JAMES COOK
UNIVERSITY
AUSTRALIA

Author: Hendrik A. Dreyer

Student ID: 13622464

Course: Master of Data Science

Faculty: School of Science, Engineering and IT

Subject: MA5851 – Data Science Master Class 1

Assessment: 3 – Web Crawler Report

Due Date: 1 December 2019

Introduction

Hacker News, (https://news.ycombinator.com/), is a social news website focusing on computer science and entrepreneurship. It is run by Paul Graham's investment fund and startup incubator, Y Combinator. In general, content that can be submitted is defined as "anything that gratifies one's intellectual curiosity (“Hacker News,” 2019).

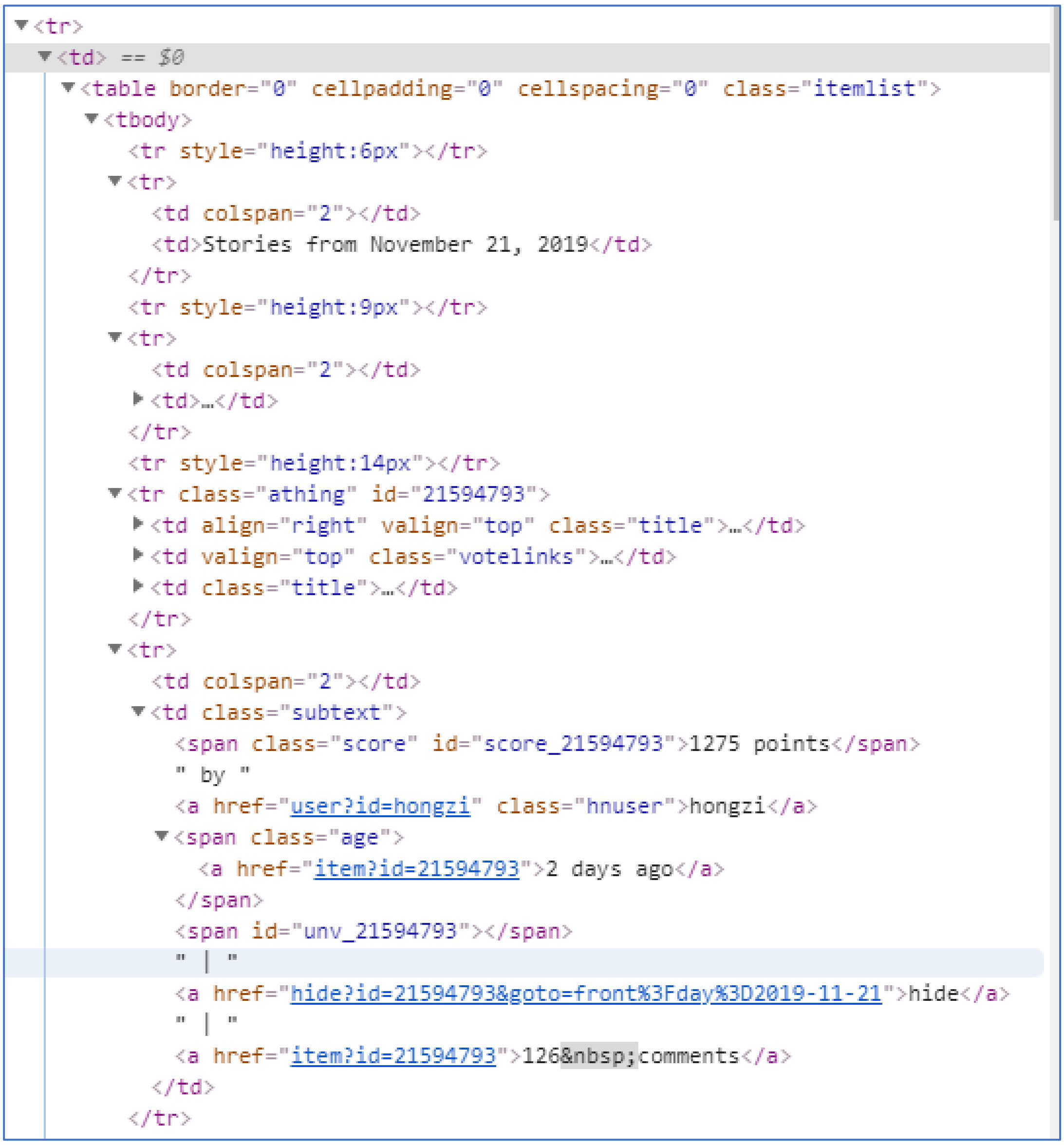
HN is a public website whereby people mostly post links to media articles. Users can then comment on the posted links and thus public dialogue is encouraged. HN devised a complex scoring algorithm whereby posts are assigned points. However, it is important to mention that the posts are scored and not the comments, although the number of comments on a post contributes towards the score of the post. The other important point to remember about the posts is that the actual wording of a post is not (probably 99% of the time) the words posted by the user, it is in fact the words directly related to the title of the article that the post refers to. In other words, it is the written word as published by the media. Therefore, any corpus derived from the scraped titles are content generated by the media. Herein lies an interesting phenomenon as we now can look into the narrative that the media shapes around all things contemporary.

