Setup Instructions & How to run codes

###Please follow the instructions below to set up the environment

python version : python3.6.9

##########################################################################

## Run the following commands in terminal to install required libraries.

python3 -m pip install numpy==1.19.0

python3 -m pip install pandas==1.0.5

python3 -m pip install scikit-learn==0.23.1

python3 -m pip install joblib==0.15.1

python3 -m pip install matplotlib==3.2.2

python3 -m pip install nltk==3.5

python3 -m pip install beautifulsoup4==4.9.1

python3 -m pip install spacy==2.3.2

python3 -m pip install scispacy==0.2.5

python3 -m pip install tensorflow==2.2.0

## Run the following commands in terminal to download spacy models

python3 -m spacy download en\_core\_web\_sm

python3 -m pip install https://s3-us-west-2.amazonaws.com/ai2-s2-scispacy/releases/v0.2.5/en\_core\_sci\_sm-0.2.5.tar.gz

#########################################################################

## Run the following commands in python session to download NLTK models

# specify the directory (inside quotes) for downloading NLTK models. You have to specify the folder name later in config.py

nltk\_dir = ''

# download the models

import nltk

nltk.download('punkt',download\_dir = nltk\_dir)

nltk.download('stopwords',download\_dir = nltk\_dir)

nltk.download('averaged\_perceptron\_tagger',download\_dir = nltk\_dir)

nltk.download('universal\_tagset',download\_dir = nltk\_dir)

########################################################################

## How to run the codes?

* Keep all the python scripts (8 files) in one folder.
  + (clean\_text.py,feature\_create.py,rules\_final.py,classification.py,config.py,main.py,rewrite\_html.py,x\_prep.py)
  + You need to launch python from the same folder.
* Keep all the raw html files in one folder. Please make sure that only html files are present in the folder. No duplicate names for files.
  + You will have to modify the final section of x\_prep.py in case the input html files follow any other structure (like each file in separate folders etc...)
  + Reduce the number of files and run again if the code breaks in between due to memory issue, or it takes lot of time.
    - The processing took <3 hours for 416 files on a system with 16 core CPU and 32 GB RAM(just to give an idea).
* Keep all the model objects etc.. from mathlogic in one folder(model\_objects\_dir in config file).
* Keep the nltk models in one folder(nltk\_dir in config file).
* config.py is the only file you need to change. The required details are given in the config.py file itself.
* main.py is the file to be executed. All the other files will be called internally. It can be run interactively or in nohup.
  + You can run the code in nohup using command given below:
    - nohup python3 -u main.py &
* The intermediate files will be saved in 'temp\_dir' specified in the config file. You can delete them once the final output is ready.
  + Tip: If you create a separate temp directory, you can delete the folder altogether.
* You can find the files with bodymatter related errors in null\_bodymatter.csv in ‘temp\_dir’. These files will not be processed.
* The corrected html files are written to 'rewrite\_dir' specified in the config file.