-----------+----------------+----------------+-------+------+
          | threshold |      fpr       |      tpr       |   p   |  n   |
          +-----------+----------------+----------------+-------+------+
          |    0.0    |      1.0       |      1.0       | 27976 | 5328 |
          |   1e-05   | 0.909346846847 | 0.998856162425 | 27976 | 5328 |
          |   2e-05   | 0.896021021021 | 0.998748927652 | 27976 | 5328 |
          |   3e-05   | 0.886448948949 | 0.998462968259 | 27976 | 5328 |
          |   4e-05   | 0.879692192192 | 0.998284243637 | 27976 | 5328 |
          |   5e-05   | 0.875187687688 | 0.998212753789 | 27976 | 5328 |
          |   6e-05   | 0.872184684685 | 0.998177008865 | 27976 | 5328 |
          |   7e-05   | 0.868618618619 | 0.998034029168 | 27976 | 5328 |
          |   8e-05   | 0.864677177177 | 0.997998284244 | 27976 | 5328 |
          |   9e-05   | 0.860735735736 | 0.997962539319 | 27976 | 5328 |
          +-----------+----------------+----------------+-------+------+
          [100001 rows x 5 columns]
          Note: Only the head of the SFrame is printed.
          You can use print_rows(num_rows=m, num_columns=n) to print more rows and columns.}
```

```
In [34]: sentiment_model.show(view='Evaluation')
```

Most recent model evaluation with dataset *test_data*



True Positive	False Negative	Accuracy	Precision
26521	**1455**	**0.916**	**0.952**
False Positive	True Negative	Recall	F1 Score
1327	**4001**	**0.948**	**0.95**

Threshold
0.501

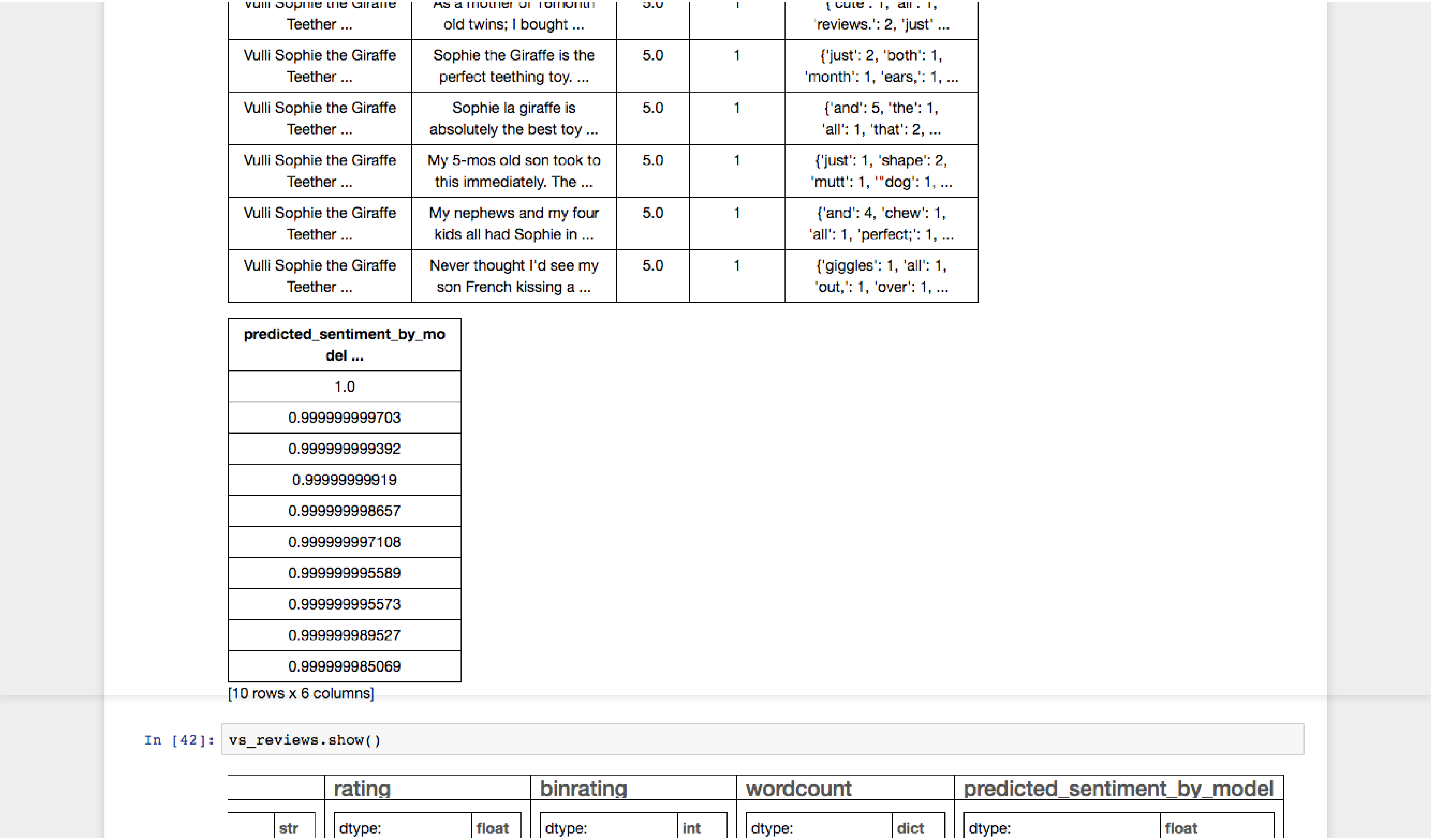
AUC
0.944

