In response to feedback:

1. Data (2.5 points) Student talks about how they will access the data (1.5 points) Data set is clearly defined (1 point)

0 / 2.5

You indicate: "My source of data is the web. I will scrape about 40 journal level articles to start building my summarizer. "

So you talk generally about "how" but not in the specifics necessary, and the data set is not clearly defined. Please describe which journals (many are not open access) and how you will retrieve the data. What is the form of the data? etc. If this part is not clear it may be very difficult to get the project started.

**Here is where I will find my data:**

\

**COVID-19 Open Research Dataset (CORD-19)**

**DATASET BY**[**KELLY GARRETT**](https://data.world/kgarrett)

Comment

257

**Explore this dataset**

* [Overview](https://data.world/kgarrett/covid-19-open-research-dataset)
* [Contributors](https://data.world/kgarrett/covid-19-open-research-dataset)
* [Discussion](https://data.world/kgarrett/covid-19-open-research-dataset)
* [Activity](https://data.world/kgarrett/covid-19-open-research-dataset)

**Overview**

**DESCRIPTION**

**Corpus of 29,000 scholarly articles about COVID-19**, use natural language processing to generate new insights

Manu, if not all of these articles are in PDF.

There is a Python PDF reader that I will use to load the data

Project goals are clearly articulated and achievable (1 point)

0/1

The biggest challenge highlighted is the biggest weakness. You outline approaches that could be taken to evaluate the summarizer but do not indicate which you will use. This is very difficult because you do not have a target/ gold data to evaluate against. I suggest reframing the problem where you can exploit supervised learning a target variable.

I will use ROGUE-1, ROGUE-2, ROGUE-L and Language Diversity to evaluate my summarizer.