Number of URLs: 12333 including evoke

1. How do you define the token in your project?
   1. We first omit the English Stop Words; the numbers does not count as token because there’s one third of the token count that are numbers, which makes it useless; one word does not count as token. Ex. “b”, “c”. These does not occur in the English stop words but are also useless.
2. How long does your project run?
   1. Our project run for 2.5 hours since we do not have enough time; We stop running on evoke.ics.uci.edu and all other urls are successfully extracted. There’s a lot of live links like <https://mt-live.ics.uci.edu> that we haven’t crawled yet.
3. How do you ignore low information pages?
   1. We count the word for page, if it is lower than 20, then skip the page.
      1. Why 20? Because we find out that <https://today.uci.edu/department/information_computer_sciences> is 404 page not found, and the word count for this is 15, so any pages that does not found count as low information pages.
   2. For very long pages like evoke…, we find out that these pages contain lots of user reply, so we intent to extract only the first 100 reply to increase crawling speed.
4. How do you detect duplicates urls/avoid trap?
   1. We put all urls in a set.
5. How do you detect duplicates pages?
   1. We didn’t do that.
6. How do you avoid dead URLs that return a 200 status but no data?
   1. if response is 200 and content is None return empty list
7. How do you set up your program politeness?
   1. 0.5
8. Why do you have computewordfrequency and computelocalfrequency?
   1. Local is used for computing the longest word in one page (for question 2)
   2. Word frequency is used for computing the 50 most common words for the entire pages (for question 3)
9. Why are you setting isVaild function inside extract\_next\_link?
   1. Just to make sure that the url does pass our criteria
10. What is the ban list? How do you come up with this list?
    1. We didn’t use it
11. Why your subdomain doesn’t include evoke?
    1. We compute the url for evoke and for the rest separately. The subdomain only contains url for the rest.
12. Do you ban some URLs that are hard/takes time to crawl?
    1. Yes. We have banned some of URLs which include evoke. For some large evoke, we extract only the first 100 customer reply and the content