A. Hiring Assessment: Text Normalization

The dataset contains pairs of raw and normalized text like the below example:
RAW TEXT: **Mike Hoyer/JERRY CHESNUT/SONY/ATV MUSIC PUBLISHING (UK) LIMITED**
Normalized Text: **JERRY CHESNUT/Mike Hoyer**

We need to normalize text containing composer and writer information. This includes removing redundant data such as publisher names, irrelevant placeholders (e.g., <Unknown>), and special characters, while preserving and formatting writer names consistently. The main peculiarity of the data is that the names may be in different orders (e.g. “Lastname, Firstname”, “Firstname Lastname” etc). Obviously there are also redundant words and special characters which need to be removed except for the character ‘/’ which needs to separate the different names.

Basic preprocessing

There are several basic steps for initial preprocessing which include:

- Tokenization
- Convert to Lower case
- Removing stop words (“the”, “a”, “with”, “I”, etc)
- Removing special characters and punctuation (except for “/”)

This can be done with nltk or spaCy libraries which are used for natural language processing in python.

NER models

After that, we are going to use a pretrained NER (named entity recognition) model that can identify names among a set of words. Specifically, we can use the pre-trained bert-base-NER model from HuggingFace which can effectively recognize locations (LOC), organizations (ORG), persons (PER) and Miscellaneous (MISC) and keep all tokens that correspond to persons. Additionally, spaCy provides pre-trained NER models that can identify names and other entities in text as well. We could compare these two approaches or combine them by using them in order, though the BERT based model will be probably more effective.

For better performance we can also fine-tune the NER model on custom data, but this would require a way to annotate the dataset first, so the model can learn which of the new tokens refer to words.

Final formatting & Evaluation

After the names are extracted we can use again the spaCy library to format them in the desired way. This includes capitalization, adding whitespaces and ensuring there are slashes used as separators between different names.

For evaluating the model we could use the following metrics:

- Precision: Percentage of correctly identified names out of all extracted components.
- Recall: Percentage of true writer names extracted from the raw text.
- F1-Score: Harmonic mean of precision and recall.