Looking at the top scored posts, the industry related question(s) that arises from this information revolves around the question of why the media would word articles the way they do? And, equally important, firstly, why are users of the forum posting these specific articles, and secondly, why are users responding more to certain articles than others. By understanding the forces that push and pulls users to react in certain ways to posted links on the forum can be utilised as an effective insight into various industry activities such as, marketing, opinion polls, voting, etc.

Content Layout

The site, https://news.ycombinator.com/, has a simple layout as far as content is concerned. All posts, dating back to the initial start-up date of the forum, can be accessed via a top-level menu option eloquently labelled, **past**. It is through this option that the author managed to harvest the last 365 days’ worth of front-page posts. Each day’s front page contains the top 30 scored posts for the day. Thus, the scraped data represents the highest scored posts per day for the last year.

Figure 1 below, illustrates the layout of a single post in the Hacker News forum. All post entries are encapsulated and ordered via an HTML table object (tb). Each post is, in turn, is encapsulated by two HTML table row objects (tr). This layout caused some difficulty in scraping the information from posts as they are segregated by an id. Therefore, two separate scraping runs had to be executed.



```
▼<tr>
  ▼<td> == $0
    ▼<table border="0" cellpadding="0" cellspacing="0" class="itemlist">
      ▼<tbody>
          <tr style="height:6px"></tr>
        ▼<tr>
            <td colspan="2"></td>
            <td>Stories from November 21, 2019</td>
          </tr>
          <tr style="height:9px"></tr>
        ▼<tr>
            <td colspan="2"></td>
          ▶<td>…</td>
          </tr>
          <tr style="height:14px"></tr>
        ▼<tr class="athing" id="21594793">
          ▶<td align="right" valign="top" class="title">…</td>
          ▶<td valign="top" class="votelinks">…</td>
          ▶<td class="title">…</td>
          </tr>
        ▼<tr>
            <td colspan="2"></td>
          ▼<td class="subtext">
              <span class="score" id="score_21594793">1275 points</span>
              " by "
              <a href="user?id=hongzi" class="hnuser">hongzi</a>
            ▼<span class="age">
                <a href="item?id=21594793">2 days ago</a>
              </span>
              <span id="unv_21594793"></span>
              " | "
              <a href="hide?id=21594793&goto=front%3Fday%3D2019-11-21">hide</a>
              " | "
              <a href="item?id=21594793">126&nbsp;comments</a>
            </td>
          </tr>
```

Figure 1 - Hacker News: Inspect Code

The first scraping run harvested the following information from each post and saved to a csv file labelled, ***hacker_news_1.csv***:

a) Post id

b) Title of the article of the posted link

c) The web link to the article

The second scraping run harvested the following information from each post and saved it to a csv file labelled, ***hacker_news_2.csv***:

a) Post id

b) Points scored per post

Once the above two scraping runs completed successfully, post processing was performed on the contents of the two files whereby a join was performed on the “id” fields. This process resulted in an amalgamated data frame, with the following layout:

a) Index

b) Post id

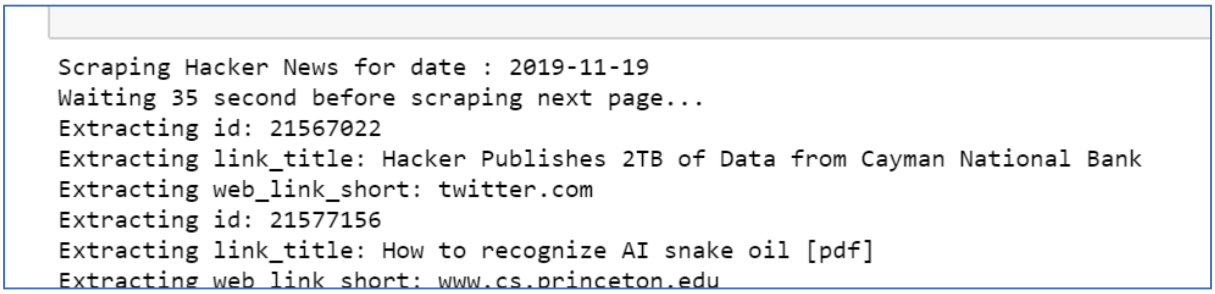
c) The title of the posted article

d) The actual web link to the article (only the base url is captured, for example if the link was pointing to https://blog.mozilla.org/the-story-about-blah/#$^538723/page?=34, then only the following was captured into the field -> blog.moziall.org)

e) Points scored by the post

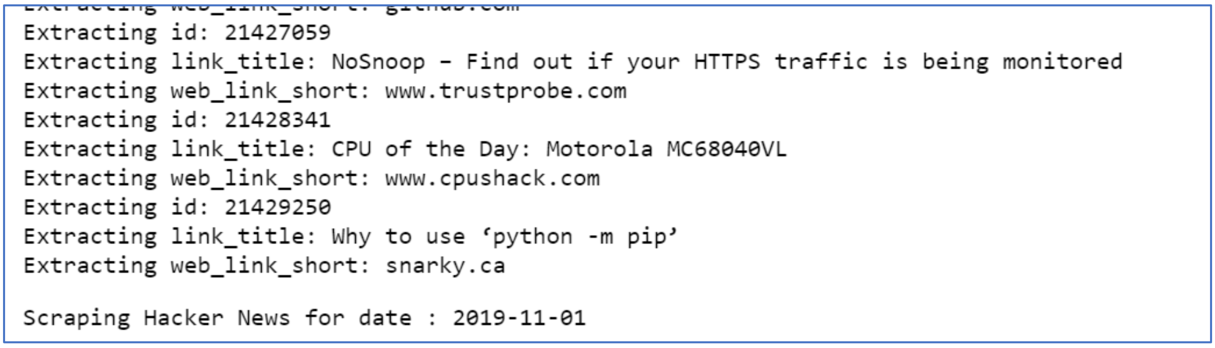
The final amalgamated data frame was saved to a csv file labelled, ***hacker_news_post_process.csv***. This file will serve as entry data to Assessment 4 and contains 10920 entries.

Evidence of the scraping process can be viewed in Figure 2 and Figure 3.



```
Scraping Hacker News for date : 2019-11-19
Waiting 35 second before scraping next page...
Extracting id: 21567022
Extracting link_title: Hacker Publishes 2TB of Data from Cayman National Bank
Extracting web_link_short: twitter.com
Extracting id: 21577156
Extracting link_title: How to recognize AI snake oil [pdf]
Extracting web_link_short: www.cs.princeton.edu
```

Figure 2 - HN - Scraping Evidence 1



```
Extracting web_link_short: github.com
Extracting id: 21427059
Extracting link_title: NoSnoop - Find out if your HTTPS traffic is being monitored
Extracting web_link_short: www.trustprobe.com
Extracting id: 21428341
Extracting link_title: CPU of the Day: Motorola MC68040VL
Extracting web_link_short: www.cpushack.com
Extracting id: 21429250
Extracting link_title: Why to use ‘python -m pip’
Extracting web_link_short: snarky.ca

Scraping Hacker News for date : 2019-11-01
```