```
In [40]: # now that model is ready, let's use it.
         # let's see how it predict each review sentiment for Vullie Sophie Giraffe toy.
         # we will add a column for each review.  That will hold predicted sentiment by the model we built.

         vs_reviews['predicted_sentiment_by_model']=sentiment_model.predict(vs_reviews, output_type='probability')
```

```
In [51]: vs_reviews.head()
```

Out[51]:

name	review	rating	binrating	wordcount
Vulli Sophie the Giraffe Teether ...	Sophie, oh Sophie, your time has come. My ...	5.0	1	{'giggles': 1, 'all': 1, "violet's": 2, 'bring': ...
Vulli Sophie the Giraffe Teether ...	I'm not sure why Sophie is such a hit with the ...	4.0	1	{'adoring': 1, 'find': 1, 'month': 1, 'bright': 1, ...
Vulli Sophie the Giraffe Teether ...	I'll be honest...I bought this toy because all the ...	4.0	1	{'all': 2, 'discovered': 1, 'existence.': 1, ...
Vulli Sophie the Giraffe Teether ...	We got this little giraffe as a gift from a ...	5.0	1	{'all': 2, "don't": 1, '(literally).so': 1, ...
Vulli Sophie the Giraffe	As a mother of 16month	5.0	1	{'cute': 1, 'all': 1



