# Analysis for Question3

2. Test your function with "amazon_review_300.csv" and a few reviews from this file.
   - Check the most similar review discovered for each of the selected reviews
   - Can you use the calculated similarity score to determine if two documents are similar?
   - Do you think this function can successfully find similar documents? Why does it work or not work?
   - If it does not work, what can you do to improve the search?
   - Write down your analysis along with some evidence or observations you have in a pdf file and submit this pdf file along with your code.

In order to analysis this function's performance on this "amazon_review_300.csv" documents, I need to apply this function on each document of data. For each row of the data, I computed the 'top_sim_index' and 'top_sim_score'. After that, I join this two columns to the origin data. The result DataFrame is like this:

Out[6]:

| | top_sim_index | top_sim_score |
|---|---|---|
| 0 | 2.0 | 0.196306 |
| 1 | 2.0 | 0.147997 |
| 2 | 3.0 | 0.223204 |
| 3 | 2.0 | 0.223204 |
| 4 | 2.0 | 0.200302 |
| ... | ... | ... |
| 295 | 299.0 | 0.433092 |
| 296 | 120.0 | 0.269875 |
| 297 | 291.0 | 0.101940 |
| 298 | 296.0 | 0.156068 |
| 299 | 295.0 | 0.433092 |

300 rows × 2 columns

The index is the origin index of these 300 reviews. The 'top_sim_index' and 'top_sim_score' columns are the corresponding to the most similar review. We will focus on the 'top_sim_score' column. Basing on the 'top_sim_score' , we will check on the document to see the performance of the find_similar_doc function.

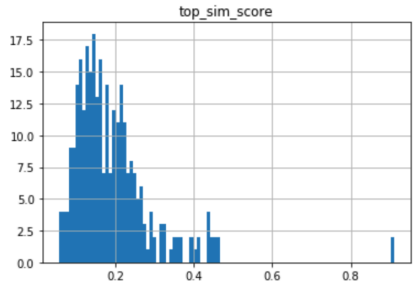
First we see the description and the histogram of the result:

```
In [7]: result.describe()
```

Out[7]:

| | top_sim_index | top_sim_score |
|---|---|---|
| count | 300.000000 | 300.000000 |
| mean | 153.280000 | 0.191250 |
| std | 91.399752 | 0.104174 |
| min | 1.000000 | 0.055308 |
| 25% | 69.000000 | 0.125883 |
| 50% | 156.000000 | 0.166524 |
| 75% | 232.500000 | 0.223430 |
| max | 299.000000 | 0.911463 |

```
In [12]: result.hist('top_sim_score',bins = 100)
Out[12]: array([[<matplotlib.axes._subplots.AxesSubplot object at 0x1168b6cc0>]],
       dtype=object)
```




We find that except one odd value 0.91, most value are in the interval [0.1,0.3]. We will test some representative examples in this data.

First we check the well-behave group, which has a higher top_sim_score (top5%) :

```
In [20]: sort = result.sort_values(by=['top_sim_score'],ascending = False)
In [23]: sort.head(15)
```

Out[23]:

| | top_sim_index | top_sim_score |
|---|---|---|
| 117 | 118.0 | 0.911463 |
| 118 | 117.0 | 0.911463 |
| 266 | 263.0 | 0.462068 |
| 263 | 266.0 | 0.462068 |
| 67 | 69.0 | 0.451566 |
| 69 | 67.0 | 0.451566 |
| 255 | 252.0 | 0.440894 |
| 252 | 255.0 | 0.440894 |
| 158 | 159.0 | 0.439085 |
| 159 | 158.0 | 0.439085 |
| 299 | 295.0 | 0.433092 |
| 295 | 299.0 | 0.433092 |
| 131 | 127.0 | 0.407373 |
| 127 | 131.0 | 0.407373 |
| 253 | 255.0 | 0.401466 |

Remind that these 15 pairs have repetition, which means like A is the most similar to B and also B is the most similar to A. So we check doc_id = [117, 266, 67, 255, 158, 299, 131]

The results are below:

```
Similarity between 117 and 118 is 0.91:

selected doc:  I did not know this was the audio cd of the game, I thought it could be used on the playstation. Maybe
the other guys are from the US or somewhere where it is actually easy to get the game,and play it,but this was all in
japanese and was not compatible with my machine. I am deeply disappointed as i love the arcade game, but the Dance st
age euromix (Dance Dance Revolution equivalent) is lacking in decent tracks, many are jungle-y and are difficult to d
ance to them. Also, most of these tracks are unheard of here and are therefore not very enjoyable. However, my favour
ite are Keep on moving, make a Jam, and Video Killed the Radio star. But it gets a bit pedantic if you keep playing o
n these tracks. If anyone is in the same situation as I am, or those of you who are lucky enough the play on the newe
st version, can you please let me know where I can order one? I am in dance deprivation/boredom!

similar doc:  I did not know this was the audio cd of the game, I thought it could be used on the playstation. Maybe
the other guys are from the US or somewhere where it is actually easy to get the game,and play it,but this was all in
japanese and was not compatible with my machine. I am deeply disappointed as i love the arcade game, but the Dance st
age euromix (Dance Dance Revolution equivalent) is lacking in decent tracks, many are jungle-y and are difficult to d
ance to them. Also, most of these tracks are unheard of here and are therefore not very enjoyable. If anyone is in th
e same situation as I am, or those of you who are lucky enough the play on the newest version, can you please let me
know where I can order one? I am in dance deprivation/boredom!
```

