**Improve Optical Character Recognition of Periodicals Database**

Two types of articles

High quality transcriptions (>50% of words transcribed)

Low quality transcriptions ( <50% of words are in English dictionary)

Focusing on the first type of article, can you write a spell corrector program that fixes the spelling errors in articles?

Ex. [text-davinci-002](https://beta.openai.com/playground)

Focusing on the second type of article, what are systematic errors made during transcription

Ex. Double “ss” is turned into “ff”

Identify these errors and fix them

Two other sub projects include:

Topic modeling - treating every 1000 words in an article as a topic

SQL for analysis - creating tables from the database

Approaches to the OCR problem

The first few words of every article is the article title

*Could this be leveraged to train an error correcting program?*

There are advanced OCR models that can use inference to transcribe a pdf

*Use OCR model on 5,000 articles and*