Vulli Sophie the Giraffe Teether ...	As a mother of 16month old twins; I bought ...	5.0	1	{'cute': 1, 'all': 1, 'reviews.': 2, 'just' ...
Vulli Sophie the Giraffe Teether ...	Sophie the Giraffe is the perfect teething toy. ...	5.0	1	{'just': 2, 'both': 1, 'month': 1, 'ears,': 1, ...
Vulli Sophie the Giraffe Teether ...	Sophie la giraffe is absolutely the best toy ...	5.0	1	{'and': 5, 'the': 1, 'all': 1, 'that': 2, ...
Vulli Sophie the Giraffe Teether ...	My 5-mos old son took to this immediately. The ...	5.0	1	{'just': 1, 'shape': 2, 'mutt': 1, '"dog': 1, ...
Vulli Sophie the Giraffe Teether ...	My nephews and my four kids all had Sophie in ...	5.0	1	{'and': 4, 'chew': 1, 'all': 1, 'perfect;': 1, ...
Vulli Sophie the Giraffe Teether ...	Never thought I'd see my son French kissing a ...	5.0	1	{'giggles': 1, 'all': 1, 'out,': 1, 'over': 1, ...

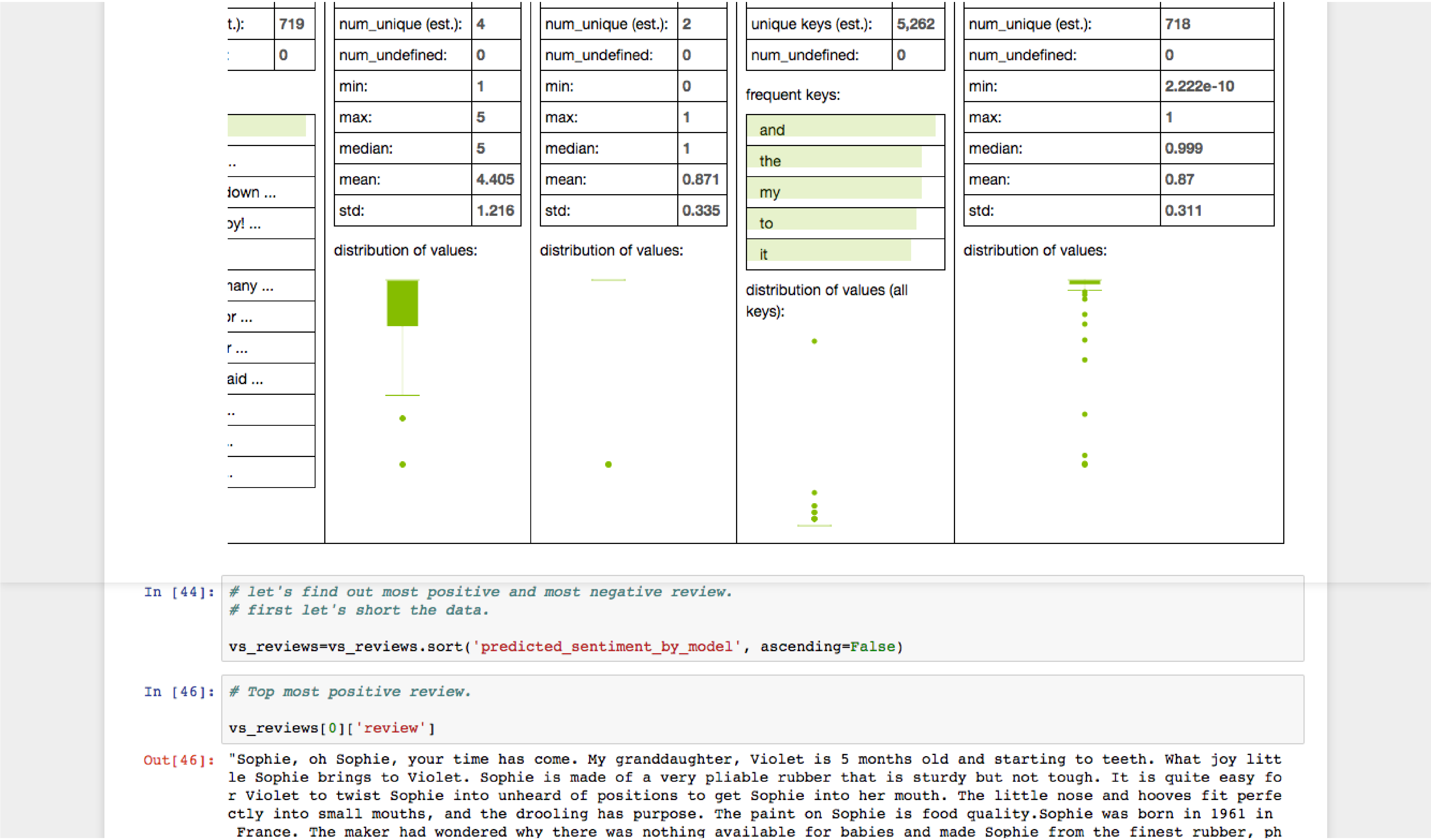
predicted_sentiment_by_model ...
1.0
0.999999999703
0.999999999392
0.99999999919
0.999999998657
0.999999997108
0.999999995589
0.999999995573
0.999999989527
0.999999985069

[10 rows x 6 columns]

In [42]:

