Net Scraping a Corpus

Let's say you have a natural language model, in the form of a chat bot. To ask it a question it must be given the relevant data, read it quickly, and spit it back to you in a natural way. That relevant data is taken from a corpus, and one way to build a corpus is to search the *near* infinite web for text related to whatever subject we want to build a language model over. Let's say you are building a little minion robot from the film *Despicable Me*. We can use articles on the web to scrape for relevant details about the film, and the little minion could answer questions about its master Gru, or what whatever may have happened during the film.

Building a Knowledge Base

Given an understanding of the foundations of the web, things like HTML, CSS, and HTTP, it's easy to build a python script to find relevant articles to a given topic. So, to build a knowledge base for our Minion chat bot, we scraped the web. Considering our interest in Despicable Me, we started out with a root url for the *Despicable Me* Wikipedia page. Using a library for handling requests from the web like *urllib*, we can open and read URLs. So, we created a queue and searched our root URL, adding urls that met certain criteria to the queue. If a link (another URL) on a page contained the keyword despicable, wasn't just a google page, and met the proper formatting, we added it to the queue. Then, kept popping the queue while adding relevant URL's back into it. Doing this 30 times, we had built a list of relevant links to the original wikipedia article.

For each link in this queue, we searched its content for text using the library BeautifulSoup (BS). BS allowed us to easily interface with HTML elements acquired with *urllib* and scan any URL we opened for paragraph tags. If a page successfully yielded text, we wrote it into a numbered text file and moved on through the queue. This text, being a little rough after just being scraped off random web sources, is then cleaned with the text processing tools regex and nltk's sentence tokenizer. We simply "chunked" the text by splitting it between lines and removing whitespace. Then, we applied nltk.sent_tokenize to further divide the text by its sentences. The clean result was then printed in a clean text file, with each sentence separated by newlines.

This text, while readable, is not cleanly labeled for a model to reference however. So, we processed the text further to extract important words That could be used to index the knowledge base. By removing stop words and reducing the text to lower case alpha characters, we were able to scan each clean file for the most common words across all

text files. The 30 most relevant links revealed the top 25 words:

```
Top 25 most common words and their counts: [('despicable', 185), ('de', 171), ('film', 136), ('gru', 126
), ('en', 115), ('minions', 103), ('fan', 92), ('op', 90), ('e', 85), ('animation', 81), ('yn', 78), ('d
an', 72), ('annecy', 66), ('illumination', 57), ('one', 50), ('universal', 47), ('new', 46), ('minion',
46), ('een', 46), ('films', 44), ('dvd', 44), ('movie', 44), ('se', 40), ('animated', 40), ('het', 39)]
```

It would take some work to automate removing words from other languages, so we just manually sort out the top ten words:

```
topTenWords = ['despicable', 'film', 'gru', 'minions', 'animation',
               'fan', 'illumination', 'universal', 'movie', 'animated']
```