Figure 3 - HN - Scraping Evidence 2

Preliminary Findings

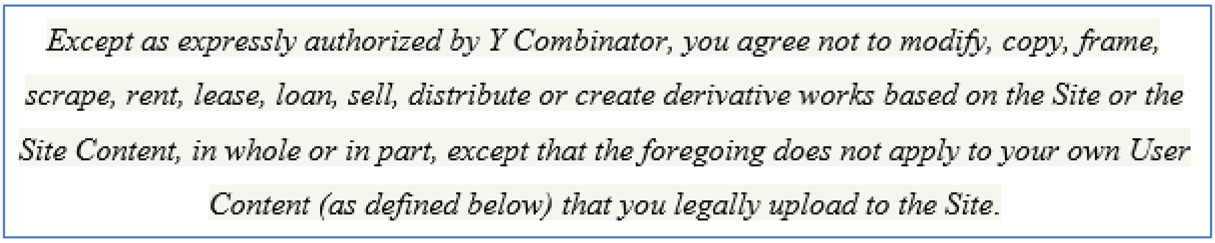
By ordering the posts by points scored in an ascending order, revealed that the top scored post during the last 365 days was pointing to an article titled, “*Switch from Chrome to Firefox”* followed by *“I sell Onions on the Internet”* and “*Announcing unlimited free private repos”*. These titles are informative, quirky and disturbing.

The second interesting thing to spot was the fact that **github.com** features three times in the top 20 posts for the timeframe of the scraped posts, which is more than any of the other URL listed in the top 20.

The deepest insights to be gained from the harvested data would be the insight drawn from the text corpus as assembled from the article link words. This activity, however, would be the focal point of Assessment 4.

Copy Rights

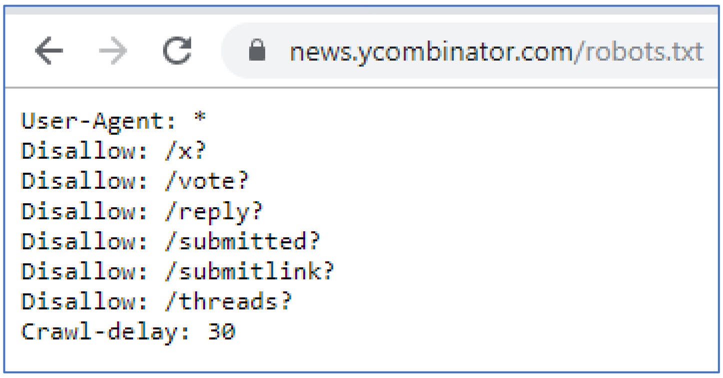
Figure 4 below lists a snippet from the section labelled, Intellectual Property Rights as taken from the page labelled, **legal,** in the Hacker News forum.



Except as expressly authorized by Y Combinator, you agree not to modify, copy, frame, scrape, rent, lease, loan, sell, distribute or create derivative works based on the Site or the Site Content, in whole or in part, except that the foregoing does not apply to your own User Content (as defined below) that you legally upload to the Site.

Figure 4 - HN - Intellectual Property Rights

The author finds this statement to be very restrictive and ill-posed in the light of the following information. The robots.txt file for the forum, as published under the URL https://news.ycombinator.com/robots.txt , explicitly specifies a crawl delay of 30 seconds. This implies that crawling of the sight is permitted if the crawl rate is not faster than 30 seconds per page. Also, all User-Agents are allowed as specified by the robots.txt file. Figure 5 illustrates the listing as published by the robotsd.txt file:



```
User-Agent: *
Disallow: /x?
Disallow: /vote?
Disallow: /reply?
Disallow: /submitted?
Disallow: /submitlink?
Disallow: /threads?
Crawl-delay: 30
```

Figure 5 - Hancker News - Robots.txt

The statement under the IP section might be contentious but, the author argues that crawling the Hacker News forum at a rate slower than 30 seconds per page, does not, in fact, constitutes illegal crawling and therefore, the information scraped does not violate any copy right laws, especially because, none of the scraped material is altered or modified and re-used for commercial purposes. Also, very important to the author's argument, is that Hacker News decided in March 2016 to publish all stories and comments from Hacker News to the public data section of Google's BigQuery. Visit Google BigQuery to view Hacker News' full published data base labelled, "***hacker_news"***, in the ***"bigquery-public-data"*** section.

References

Hacker News. (2019). In *Wikipedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Hacker_News&oldid=924206938