CS 336: Language Modelling from Scratch

Assignment 4

May 22, 2024

SUNet ID: mattreed
Name: Matt Reed

1 look_at_cc

(a) The first page in this sample from CC is from the host "000-084.smartcode.com," which is no longer active. This home page likely has a bunch of links to download software packages for free.

(b) The WET content contains lists like "Home, Categories, New, Popular, Submit, RSS, Contact, false" which likely reflect different tabs of the webpage's header and aren't particularly relevant for a language model trying to learn natural language. In that respect, random HTML info might be considered lower quality data than books or articles. A model that is not sufficiently complex to learn the difference between natural and non-natural text will take this text which includes random numbers, tags, and lists and will start generating content that looks like random HTML-adjacent text. With a LLM, it might be helpful to learn all kinds of data (including HTML), but with a smaller model (and especially with data that isn't HTML but merely HTML aliasing as natural text), this is definitely lower quality data than we'd prefer.

(c) A training example like this could be helpful for an LM trying to learn about the structure of websites, but it would not be helpful for an LM trying to learn natural language.

(d) Got it. Here is the expanded table with all languages set to English:

Number	Language	Domain	Type
1	English	000-084.smartcode.com	Home Page
2	English	http://00000000sexbom.quickloansforbadcredit.org/...	Porn
3	Madarin?	http://00000028.cn/news.asp?id=1279	idk?
4	Italian?	01click.ru/per/... risposte-di-unghie.html	idk
5	idk	01click.ru/trattamento-unghie/...	idk
6	English	01webdirectory.com/baked-goods.htm	Baking Site
7	German?	0255.e-spots.nl/?id=1564	Home Page
8	Russian	03online.com/news/beremennost/2014-7-29-30380	idk
9	German?	0485.e-spots.nl/?categorie=34	idk
10	Mandarin	0509.7540.info/index.phtml...	idk
11	Mandarin	06.99bon.net/	idk
12	Mandarin	http://0708.com/t3/48/1531/1500000878.html	idk
13	English	0755zhuangxiu.com/pdetail.html?newsId=202906	idk
14	Mandarin	080.jisi.info/d/?cid=NzQ5OA==	idk
15	English	080baby.com/zhanhui/2627.shtml	Home Page
16	Mandarin	086ic.cn/tradeinfo/trade/8.html	Home Page
17	Mandarin	0898hnly.com/yglhj/1312.html	Home Page
18	Mandarin	0aw.net/index.php/index/goods/...	Home Page
19	Mandarin	0sq.whzxyy.cn/	Home Page
20	Mandarin	0ycmu.bifim.com	Home Page
21	French?	1-72.forumgratuit.org/...	Home Page
22	English	1.bp.blogspot.com/...	idk
23	Russian	100-porno.wtfdexter.com/public/	idk
24	Random?	/1001-rimes.com/listeperson.php?letter=o+eau+ots	idk
25	French	1001-votes.com/vote/sondage-le-rock-22393.html	idk

It took 45 documents to reach one that was reasonably high quality since it had coherent english sentences and quotes from real people like: "The best comment that I received was from our Regional Managing Partner (he is the highest ranking in our region). He said the food was the best we have ever had at any of our events. Thanks again for all of your help. The team was great to work with over there!" – Heather R. - BDO

The vast majority of pages were porn, just images, empty, or incomprehensible.

2 extract_text

(a) CODE ONLY

(b) The primary difference is that the resiliparse library formats the text a little better than the CC WET file. For instance, it has more readable lists using bullet points and more consistent spacing and line breaks. It still contains some of the extraneous

HTML content like some weird numbers and "false"s, but these outliers are more clearly differentiated from the rest of the natural language due to the clearer formatting.

3 language_identification

(a) CODE ONLY

(b) One problem is with the language identification model not working well enough. If there are too many false positives, then you will be training on multilingual data when that may not be intentional. If there are too many false negatives, then you will have less high quality data than you might need. Another potential issue is with marginalizing underrepresented groups in the training data. For example, if you are only accepting documents with a confidence of English above 0.9, then you might start getting rid of documents from people with non-English names or about non-English speaking countries, even if the documents themselves are in English.