With those top ten words we finally can sort back through our scanned files of text, and store sentences containing those words into a python dictionary. This dictionary could later be referenced to quickly find content relating directly to minions in the context of the film. For example, if we just read text containing minions we get a wall of text just related to the little yellow guys (or whatever they are):

```
able me will produce some progeny.', 'gru must rescue his legion of obedient yellow minions.', 'an early sign that gru has a formi
dable foe is the capture by magnet of his legion of cute, obedient yellow minions, whom the serum transforms into an enemy force o
f furry purple warriors.', "that would collectively be gru's minions, those two-tone time release capsules with goggles whose gibb
erish and pratfalls make children squeal with delight.", 'one child equals two adults in box office math, so this way, minions, to
 center stage.', 'butt out, minions.', "i'm guessing the emphasis on minions had something to do with it.", "that would collective
ly be gru's minions, those two-tone time release capsules with goggles whose gibberish and pratfalls make children squeal with del
ight.", 'one child equals two adults in box office math, so this way, minions, to center stage.', 'butt out, minions.', "i'm guess
ing the emphasis on minions had something to do with it.", 'gru, his girls and his vast army of minions return in despicable me 2,
 the follow-up to the blockbuster 3d cgi feature that grossed more than $540 million at the worldwide box office and became the te
nth biggest animated film in domestic history.', 'gru, his girls and his vast army of minions return in despicable me 2, the follo
w-up to the blockbuster 3d cgi feature that grossed more than $540 million at the worldwide box office and became the tenth bigges
t animated film in domestic history.', 'plus, there's more minions, spotted in the trailer having a grand ol' time behind bars.',
'"the minions have been patiently waiting for gru to get over this phase of being good, but when that doesn't come to fruition, me
l is the one who can't take anymore and is speaking out, and so all the other minions get behind him to be like, 'preach it!'" he
said.', 'there is only one thing certain about a new movie featuring those pellet-shaped jaundiced critters unimaginatively called
 minions, and it's that every new film will be as creatively bankrupt as the last.', 'elsewhere, the minions go on their own separ
ate adventure, tumbling from one mishap to another in varying degrees of comedic success.', 'announced by the annecy festival as a
 world premiere, "despicable me 3" continues chris meledandri's close relationship with the french festival which has hosted the b
ows of the first two parts of the franchise and "minions" – to a $2.7 billion worldwide box office to date.', 'despicable me 2 min
ions laughing - h 2013', 'news of a third despicable me comes as the second film in the franchise celebrates a dazzling run at the
 global box office, earning north of $935 million to date, one of the best showings of all time for an animated film (illumination
's minions spinoff opens in theaters july 10, 2015).', 'despicable me 2 minions laughing - h 2013', 'news of a third despicable me
 comes as the second film in the franchise celebrates a dazzling run at the global box office, earning north of $935 million to da
te, one of the best showings of all time for an animated film (illumination's minions spinoff opens in theaters july 10, 2015).',
'after comcast corp.'s universal pictures acquired dreamworks animation last year, executives there immediately began talking abou
t chris meledandri taking on an expanded role, giving the "minions" master responsibility for "shrek" and "kung fu panda," too.',
'meledandri has a "genius for creating deeply memorable and insanely hysterically funny characters," such as the minions and scrat
 from "ice age," he said.', 'meledandri's hits, including the "ice age" series he worked on for 21st century fox inc. and last yea
r's "the secret life of pets," an $875 million hit for universal, have charmed their way into popular culture, with characters lik
e the lovable super villain gru and his squeaky, yellow henchmen the minions popping up on everything from backpacks to theme-park
 rides.', 'there are display cases in the lobby filled with minions merchandise.', 'on one wall are storyboards for weekly cartoon
s that appeared on twitter.com before the release of "despicable me 3." on another wall were drawings for minions-themed puma snea
kers, designer apparel and a theme-park attraction in beijing.', 'many are still surprised, he said, to find out that his 2015 fil
m "minions" is the second-highest grossing animated picture of all-time, behind only disney's "frozen." "despicable me 3" has take
n in almost three times as much as pixar's "cars 3," which also came out in june.', '"despicable me 3" did fall 23 percent short o
f its predecessor, "minions," in the north american box office, though it's done better overseas.', 'hollywood's dependence on fra
nchises with familiar characters can be limiting, said meledandri, who has "minions 2," "sing 2" and "secret life of pets 2" in th
e works.', 'the despicable me franchise, launched in 2010, follows gru (steve carell)—a reformed super-villain who ultimately beco
mes a father, husband, and secret agent—and his yellow-hued pack of minions.', 'illumination set a july 1, 2022 release date for i
ts covid-delayed despicable me spin-off sequel, minions: the rise of gru, back in march.', 'the\xa0despicable me\xa0and\xa0minions
\xa0movies have thus far collectively grossed more than $3.7 billion worldwide.', 'the despicable me franchise, launched in 2010,
follows gru (steve carell)—a reformed super-villain who ultimately becomes a father, husband, and secret agent—and his yellow-hued
 pack of minions.', 'illumination set a july 1, 2022 release date for its covid-delayed despicable me spin-off sequel, minions: th
e rise of gru, back in march.', 'the\xa0despicable me\xa0and\xa0minions\xa0movies have thus far collectively grossed more than $3.
7 billion worldwide.', 'after helping despicable me and despicable me 2 make all the money, the minions got their own movie.', 'as
 the trailer shows, minions tells the little yellow creatures' history of accidentally murdering the greatest villains of all time
.', 'the main plot focuses on new york city in the late '60s, where three minions are looking for their next villain to "help." ma
ybe it's don draper!']
```

We see that we find both reviews that mention the minions in passing, the minions films

in relation to their success in Despicable Me, and more. For a larger knowledge base that is more usable for a chat bot, I'd consider actually reading all languages and just translating everything that comes in.

How Might We Use this?

The plan is to simply convert all of this to the minion language using online API's like Lingo Jam. While this simple corpus isn't enough to form fully informed messages, imagine just chatting with a Minion chat bot you don't understand. If we were to hypothetically just talk to the minion we could get results like:

```
> Hi there what is your name?
     > Bello, ka'm Bob → "Hello, I'm Bob"
> Oh. What?
   > to domo dub ta lingu → "You don't speak the language"
> What am I talking to?
   > a minion da ta watton hyp "despicable me"
     → "A minion from the popular film Despicable Me"
> Oh God. I hate that movie.
   > pik's a gopa ore, Yi kai yai yai! aca nama tem titdak
phiens:
   → "It's a great movie! Here are some critics reviews:"
```

In this use case the language model doesn't have to be that good, but it does help that it could properly throw in proper nouns, and speak with some relevance to the source material.

Conclusion:

I can see this being used to generate some wonderfully ironic Minion memes. Regardless, here lies the structure to create and expand a larger corpus that could be used to train a language model. Which is super cool!