Create a parser to extract data from the following link:

<https://www.ndtv.com/india-news/details-of-dissent-letter-to-sonia-gandhi-steady-decline-no-honest-inspection-2286399>

Data extraction should be able to eliminate noise to a reasonable extent. To eliminate noise, you need to deal with a few situations like:

1. Text from ads coming b/w news text
2. Text from links to other pages coming b/w news text

**Hint** : Can we build/use a model that can predict whether an intermediate sentence is in accordance with the content of the news article ?

And finally, take a URL as an input and compare if this input URL has "**similar**" content as the one on top. You can be creative with how you want to define similarity. We are looking for well thought out ways to understand content similarity.

***Note : Don’t try to make a universal parser, we ideally want something that works for a few websites (at least 3) of your choice ( You can state that in your submission.)***

Your final submission should be well modularised so that we can just pass a url and check the similarity. Please document your submission well and include a README.md where you explain how to run the project.

The last date for submission would be **Monday, 9th of November.** Any submission made post this would not be accepted for evaluation.