(c) My random sample of 20 documents had some Russian, English, German, Japanese, and Mandarin and it correctly identified all of them. 11 of the 20 documents were in English. I believe an English classifier threshold of 0.5-0.6 would be suitable since documents with this level of confidence seemed to still clearly be English. This may not be the case for other languages where the confidence levels seem to be higher overall.

4 mask_pii

(a) CODE ONLY

(b) CODE ONLY

(c) CODE ONLY

(d) One problem with removing all instances of emails, phone numbers, and ip addresses is that the language model will not learn anything about what these things are or how they are formatted. Another issue is that the model might start literally writing

`IP\_ADDRESS`

or the other custom tokens when generating text. This could be mitigated by either replacing them with random ip addresses, emails, and phone numbers, (like 123-4567 or example@me.com), or it could be treated as a special token in the tokenizer.

(e) Overall the replacements worked pretty well, but the phone number regex was a little over sensitive. Here were some false positives that occurred using my original functions:

000-084 2012 = PHONE_NUMBER

715509390898681801.jpg = PHONE_NUMBER98681801.jpg

I made a slight alteration to the phone regex to exclude cases where there's a preceeding digit or subsequent digit which cut down on a lot of false positives, but there were still a few like:

6101555552016-02-26 21:20:29 = PHONE_NUMBER-02-26 21:20:29

The IP Addresses and Emails worked great:

7.21.0.0 Portable = IP_ADDRESS Portable
webmaster@01webdirectory.com = EMAIL_ADDRESS

I was not able to run into any false negatives since my regexes are overly-sensitive if anything.

5 harmful_content

(a) CODE ONLY

(b) CODE ONLY

(c) One problem that might arise is that the training data might exclude data from marginalized communities who sometimes use more derogatory language compared to WASPy spaces. Also by removing all NSFW and toxic content, the language model doesn't understand what toxic content is, so it may not be able to avoid it later or interpret it. One way of mitigating this would be to train the model on labelled toxic data so that it still learns it but it knows that it is toxic.

(d) Out of the first 1000 documents, 72 were NSFW, 184 were toxic, and 71 were both. So for the most part, NSFW is a subset of toxic. Out of the documents that were classified as toxic, pretty much all of them were pornography or a different language (but probably still pornography). The toxic filter works pretty well with the default threshold, but the nsfw threshold might need to be decreased to 0.4. Some documents that were literally straight hardcore pornography were classified as non-nsfw with confidence 0.614013, 0.69537, and 0.63835, which is absurd. The NSFW classifier as it stands is almost useless because the only websites it caught that the toxic classifier did not were Russian pornography sites.

6 gopher_quality_filters

(a) CODE ONLY

(b) The Gopher Rules tend to work pretty well. Most of the documents that it filters out are due to the alphabetic requirement which filters out pages that are not primarily natural language. It is definitely a pretty strong filter as it gets rid of close to 90% of the documents, but the ones it gets rid of seem to be qualitatively lower quality data. It will end up getting rid of a lot of data, however, since a lot of the remaining documents are not even in English.

7 quality_classifier

(a) CODE ONLY

(b) CODE ONLY

8 exact_deduplication

(a) CODE ONLY

9 minhash_deduplication

(a) CODE ONLY

10 filter_data

(a) CODE ONLY

(b) I didn't realize there was an a4 partition with a higher job limit so I only parallelized within a single task on the batch partition. This took 5 jobs lasting 8 hours each, so approximately 40 hours of filtering. Originally, it was even higher than this, so I did some benchmarking and determined which classifiers were taking the longest vs. how much they were filtering the data. The quality filter was among the faster steps (faster than the language and gopher filters) and it did a vast majority of the filtering, so I did that step first.

In total, the filter let through 2424873/184560645 documents which is about 1.3%.