```
vs_reviews.show()
```

	rating	binrating	wordcount	predicted_sentiment_by_model
str	dtype: float	dtype: int	dtype: dict	dtype: float



t.):	719
:	0

..
lown ...
oy! ...
nany ...
or ...
r ...
aid ...
..
..
.

num_unique (est.):	4
num_undefined:	0
min:	1
max:	5
median:	5
mean:	4.405
std:	1.216

distribution of values:

num_unique (est.):	2
num_undefined:	0
min:	0
max:	1
median:	1
mean:	0.871
std:	0.335

distribution of values:

unique keys (est.):	5,262
num_undefined:	0

frequent keys:

and
the
my
to
it

distribution of values (all keys):

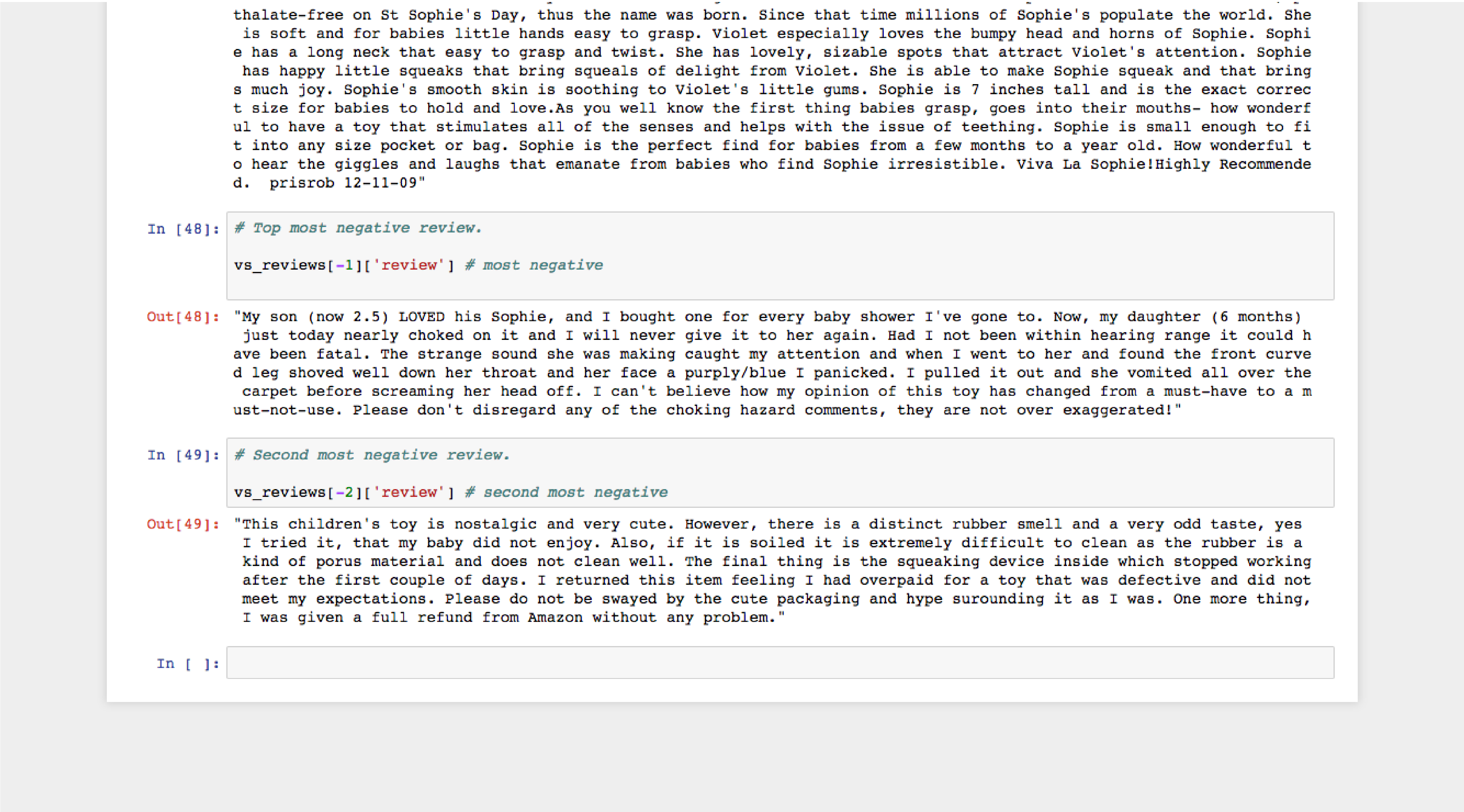
num_unique (est.):	718
num_undefined:	0
min:	2.222e-10
max:	1
median:	0.999
mean:	0.87
std:	0.311

distribution of values:

In [44]:
```
# let's find out most positive and most negative review.
# first let's short the data.