These two documents are almost the same, no wonder the score is 0.91. The selected doc only has two more sentences than the similar doc. So the function works pretty well in this case.

```
Similarity between 266 and 263 is 0.46:

selected doc:  I agree, these are cut small... Buy an inch or two larger than you normally do. The waist band does NO
T stretch.

similar doc:  I have bought dockers for years but these are cut smaller than they usually are. Even with the "stretch
" waist band they are tight. So much for an "extra" inch. Beware. Try these in a store before you buy.
```

These two sentences mean almost the same. Each express that the cut is small and the waist band is tight. These two customers have the same respond to the product . So the function works well in this case.

```
Similarity between 67 and 69 is 0.45:

selected doc:  My four year old daughter loves everything Barbie and loves the Rapunzel movie. This game is tons of f
un, even for a 42 year old. We love playing it together. We love decorating all the rooms and finding the gems. What
even better is, she can play it alone and I get some me time!

similar doc:  This is such a great game both my 3 year old son and 7 year old daughter love it. I like to play to if
they would let me! So much fun decorating the rooms and so many choices to keep you playing Great game. Not at all wh
at you would think a Barbie game would be. Great fun for all!!
```

These two sentences mean almost the same. Each express that this is a fantastic toy and the children all like it. Also the parent can attend the game with the children. The different is that the first customer like this toy because the children can play it alone so he or she can get some time,

while the second customer prefer to attend the children and play together. So the function works just well in this case, although a little bit bias.

```
Similarity between 255 and 252 is 0.44:

selected doc:  I bought this thinking it would be packed with Thomas stories. Turns out there are only ten, which is
a typical Thomas DVD. However, between each story, it also has interviews with parents and their kids about what they
like about Thomas. As I parent, I'm not interested in this, and I know that my 3 year old would rather watch Thomas t
han a bunch of strangers any day. Plus, the sing-along section has one song. One. We have other Thomas videos that ha
s at least 4, and they aren't special DVD's, just normal Thomas collections. I wouldn't recommend this to anyone, if
you are looking for solely Thomas stories and can find other collections.

similar doc:  This video is really a disappointment, it came with a Thomas piece so naturally my daughter picked it a
nd I thought the bonus "feature" wouldn't be mixed in with the actually Thomas stories but I was wrong. My daughter e
asily lost interest when the people would come on and talk about loving Thomas and getting to ride the real Thomas tr
ain... I was totally annoyed.
```

These two sentences have the same sentiment. Both are not satisfied with the product. Also they both said their children will lose interest on this product and this product basing onThomas stories . So the function works just well in this case.

```
Similarity between 158 and 159 is 0.44:

selected doc:  I didn't refer to Baseball America as such. It was Cardinals GM Walt Jocketty who said it. And who am
I to contradict? The best columnists in the game, from Peter Gammons, Tracy Ringolsby and Jayson Stark, to detailed i
nformation in ALL of baseball, from the Major Leagues to Japan, stats, scores, and the best and most extensive resour
ce on prospect information, there's not a single serious fan in all of baseball who deserves such a title if is not a
Baseball America subscriber. You got to read it if you want to know the stars of tomorrow (being tomorrow 6 years or
2 months). Top 10 prospects from every team till its massive Top 100 list, Baseball America covers it all. I know, I
have been subscribed for 2 years now, and I can't think of living without it!

similar doc:  I have read Baseball America off and on over the years and finally took the plunge and got a subscripti
on. This is the best place to get coverage of the minors, college, and high school baseball so you know about the top
prospects before others do.
```

These two sentences both praise the Baseball America is a good magazine and they both have been subscribed it for many years. So the function works well in this case.

```
Similarity between 299 and 295 is 0.43:

selected doc:  lil wayne and turk set the whole cd off the untamed killer seem to know how to flow. his beats are for
real and i cant wait for lil wayne to come out with his solo cd

similar doc:  most underrated rap group, when lil wayne actually talked real s*** besides aliens and skateboards..if
u like lil Wayne u should deffinetly love this album
```

These two sentences both praise this album from Lil Wayne. So the function works well in this case.

```
Similarity between 131 and 127 is 0.41:

selected doc:  My wife and I built a hybrid timber frame home using this book as a guide. Although we have constructi
on skills, this book was invaluable in guiding us through the process. We can't wait to build another. We highly reco
mmend this book to anyone considering a timber frame house.

similar doc:  My husband loves this book. He said he has learned so much and that this was the best book to learn how
to build a timber frame house by.
```

These two sentences both praise that this book is a good guide to build a timber frame house. So the function works well in this case.

These 7 examples are the evidence that the function can successfully find similar documents. It works because the tf_idf matrix can extract the main words from a document. And we use cosine distance to measure each document and find the nearest document to the target. The similarity score is 1- cosine_distance. So the higher the similarity score, the more similar of two document. The result indeed shows this point. These top 5% examples perform well under this function.

The reason why I take the top 5% similarity score as example is that these can show us that our function really work in some cases. And that is enough to prove the accessibility of the method. I think the point is to determine a threshold. When the similarity score above the threshold, this function works well. When the similarity score below the threshold, this function doesn't works well. It makes sense. Because this is only 300 documents. We cannot promise that each document can married with a good partner. Maybe some documents are just unique and can't match others.

In conclusion, we can use the calculated similarity score to determine if two documents are similar.This function can successfully find similar documents. When the similarity score is high enough, the two document are really similar.