Problem Statement

like to build towards a system that, as stories are uploaded, is able to flag them if they are suspected to contain copyrighted content.

Assume you are given a finite set of text files (in a directory), each of which contains the complete text for one copyrighted story:

~/copyrighted\_works:

harry\_potter\_and\_the\_philosophers\_stone.txt

the\_fellowship\_of\_the\_ring.txt

twilight.txt

...

Design a program that, as a new story is uploaded by a user:

user\_input\_1.txt

...is able to generate some sort of similarity scores between the user upload and the predefined list of copyrighted works. These scores may then be used to flag stories for manual review.

We will call your program via a bash terminal (on macOS) as follows:

$: ./your\_program -c ~/copyrighted\_works -u ~/user\_input\_1.txt

...and expect output to be printed to stdout, in order of most to least similar copyrighted work:

harry\_potter\_and\_the\_philosophers\_stone.txt : 0.5

twilight.txt : 0.2

the\_fellowship\_of\_the\_ring.txt : 0.1

...

We will refer to an optionally provided README.txt for instructions regarding the running of your program (e.g. any required dependencies).

**Please document any assumptions or tradeoffs that you are making** with respect to both performance and output accuracy**.** This exercise is intentionally open ended so that you may build the program in a platform or language of your choosing. We encourage creative solutions to the problem at hand but aren’t necessarily looking for a fully complete and robust approach. We will use your solution as a starting off point for our discussions during the next stage of the interview process.