Assignment 3

(i) Describe the model you used to encode the genres of the books

- All variables which is provide the values necessary to perform parse genre field is in utils.py.
- The 'find_str' function returns the starting index or the last index with the control parameter, which is the boolean expression, whether there is a string sent as 'char' in the 's' parameter and the its index.
- The 'find_genres' function finds all genre name with html tag, then with use of 'clean_html' function, all genre strings are cleared from html tag.
- I use tf_idf model to encode the genres of the books.
- The steps behind the tf_idf model are find tf value(for all genre word in one url data) firstly, then find idf value(for all genre word), lastly find tf_idf value by multiplying tf and idf value.

(ii) Describe the model parameters (minimum/maximum thresholds, number of terms, weight variants, α, etc.)

- $tf_{t,d}$ -> t means description word or genre word, d -> single url description or genres
- $w_{t,d}$ -> log frequency weighting
- df_t -> number of url description contain description word or number of url genre contain description genre
- tf-idf weighting is the product of its tf weight and its idf weight.
- I use α value to find cos_similarity. Purpose behind use α is to combine tf_idf value getting from description data and genres data.
- To select α, I start a with 0 and each step increase a by 0.025. I compare cos_similarity, then I find highest cos_similarity at α 0.50 value.