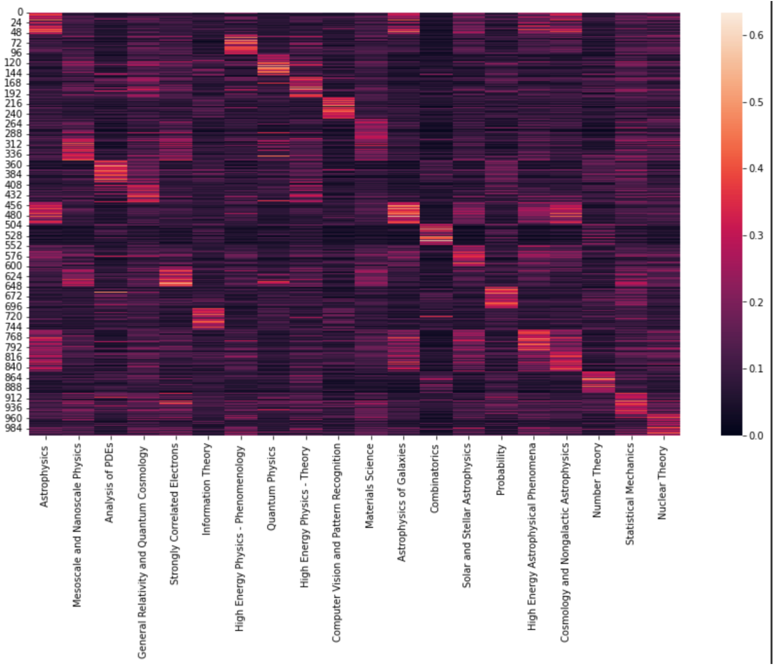
Task1:
From the bottom of the output.txt. We can see the final hit rate is 0.915, and there are some missed examples. Only 0.44 of the missed examples.
Here are three reasons I analyze for the example:

1. The relevance score of the hit and the missed example are quite close, the hit just loses the game and become the runner-up:
Looking deep into the case, for example, we can see the first missed case, query[0]: the top 3 documents and the corresponding score is quite close, only one or two words may influence the result.
[(485, 0.005079931858097432), (0, 0.004980543351988675), (818, 0.004533811094722729)]

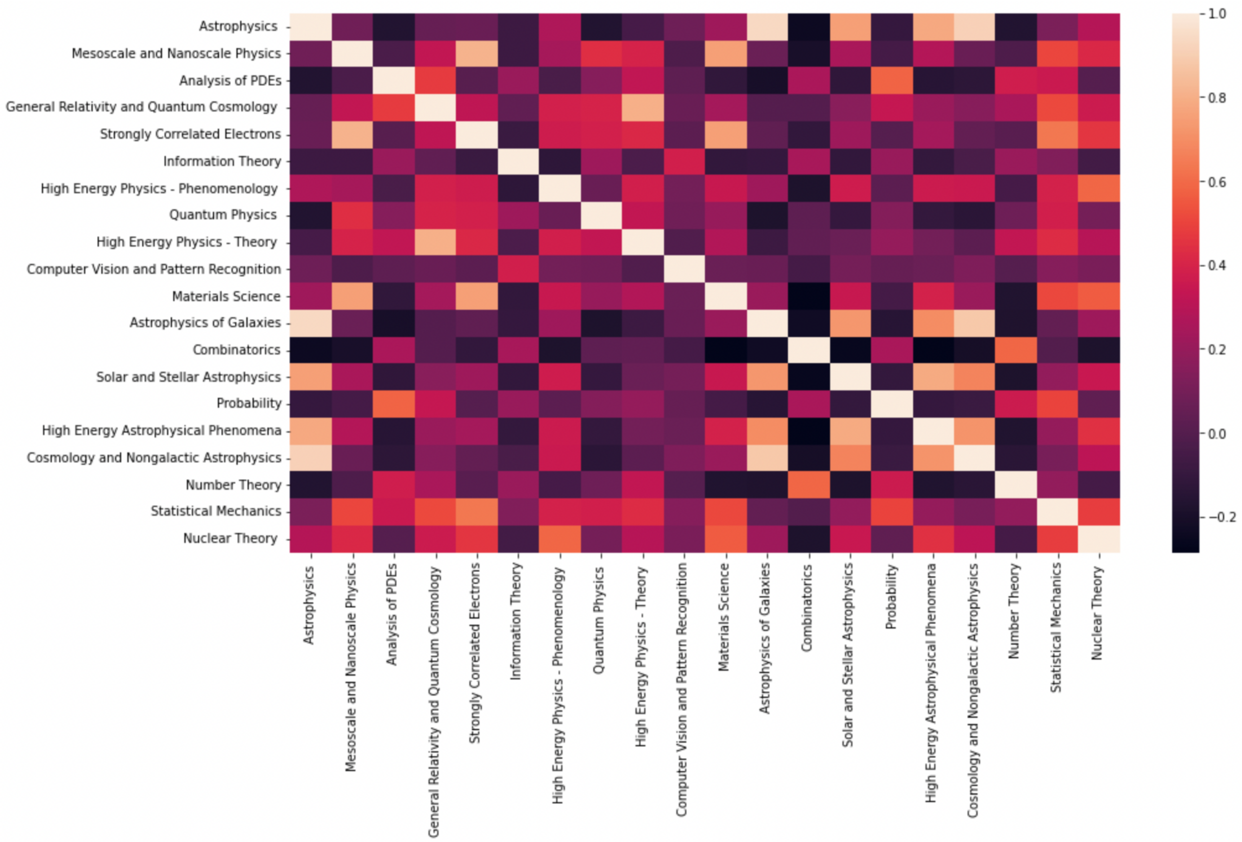
2. The characteristics of the title are not prominent enough, or the narrative content of the title has been studied in many articles.
Let us take query[22] for example, which title is “time resolve grb spectroscopy” after processing. From the title, we can hardly make sure what the article is going to talk about. Words like time, resolve are quite normal, and the title doesn’t consist too much words, thus, if the field or the topic is a popular one, it is hard to become hint.

3. Some articles with higher matching accuracy in the same field have long abstracts, which are likely to increase the TF value, while the abstracts of the hit itself are very short.
Let us take the missed example query[10] for example, in this case, all top3 examples have a quite long abstract, which significantly increases their score, and the abstract[10] itself, is even not founded in the top3 list.

Task2:
The heat map of categories and documents is shown below:



To study the correlation of the categories, I compute the correlation of every category and get the heat map of the correlation matrix as below:



The color of the corresponding cells represents the relevance of the two categories. The lighter the color, the higher the relevance. From the heat map, we can easily find some categories, such as (“Astrophysics of Galaxies”, ”Cosmology Nongalactic Astrophysics”, “Astropysics”), (“Strongly Correlated Electrons”, “Mesoscale and Nanoscale Physics”), have close connection, while some categories, like “Quantum Physics”, seems have few connection in the categories the sample given.

It is worth noting that, in the heat map, what is the most irrelevant is the color which corresponding value is around 0. The negative (deep color cells) tell us some information, like if the category is A then it is less likely to mention topics in category B. There could be a negative correlation.