Final Project Report 3: Application of TF-IDF Weighting

As this part of code was written as per instruction, the adaptation of code examples basically contains two parts: first, modifying the week 8 assignment to load the text; second, integrating applicable portions of the provided code samples. For the first part, I improved the way of loading text by employing BeautifulSoup instead of my previous choice of Regex, which made up for my previous unfamiliarity with the use case of the former. The remainder of the first part appears to be similar to the week 8 assignment, with a combined filter eliminating the three categories of words as detailed in 8.1slides.

My familiarity with Pandas eased the work with the second part (if my understanding and coding were correct) — just to configure the vectorizer as detailed in the last several pages of 9.2 slides, with the consideration for setting both the maximum and minimum range of the parameter. On top of that, I established two directories to organize the outputs for the subsequent two questions, and write their respective files later under each loop. Another notable adjustment was the use of .iloc() instead of .loc() to facilitate the retrieval of the prescribed range based on position (rather than on label), as the first is a redundant comparison with a work itself when identifying the top similarities. They were lastly written in the file with only the title and weighting, a practice similar to an assignment in previous week.

This adaptation of using TF-IDF vectorizer indeed produced intriguing insights, particularly when juxtaposed with findings from PhiloLogic. The list of top 20 words serves as a quick tool to grasp the essence and backdrop of a literary piece, regardless of one's prior knowledge of the text. For instance, the frequent appearance of words like “samurai” and “daimyo” in stories set in medieval times hints at the historical setting and the social hierarchy of the characters often depicted in these narratives. Additionally, terms associated with the erotic grotesque genre—such as “goblin,” “fairy,” “legend,” “spectacle,” “nude,” and “fantastic”—alongside religious terms like “sutra,” “chant,” “hermit,” and “monk,” are also prevalent, offering clues to the work’s theme, thus serving as a useful complement to PhiloLogic.

The analysis of the top 3 similar works, with similarity scores ranging roughly from 0.13 to 0.39, indicates a well-balanced comparison. Notably, three twentieth-century fictions emerged as mutually similar, each with similarities exceeding 0.3, suggesting a commonality not just in their modern settings but also in narrative style (e.g., modern erotic grotesque can be evoked by topics like overdose and groupsex, which are not common in medieval contexts). This observation also tentatively supports my hypothesis regarding the adaptation of medieval themes in modern renditions, indicating underlying connections that merit deeper investigation. Moreover, the analysis can reveal an author’s distinctive writing style, as seen in cases where one author’s work is similar to another of himself. TF-IDF analysis, overall, validates some preliminary assumptions, however, more advanced examination is essential to uncover specific examples.