vs_reviews=vs_reviews.sort('predicted_sentiment_by_model', ascending=False)
```

In [46]:
```
# Top most positive review.

vs_reviews[0]['review']
```

Out[46]:
```
"Sophie, oh Sophie, your time has come. My granddaughter, Violet is 5 months old and starting to teeth. What joy litt
le Sophie brings to Violet. Sophie is made of a very pliable rubber that is sturdy but not tough. It is quite easy fo
r Violet to twist Sophie into unheard of positions to get Sophie into her mouth. The little nose and hooves fit perfe
ctly into small mouths, and the drooling has purpose. The paint on Sophie is food quality.Sophie was born in 1961 in
 France. The maker had wondered why there was nothing available for babies and made Sophie from the finest rubber, ph
```



```
thalate-free on St Sophie's Day, thus the name was born. Since that time millions of Sophie's populate the world. She
 is soft and for babies little hands easy to grasp. Violet especially loves the bumpy head and horns of Sophie. Sophi
e has a long neck that easy to grasp and twist. She has lovely, sizable spots that attract Violet's attention. Sophie
 has happy little squeaks that bring squeals of delight from Violet. She is able to make Sophie squeak and that bring
s much joy. Sophie's smooth skin is soothing to Violet's little gums. Sophie is 7 inches tall and is the exact correc
t size for babies to hold and love.As you well know the first thing babies grasp, goes into their mouths- how wonderf
ul to have a toy that stimulates all of the senses and helps with the issue of teething. Sophie is small enough to fi
t into any size pocket or bag. Sophie is the perfect find for babies from a few months to a year old. How wonderful t
o hear the giggles and laughs that emanate from babies who find Sophie irresistible. Viva La Sophie!Highly Recommende
d.  prisrob 12-11-09"
```

In [48]:

```
# Top most negative review.

vs_reviews[-1]['review'] # most negative
```

Out[48]:

```
"My son (now 2.5) LOVED his Sophie, and I bought one for every baby shower I've gone to. Now, my daughter (6 months)
 just today nearly choked on it and I will never give it to her again. Had I not been within hearing range it could h
ave been fatal. The strange sound she was making caught my attention and when I went to her and found the front curve
d leg shoved well down her throat and her face a purply/blue I panicked. I pulled it out and she vomited all over the
 carpet before screaming her head off. I can't believe how my opinion of this toy has changed from a must-have to a m
ust-not-use. Please don't disregard any of the choking hazard comments, they are not over exaggerated!"
```

In [49]:

```
# Second most negative review.

vs_reviews[-2]['review'] # second most negative
```

Out[49]:

```
"This children's toy is nostalgic and very cute. However, there is a distinct rubber smell and a very odd taste, yes
 I tried it, that my baby did not enjoy. Also, if it is soiled it is extremely difficult to clean as the rubber is a
 kind of porus material and does not clean well. The final thing is the squeaking device inside which stopped working
 after the first couple of days. I returned this item feeling I had overpaid for a toy that was defective and did not
 meet my expectations. Please do not be swayed by the cute packaging and hype surounding it as I was. One more thing,
 I was given a full refund from Amazon without any problem."
```

In [ ]: