Houston Holman

3/11/23

Web Scraping

Running the Code

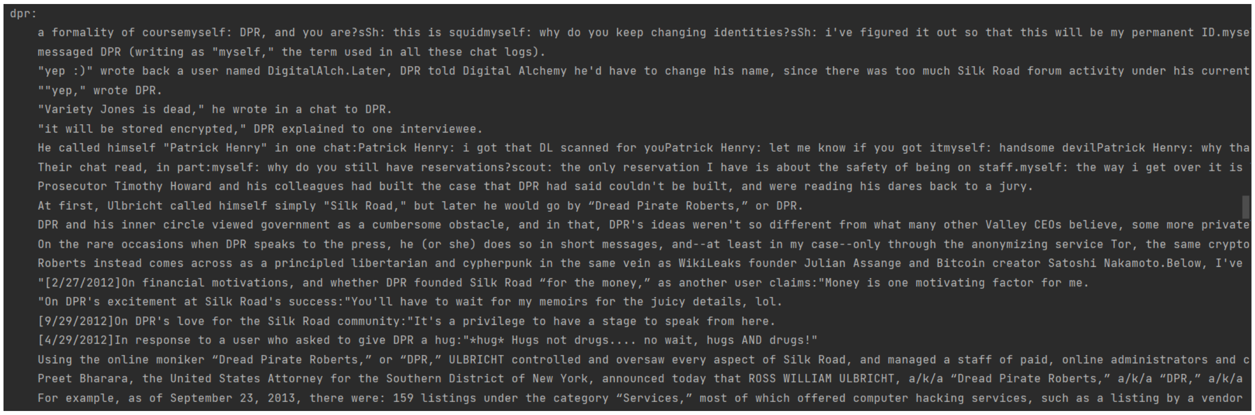
1. Run web_crawler.py to generate the list of urls stored in urls.txt
2. Run web_scraper.py to scrape text off of the urls and store them in pages
3. Run file_cleaner.py to clean the files in pages and store them in cleaned_pages
4. Run term_extractor.py to see the list of most important terms found in cleaned_pages
5. Run knowledge_base_creator.py to create the knowledge base from the selected best words from term_extractor.py

Building the Knowledge Base

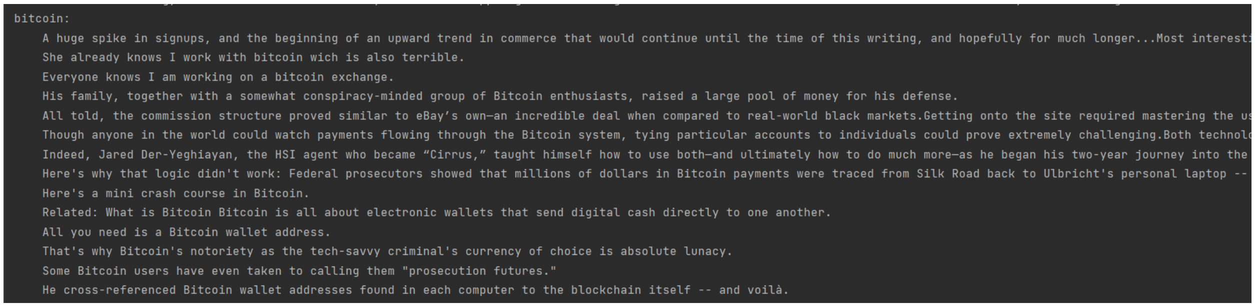
I began by choosing the 10 best words

```
best_words = ['ulbricht', 'silk', 'dpr', 'bitcoin', 'drug', 'fbi',
'tor', 'market', 'computer', 'court']
```

For each word, I then loop through every sentence of each document. If the word is contained in the sentence, that sentence is added to an array. The result is a dictionary where the key is the word and the value is an array of sentences that contain that word. Below are some examples taken from the knowledge base. The entire dictionary is contained in the file knowledge_dict.p that is created by knowledge_base_creator.py



```
dpr:
    a formality of coursemyself: DPR, and you are?sSh: this is squidmyself: why do you keep changing identities?sSh: i've figured it out so that this will be my permanent ID.myse
    messaged DPR (writing as "myself," the term used in all these chat logs).
    "yep :)" wrote back a user named DigitalAlch.Later, DPR told Digital Alchemy he'd have to change his name, since there was too much Silk Road forum activity under his current
    ""yep," wrote DPR.
    "Variety Jones is dead," he wrote in a chat to DPR.
    "it will be stored encrypted," DPR explained to one interviewee.
    He called himself "Patrick Henry" in one chat:Patrick Henry: i got that DL scanned for youPatrick Henry: let me know if you got itmyself: handsome devilPatrick Henry: why tha
    Their chat read, in part:myself: why do you still have reservations?scout: the only reservation I have is about the safety of being on staff.myself: the way i get over it is
    Prosecutor Timothy Howard and his colleagues had built the case that DPR had said couldn't be built, and were reading his dares back to a jury.
    At first, Ulbricht called himself simply "Silk Road," but later he would go by “Dread Pirate Roberts,” or DPR.
    DPR and his inner circle viewed government as a cumbersome obstacle, and in that, DPR's ideas weren't so different from what many other Valley CEOs believe, some more private
    On the rare occasions when DPR speaks to the press, he (or she) does so in short messages, and--at least in my case--only through the anonymizing service Tor, the same crypto
    Roberts instead comes across as a principled libertarian and cypherpunk in the same vein as WikiLeaks founder Julian Assange and Bitcoin creator Satoshi Nakamoto.Below, I've
    "[2/27/2012]On financial motivations, and whether DPR founded Silk Road “for the money,” as another user claims:"Money is one motivating factor for me.
    "On DPR's excitement at Silk Road's success:"You'll have to wait for my memoirs for the juicy details, lol.
    [9/29/2012]On DPR's love for the Silk Road community:"It's a privilege to have a stage to speak from here.
    [4/29/2012]In response to a user who asked to give DPR a hug:"*hug* Hugs not drugs.... no wait, hugs AND drugs!"
    Using the online moniker “Dread Pirate Roberts,” or “DPR,” ULBRICHT controlled and oversaw every aspect of Silk Road, and managed a staff of paid, online administrators and c
    Preet Bharara, the United States Attorney for the Southern District of New York, announced today that ROSS WILLIAM ULBRICHT, a/k/a “Dread Pirate Roberts,” a/k/a “DPR,” a/k/a
    For example, as of September 23, 2013, there were: 159 listings under the category “Services,” most of which offered computer hacking services, such as a listing by a vendor
```



```
bitcoin:
    A huge spike in signups, and the beginning of an upward trend in commerce that would continue until the time of this writing, and hopefully for much longer...Most interesti
    She already knows I work with bitcoin wich is also terrible.
    Everyone knows I am working on a bitcoin exchange.
    His family, together with a somewhat conspiracy-minded group of Bitcoin enthusiasts, raised a large pool of money for his defense.
    All told, the commission structure proved similar to eBay’s own—an incredible deal when compared to real-world black markets.Getting onto the site required mastering the us
    Though anyone in the world could watch payments flowing through the Bitcoin system, tying particular accounts to individuals could prove extremely challenging.Both technolo
    Indeed, Jared Der-Yeghiayan, the HSI agent who became “Cirrus,” taught himself how to use both—and ultimately how to do much more—as he began his two-year journey into the
    Here's why that logic didn't work: Federal prosecutors showed that millions of dollars in Bitcoin payments were traced from Silk Road back to Ulbricht's personal laptop --
    Here's a mini crash course in Bitcoin.
    Related: What is Bitcoin Bitcoin is all about electronic wallets that send digital cash directly to one another.
    All you need is a Bitcoin wallet address.
    That's why Bitcoin's notoriety as the tech-savvy criminal's currency of choice is absolute lunacy.
    Some Bitcoin users have even taken to calling them "prosecution futures."
    He cross-referenced Bitcoin wallet addresses found in each computer to the blockchain itself -- and voilà.
```

Sample Dialog

> Who is DPR?

“Dread Pirate Roberts” or “DPR” was the online moniker of Ross Ulbricht, who controlled and oversaw every aspect of Silk Road.

> What was Silk Road?

Silk Road was an unregulated online marketplace where buyers paid using bitcoin, an electronic currency that is hard to trace.

> How successful was Silk Road?

Silk Road facilitated over 1.5 million transactions totaling $214 million.

> How does TOR work?

Tor, technology originally developed by the US Navy and now overseen by a nonprofit, helped to anonymize Internet use by routing requests through multiple servers, adding and removing layers of encryption along the way.

> What was Ross Ulbricht charged with in court?

Ulbricht was charged in federal court in New York with money laundering, drug dealing and conspiring to murder a witness.

> Where did Ross Ulbricht go to college?

Ulbricht went on to an undergraduate degree in physics at the University of Texas at Dallas and a prestigious fellowship to a graduate program in materials science at Penn State.