Filter	Number Filtered Out	Percent Filtered Out	Percent Left
Quality	172173721	93%	7%
Language	4635251	37%	5%
Toxicity	243381	3%	5%
Gopher	5082675	66%	2%
PII	744	0%	1.3%

11 inspect_filtered_data

(a) The filtered data is definitely significantly better than the raw common crawl. Most of the data is decent quality text about news, stories, or something else. A random example is:

> The Union Network International (UNI) has written to the Prime Minister John Howard expressing its alarm at the governments drastic changes to workplace laws. UNI General Secretary Phillip Jen nings has told the Prime Minister that the proposed changes "will place Australia in further breach of fundamental labour standards" and "would leave Australia at the bottom of OECD countries with respect to protection of basic workers rights".

Some examples are questionable like the following which is decent syntactically speaking, but isn't very high quality data, especially considering this probably occurs a lot in the data set.

> This page is temporarily unavailable because a device from your location is sending large amounts of web requests. Visitors from other locations can still view the page. Please try again in a couple minutes by refreshing the page.

Even still, there are bad examples that make it through the filter like:

> aaron_burr_society_home.htmlabout_the_Aaron_Burr_Society.htmlSummer_of_Change
> .htmlFree_Money_Movement.html2nd_Whiskey_Rebellion.htmlUnions_Madison
> _Wisconsin.htmlNew_Moose_Party.htmlTompkin_Square+.htmlVampire
> _Slaying.htmlvideos.htmlBibliography.htmlPublications.htmlhttp://
> contact.aaronburrsociety.orgabout_the_Aaron_Burr_Society
> .htmlshapeimage_1_link_0shapeimage_1_link_1shapeimage_1
> _link_2shapeimage_1_link_3shapeimage_1_link_4shapeimage_1
> _link_5shapeimage_1_link_6shapeimage_1_link_7shapeimage_1
> _link_8shapeimage_1_link_9shapeimage_1_link_10shapeimage_1
> _link_11shapeimage_1_link_12

Others are obviously from websites and aren't the best example of Natural Language (which the validation set is mostly):

> Iraqi Militants Wanted Bush Re-Elected Say Hostages
> Iran condemns 'stupid' accusations of interference in Iraq

Iran FM warns against sectarian boycotts of Iraq elections
Iran condemns 'stupid' accusations of interference in Iraq
Iran FM warns against sectarian boycotts of Iraq elections
Industry Reform Necessary Before Free Trade Agreements
Musharraf had no hand in N-proliferation, says Powell
Musharraf had no hand in N-proliferation, says Powell
Musharraf had no hand in N-proliferation, says Powell

And also, there are some websites that make it through that are mostly english, but still contain other languages like Korean:

[KOREAN TEXT] (18 23)Incheon Airport passenger traffic to recover during Chuseok holiday [MORE KOREAN TEXT] issues advisories to fisheries cooperatives federation for W50b investment loss [KOREAN TEXT] Hamas weapons, tactics resemble those of NK : JCS One Store attracts W20b from Krafton [MORE KOREAN TEXT]

(b) Here are five examples of documents that got filtered out from each of the filters:
The quality filter usually filters out non-english texts, but it also handles low quality english text or non-natural language:

Medicine Hat Army Cadets
- HOME
- JOIN CADETS
- WHAT YOU WILL DO
- CADET FAQ
- WHAT'S IN IT FOR YOU
- PARENTS
- VOLUNTEER
- CURRENT CADETS
- Calendar & Orders
- ROUTINE ORDERS
- CALENDAR
- STANDING ORDERS
- DRESS STANDARDS
- PROMOTIONS
- Training
- LOCAL TRAINING
- FIELD TRAINING EXERCISES

- SPECIALTY TEAMS
- WORK EXPERIENCE
- SUMMER CAMP
- ACTIVITY SIGN UP
- Cadet Instructors & Staff
- TRAINING SCHEDULE
- INSTRUCTOR RESOURCES

The Language filter filters out non-english texts like:

Technisch notwendige Diese Cookies sind zum Betrieb der Webseite notwendig, z.B. zum Schutz vor Hackerangriffen und zur Gewährleistung eines konsistenten und der Nachfrage angepassten Erscheinungsbilds der Seite. Analytische Diese Cookies werden verwendet, um das Nutzererlebnis weiter zu optimieren. Hierunter fallen auch Statistiken, die dem Webseitenbetreiber von Drittanbietern zur Verfügung gestellt werden, sowie die Ausspielung von personalisierter Werbung durch die Nachverfolgung der Nutzeraktivität über verschiedene Webseiten. Drittanbieter-Inhalte

The toxic filter didn't do a lot of heavy lifting because the warcs were already filtered by the staff, but it would've just filtered out adult content and such. Here's an example of something that got filtered out which look like some kind of lyrics:

Tattoo tattoo, tattoo tattoo... Me no mek star young gyal stay far Me fucked up drunk an' text her Get fucked up inna club Ä?Ü Ä?Ü anyone weh say me nah fi fling cash Every day me money fling, gyal dem ink up My frien' dem bleach up and a ink up Aah... Bubble up, gyal Brrrrr pa pa

The Gopher filter is nice to keep because its manual rules for filtering are a good backup for filtering out non-natural language texts like:

Number of Sequences: 9484
Number of Hits to DB: 330,688
Number of extensions: 8684
Number of successful extensions: 8684
Number of sequences better than 1.0e-10: 0
Number of HSP's gapped: 8684
Number of HSP's successfully gapped: 0

```
Length of query: 2140
Length of database: 8,280,457
Length adjustment: 17
Effective length of query: 2123
Effective length of database: 8,119,229
Effective search space: 17237123167
Effective search space used: 17237123167
X1: 11 (21.8 bits)
X2: 15 (29.7 bits)
X3: 50 (99.1 bits)
S1: 12 (24.3 bits)
S2: 34 (67.9 bits)
```

The PII filter I added removes websites that contain a lot of personal information which would not be good websites to train on anyway:

$$\text{PHONE}_NUMBER65535fPHONE_NUMBER00000nPHONE_NUMBER00000nPHONE_NUMBE$$

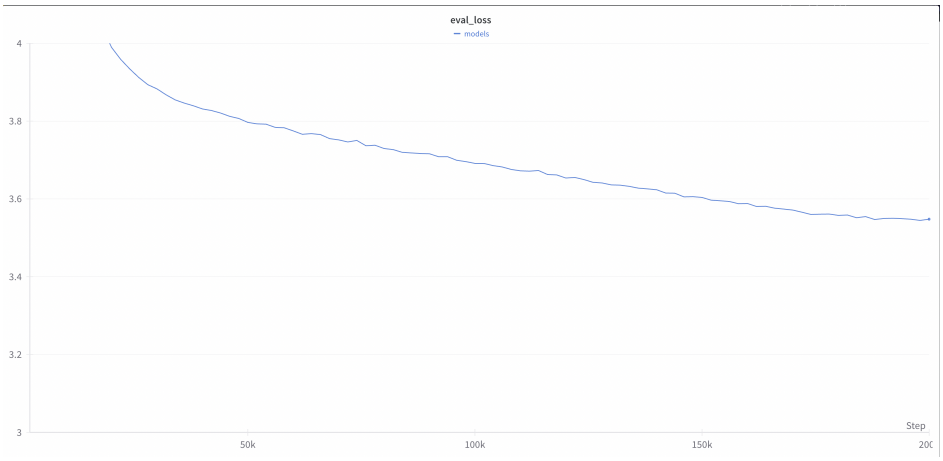
(c) I realized that a lot of the websites that were making it through the filter were just long lists of titles or subjects (probably from website tabs or categories) which are not great to train on. So, I removed lines with less than 50 characters in order to focus more on long form natural language.

12 tokenize_data

(a) CODE ONLY

13 train_model

(a)



The best validation loss I recorded was 3.545.
To filter the common crawl, I chose documents that satisfied the following properties:

(a) Quality Classifier labelled them as high quality

(b) Language Classifier labelled language as English

(c) Toxicity Classifier labelled them as non-toxic

(d) Follows Gopher Quality standards

(e) Has less than 50 IP addresses + phone numbers + emails

Additionally for approved documents I removed lines with less than 50 characters.