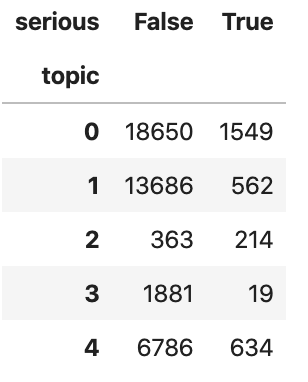
1. **Try to find topic identification that uses SVD to reduce the dimensions of the word/document matrix/table and report the topics identified.**  
     
   In Python, I found a library that performs Latent Semantic Analysis (LSA) in order to perform topic modeling. Behind the scenes, LSA actually uses SVD to identify the topics. To do so, SVD is used to create a lower dimensional approximation of the TF-IDF matrix, using the remaining tokens in the V matrix as the topics.  
     
   Here is the output using SVD to create 5 topics, showing the 5 most important terms:  
     
   A picture containing table

   Description automatically generated  
   We can cross tabulate the topics against serious, and run a Chi Square test to check if there is a non-random relationship between our topics and the ‘serious’ variable.  
     
   Running the chi square test, we get a p value of 7.423020764198414e-257, indicating that our topics are *not* independent of the serious variable.
2. **For topic identification, similar to Deliverable 1, try Latent Dirichlet Allocation. Find what you can, use as much of it as is relevant to this assignment, and write a brief report on what is available, what you did, and what you found.**Python has a decent module named ‘gensim’ that performs LDA. I tried running LDA on both the raw word counts, as well as on the TF-IDF matrix. Similar to LSA above, below are 5 topics for each case, showing the 5 most important terms.  
     
   LDA on TF:  
   A screenshot of a cell phone

   Description automatically generated  
     
   LDA on TF-IDF:  
   A screenshot of a cell phone

   Description automatically generated  
     
   There are a number of similarities between these two models, and the SVD model above. For instance, there tokens ‘report’, ‘shingrix’, and ‘medic’ appear in many different topic. Overall, these results seem poorer than the SVD results, in that there are a number of tokens that appear in multiple different topics; that is, the LDA topics appear less distinct than the SVD topics. The python LDA tool does provide a number of parameters a user can use to adjust the model.  
     
   For LDA, I also cross-tabulated the data against the serious variable, and checked the Chi Square test.  
     
   LDA on TF:  
   **A screenshot of a cell phone

   Description automatically generated**  
   The Chi square test on this crosstabulation is ~0, indicating we once again have a dependent relationship between topic and serious. This is clear, looking at topic 1, which has a huge proportion of serious (37%), and topic 0, which has a small proportion of serious (0.8%). The overall proportion of serious is 6.7%.  
     
   LDA on TF-IDF  
   A screenshot of a cell phone

   Description automatically generated  
   Again, running chi square on this cross-tab, there is a near 0 result, indicating the LDA tool did divide the population by the serious variable quite a bit. It is clear that the topics are not nearly as pure serious or non-serious.