Tiffany Loe
ECE 491B
Assignment 2

2.1

a.

Looking at the first webpage, the URL is: http://000-084.smartcode.com/ which is currently inaccessible. When looking at the raw HTML, the webpage seems to be about study resources for different exams. One that appears frequently is the IBM 000-084 exam.

b.

The WET file contains some repeated text like “Details - Download - Screenshot” that is reminiscent of the HTML file, so these parts of the file should have been filtered out by the extractor. These repeated parts of the text may produce noise if a model is trained using this data which may lead the model to learning patterns that don’t truly represent the content of the webpage. Useful information the model can extract from this data involves the study resources for various different exams.

c.

This WET file might be useful for a model that is meant to help students study for an exam especially when the model is required to navigate through various links. This WET file would not be useful for a large language model where an abundance of plain text is needed.

d.

Record:

1. English, Testkingworld.com, and it is a page about study resources.
2. Model does not exist
3. Chinese, unknown, and it is a page about compasses
4. Italian, unknown, and it is a page about treating nail fungus
5. Italian, unknown, and it is also a page about treating nail fungus

6. English, 01webdirectory.com, and it is a page about a catering business that makes cakes and other baked goods
7. Dutch, www.e-spots.nl, and it is an ad for an office space.
8. Russian, unknown, and it is a page about someone asking for advice on how to treat fungus
9. Dutch, www.e-spots.nl, and it is an ad looking for a beauty specialist/make-up artist
10. Chinese, unknown, and it is a NSFW page
11. Chinese, unknown, and it is a page marketing medicine
12. Chinese, unknown, and it is a page that lists gaming websites
13. English, YouDot.io, and it is a page that has not been fully set up yet
14. Japanese, unknown, and it is a page about finding someone's phone number or address
15. Chinese, www.565.net, and it is a page about a bridal company
16. Chinese, unknown, and it is an optoelectronics supply store
17. Chinese, unknown, and it is a slot machine website
18. Chinese, unknown, and it is a website selling storage boxes
19. Chinese, unknown, and it is a website listing products and businesses
20. Chinese, Zhongte.com, and it is a website about online gambling
21. French, unknown, and it is a website about model planes
22. No information listed, just an image which does not show
23. Russian, 100-porno.wtfdexter.com, and it is a NSFW website
24. French, unknown, and it is a website that contains all of the French rhymes
25. French, http://darken04.skyblog.com, and it is a website about polls

2.2

b.

The text extracted from the WARC file is almost identical to the WET record that was extracted. The extracted text from the WARC file has bullet points in some sections which make the text more readable. The WET record on the other hand has the text on new lines rather than bullet pointed.

2.3

b.

Problems could arise in the language identification procedure if the text files within the dataset are not labeled with the correct language. If these are incorrect, the model will be trained using the wrong information, and therefore the results from the model would be incorrect as well.

c.

From 20 extracted files, the script was able to correctly identify all of the languages, except the files that did not have any information in them like record 2 and record 22 from the previous problem. Out of the 20 records analyzed, 15% were in english. I would say a suitable confidence value is around 0.75 since all of the files past that confidence threshold were correctly identified. Any file below the 0.75 threshold may not correctly identify the document's language and should be filtered out before training the model.

2.4

d.

Problems that may arise down the road if the phone number, email, and ip address masking functions are naively applied to the training set is if these functions do not properly mask the phone numbers, emails, or ip addresses. These issues might be mitigated by applying the masking functions to some parts of the dataset and checking the output to make sure the mask is being applied correctly.

e.

After looking through 20 files I noticed that the function for masking phone numbers created the most false positives and false negatives. The other functions for masking the emails and IP addresses had a few false positives but not false negatives.

2.5

c.

Problems that might arise downstream in a language model when NSFW and toxic speech filters are applied to the training set are overfiltering and potential bias. Some texts may be unnecessarily classified especially if it is an educational site discussing the NSFW and toxic content which may bias the model. While it is important to filter true NSFW and toxic speech content, if a particular language has more NSFW and toxic content in comparison to other languages, the data could be filtered out disproportionately causing the model to lack skills in the filtered out language.

d.

In the 20 files, the toxic speech filter was able to correctly identify records that contained toxic content and did not result in any false positives or negatives. On the other hand, the NSFW filter had a few false negatives but no false positives. After manually looking through the records, there were 2 out of 20 that contained toxic speech and NSFW content, but only the toxic speech filter correctly identified these files.

2.6

b.

The quality filter did a decent job determining high quality content that matched the criteria, but it failed to include all of the records that should have fallen under the high quality content label. Some text examples had poor quality content at the very beginning of the page, but legible content was towards the bottom of the page. I think the filter might not look far enough into the bulk of the content before determining its judgement.