RQ 4.1: In Problem 3, you did an error analysis on the task of recognizing text spans mentioning named entities. Pick one type of error that you observed. How could you improve the model's performance on this type of error? What resources (such as domain knowledge, data, compute, ...) would you need to implement this improvement?

In this problem, I recognized the FP errors dominate, which means the model predicts a lot of spans that do not exist in the gold-standard.

In order to improve the performance, we have to remove the predicted mention spans that are less likely to occur in English Wikipedia. This way, the false positive errors would be reduced.

From the model’s documentation, I found that under ”en_core_web_sm”, “NER”, there are some labels for entities. As long as I can remove entities with some labels that would less likely to appear in Wikipedia, it can work.

RQ 4.2: What does the word “context” refer to in the context of Problem 6? How does this help to disambiguate between different entities? Suggest another type of context that you could use for disambiguation.

As is mentioned in the question, the context refers to the previous and later 5 tokens around the entity. We trained a multinomial Bayes model to predict the entity according to different context, given a specific mention span. In this way, when facing a new mention span, we can put it into to the corresponsing classifier to give the appropriate prediction according to the given context.

For the last question, I did not find the appropriate another type of context.

RQ 4.3: One type of entity mentions that we did not cover explicitly in this lab are pronouns. As an example, consider the following sentence pair:

Ruth Bader Ginsburg was an American jurist. She served as an associate justice of the Supreme Court from 1993 until her death in 2020.

What facts would you want to extract from this sentence pair? How do pronouns make fact extraction hard?

The fact is Ruth Bader worked as an associate justice of Supreme Court from 1993 to 2020.

The reason for the increased difficulty is it is hard to combine the two sentences together by the name “Ruth Bader” and the pronoun “She”.