====================================

**Updates on 4/27/22**

Alignment results:

Tried 3 different approaches for Scene-Chapter alignment:

* Cosine similarity between scene-chapter embeddings using SBERT:
  + Represented each scene and chapter with an aggregated embedding using dialogues’ embedding from SBERT.
  + Time to run: 3s
  + Performance does not look good at all on a manual assessment (Reason could be too much padding)
  + [Alignment\_scene\_chapter\_aggregated\_embedding.txt](https://drive.google.com/file/d/1Ob4zOcprpihZtXkhud8tAJdm8f92blBy/view?usp=sharing)
    - Just a list representing the first element as the scene number and the second as the chapter number.
* Using Hamming’s Edit Distance for all dialogues in script vs all dialogues in book:
  + Time to run: 33s (Each computation takes about 0.176781s)
  + Performance looks much better than the aggregated SBERT approach
  + [Alignment\_hamming.csv](https://drive.google.com/file/d/1TTpLnbzQz9td9ZM5GI88kyC8ZmlavFH5/view?usp=sharing)
    - The row after each scene’s dialogue represents the final aligned chapter with the scene (Scene column and Chapter column)
    - The other rows represent each dialogue from a scene and the most matched dialogue from the chapters, as well as the dialogue itself.
    - Alignment decision based on the most frequently matched chapter.
* Using Levenshtein’s Edit distance for all dialogues in script vs all dialogues in book:
  + Time to run: 1hr 48m (Each computation takes about 8.63674s)
  + Performance seems to be quite good on a manual assessment
  + [Alignment\_levenshtein.csv](https://drive.google.com/file/d/1ZJDlMfo_4OuE5lkrQH7n-0KrHp2nE9Xo/view?usp=sharing)
    - Same structure as the previous one.

====================================

**Updates on 4/25/22**

Alignment results ([scene-chapter.csv here](https://drive.google.com/drive/u/0/folders/1QAMQclozUNDlUWneFW1qKOAwMToB8VEN)):

* 989 dialogues in scenes, 982 in chapters
* Running time < 4s, not time-consuming
* 74 out of 78 scenes have dialogues
* Among 989 dialogues in scenes: average similarity score = 79.77%, 45 dialogues with exact same text in book (score=1), 554 with >80% of similarity score
* Among 74 scenes: average vote rate for chapter = 76.93%, 21 scenes with 100% and 24 scenes with over 90% rate

====================================

**Updates on 4/19**

Next step: using dialogue as anchor to determine chapter-scene alignment, expected to be finished in 1 week

1. Script: define start of the scene by ‘S: EXT’ or ‘S: INT’ in ‘Zootopic\_parsed.txt’
2. Extract dialogues from script by ‘D:’ and from book by quotation mark
3. For ~80 scenes and ~30 chapters, derive two methods for alignment
   1. convert dialogues into embedding vectors by SBERT, compute cosine similarity for all possible pairs of dialogues (989\*982). For each chapter, choose the majority of the scene with the highest score
   2. all dialogues are aggregated and converted into 1 embedding vector by SBERT for each chapter in book and each scene in script, compute similarity for all chapter-scene combinations (80\*30). For each chapter, choose the scene with the highest score
4. Compare time and performance

Added some papers of Bamman in writeup which seems related to literary, book and annotations. Will read it to check

====================================

**Updates on 4/16**

Reviewed 6 tasks and corresponding papers from github, had a [summary](https://docs.google.com/document/d/11K4fx86WZbxQiLShbVYgNF881K1PLpwgea0UZOh328Q/edit) of annotation rules, dataset, model and experiments

Will look at google scholar

====================================

**Updates on 4/8**

Split papers and tasks with my team mate, Survey based on github is expected to be done within 1-2 weeks

Meeting minutes (4/8) with Professor and PhD student from another university:

1. Survey existing works of Bamman about book annotations, first in BookNLP github (contains most), then check anything missing/new from google scholar
   1. Read papers referenced in each task in booknlp
   2. Have a table for comparison of datasets based on statistical summary
2. Align book with script. The scene order in script basically is same with book
   1. First simply match exact word from sentences in book to the dialogue/scene
   2. Use SentenceBert for the rest
   3. Summary of chapter which is used to reduce complexity might be inaccurate and inefficient (work from OpenAI can be barely replicated). We first try all combinations in zootopia and see the performance, then decide if the chapter matching is needed and how to implement it

Things confirmed

1. Since ‘Scene Segmentation’ has not been done before, our dataset will be the first of its kind. How will this comparison (no. of books, tokens) help us prove the quality of our dataset?

First, compared with other dataset we could show that our segmentation in dataset is to split meaningful unit (scene) which is not fully explored by bamman. Second is for application. Third, we could implement experiments (human study or other kinds of exp) on our generated dataset and similar work from bamman for comparison.

1. How will the summary of supported tasks help us? (show our task is the first one?)

FIrst is for related work section, check their limitations. Second is to see if we could be inspired by details in their implementation.

1. How far back are we going to check for datasets/tasks? (recent 5-7 years?)

First github, then google scholar