CMPE493 ASSIGNMENT 3 REPORT

a) I used TF-IDF based cosine similarity approach while handling the genres. I followed the process mentioned bellowed respectively:

1. Term frequency of each genre is calculated.
2. Document frequency of each genre is calculated.
3. Inverse document frequency of each genre is calculated.
4. TF*IDF values are calculated.
5. Length normalization is applied to the TF*IDF values.

- At the end, I obtained the genres with the final scores. I used these final scores in the cosine similarity function.

b) α: To set a reasonable value to α, I calculated the average precision and precision values for each URL in the "book.txt". After calculating the precisions of a URL, I added these precisions to the precision values of the previous URL. After it iterates over all URLs in the "book.txt", I calculated the average of both precision values and kept these values in different list.

$$\frac{total_{avg_{precision}}}{number\ of\ urls}, \frac{total_{precision}}{number\ of\ urls}$$

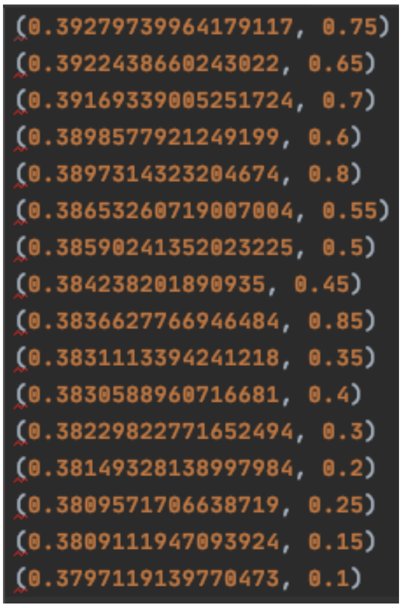
I incremented α by 0.05 and started over again to the same process for the updated α.

Basically:

1. I set α to 0.1.
2. I incremented α by 0.05 till it is equal to 0.9. → $[0.1, 0.9)$

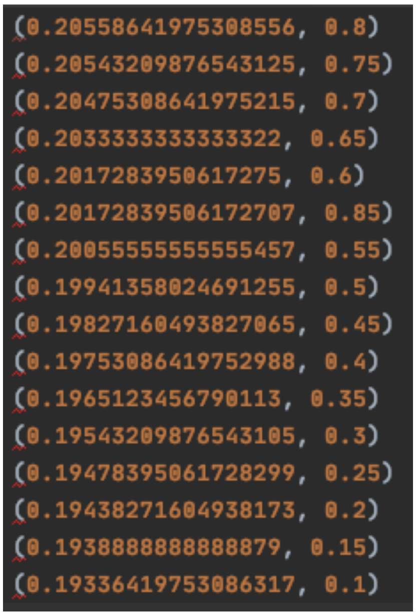
After the algorithm finished, I sorted the precision lists and output results in a "(precision, α)" format.

Average Precision Values



```
(0.39279739964179117, 0.75)
(0.3922438660243022, 0.65)
(0.39169339005251724, 0.7)
(0.3898577921249199, 0.6)
(0.3897314323204674, 0.8)
(0.38653260719007004, 0.55)
(0.38590241352023225, 0.5)
(0.384238201890935, 0.45)
(0.3836627766946484, 0.85)
(0.3831113394241218, 0.35)
(0.3830588960716681, 0.4)
(0.38229822771652494, 0.3)
(0.38149328138997984, 0.2)
(0.3809571706638719, 0.25)
(0.3809111947093924, 0.15)
(0.3797119139770473, 0.1)
```

Precision Values



```
(0.20558641975308556, 0.8)
(0.20543209876543125, 0.75)
(0.20475308641975215, 0.7)
(0.2033333333333322, 0.65)
(0.2017283950617275, 0.6)
(0.20172839506172707, 0.85)
(0.20055555555555457, 0.55)
(0.19941358024691255, 0.5)
(0.19827160493827065, 0.45)
(0.19753086419752988, 0.4)
(0.1965123456790113, 0.35)
(0.19543209876543105, 0.3)
(0.19478395061728299, 0.25)
(0.19438271604938173, 0.2)
(0.1938888888888879, 0.15)
(0.19336419753086317, 0.1)
```

I set α to 0.75, since it outputs the maximum average precision.

Notes:

- I did not use minimum/maximum thresholds, number of terms, weight variants etc. as a model parameter since I do not need them.
- Even though some of the URLs in the "book.txt" are broken, I extracted the information from URL by adding "/en/" at the end of "goodreads.com".
- I used parallelism to enhance the performance. While creating the model, 15 threads is extracting information from different URLs at the same time.

Sertay AKPINAR