Annotation Summary and Insights

We started by using Spacy NER software to identify named entities within our texts, and then manually reviewed these identifications. During this manual check, we noticed that the software missed several personal names, especially ones that are also common English words, like "Will" and "Young."

Key Observations

- **Challenges in Manual Annotation:** Manually annotating texts showed us how detail-oriented this work needs to be. Even small mistakes can lead to big inaccuracies in how entities are connected across a text. This process really highlighted how complex language is and the careful attention needed to accurately identify and link entities.
- **Handling Texts with Many Names:** Texts with a lot of names were particularly challenging. The difficulty wasn't just the number of names but also keeping our annotations consistent throughout the document. Interestingly, texts with many repeated names were easier to manage because the repetition helped us keep track of who was who. This shows that the content and structure of a text can greatly affect how difficult it is to annotate.
- **Text Length and Error Rate:** We found that longer texts tend to have more errors in annotation. The longest text we worked on, text 4, which also had the most names, had the highest error rate. This suggests that the longer a text is, the more complex and error-prone the annotation process becomes, highlighting the need for effective tools to assist with manual annotation.
- **Accuracy Across Different Texts:** The accuracy of our entity identification varied a lot between texts. The last text we looked at, number 5, was particularly challenging, with many non-person entities incorrectly identified as people. This pointed out some limitations in our model's ability to differentiate between types of entities, especially in texts where the context doesn't make this clear. It shows how important the context of a text is for Named Entity Recognition (NER) and that some texts are more challenging than others due to their specific content and structure.