Junior Data Scientist Challenge

Problem Statement

For every book we sell, we create an accompanying record in our database of product metadata. This metadata includes features like the title, author name, and physical dimensions of the book, but also descriptive information like the blurb and genre. All of this data is sent to online retailers like Amazon, who then index it for search - so that if a reader searches for the book's title, or genre, it should (hopefully!) appear in search results.

One of the most important metadata fields for search and discoverability on online platforms is the list of **plaintext keywords** we assign to each book. These keywords should describe aspects of each book's content - so, for a thriller set in ancient Rome, we might add the keyword *ancient roman thriller*, or *historical thriller*, with the idea being that when readers search for these terms, our book will appear in search results.

When assigning keywords to books, as well as using human expertise, we want a system that automatically suggests known user search queries for each of our books, so that we can add keywords which we know people are actually searching for, hopefully boosting our books' discoverability.

Your Task

Your task is to build a prototype system that assigns keywords to our books. Included with this task are two CSV files:

- A list of Book IDs and their metadata for a subset of our books
- A list of user search queries

Using the list of search queries and the book metadata, build a model that picks the most relevant search queries for each book, to be used as keywords. The search queries assigned to each book should be relevant to the themes and genres of the book as described in the metadata.

Guidelines:

- You must use Python; we want to see your coding skills on display!
- Bear in mind that this is an NLP task - we also want to see how you handle text data
- You are free to use whatever open source libraries and tools you like

Create your output in the form of a CSV file, with the following columns:

ID: Unique identifier of each book (from title_metadata.csv)
search_term: A user search query picked by your model for this book.
title: Name of the book (from title_metadata.csv)
description: Description of the book that appears on Amazon (from title_metadata.csv)
thema_codes: Codes which identify the genres of the book (from title_metadata.csv)
thema_description: String description of each THEMA Code (from title_metadata.csv)

An example with dummy data can be seen below:

ID	title	search_term	description	thema_codes	thema_description
1001	personal mba, the	personal finance books	This is the book description	KJP, KJMB, KJU, VSPM, VSI	['Business communication a
1001	personal mba, the	entrepreneur basics	This is the book description	KJP, KJMB, KJU, VSPM, VSI	['Business communication a
1001	personal mba, the	careering	This is the book description	KJP, KJMB, KJU, VSPM, VSI	['Business communication a
1001	personal mba, the	careers	This is the book description	KJP, KJMB, KJU, VSPM, VSI	['Business communication a
1001	personal mba, the	entrepreneurship	This is the book description	KJP, KJMB, KJU, VSPM, VSI	['Business communication a
1002	secret history, the	mystery thriller books	This is the book description	FF, FH, FBA, FFL	['Crime and mystery fiction
1002	secret history, the	historical crime fiction	This is the book description	FF, FH, FBA, FFL	['Crime and mystery fiction
1002	secret history, the	murder mystery books	This is the book description	FF, FH, FBA, FFL	['Crime and mystery fiction
1002	secret history, the	crime thriller books	This is the book description	FF, FH, FBA, FFL	['Crime and mystery fiction
1002	secret history, the	suspense thriller books	This is the book description	FF, FH, FBA, FFL	['Crime and mystery fiction
1002	secret history, the	detective novels	This is the book description	FF, FH, FBA, FFL	['Crime and mystery fiction

In the next interview we will ask you to

- Take us through your solution
- Explain the different approaches you considered
- Show us examples of where your solution worked well
- Show us how you think your solution could be improved.

You will be scored on

- The reasons why you chose your approach
- Your ability to effectively communicate your solution
- The quality of your code
- The areas you identified for improvement

Good luck!

Appendix

The **title_metadata.csv** file contains the following columns:
ID: Unique identifier of each book
title: Name of the book
description: Description of the book
thema_codes: THEMA Codes which identify the genres of the book
thema_descriptions: String description of each THEMA Code
fiction_flag: Flag which indicates whether the book is Fiction or Non-Fiction

The **search_terms.csv** file contains the following columns:
search_term: a user search query