

BÁO CÁO BÀI TẬP THU THẬP DỮ LIỆU CHO BÀI TOÁN SARCASM DETECTION IN NEWS HEADLINE

Giảng viên hướng dẫn:

- PGS. TS. Lê Đình Duy
- ThS. Phạm Nguyễn Trường An

Sinh viên thực hiện:

- Huỳnh Thiện Tùng 19522492
- Lê Thị Thanh Thanh 19520954
- Nguyễn Thành Vương 19522542

1. Bài toán:

Mỗi nhóm chọn 03 trang báo điện tử châm biếm tiếng Anh từ danh sách sau: https://en.wikipedia.org/wiki/List_of_satirical_news_websites
Và 03 trang báo điện tử uy tín tiếng anh từ 3 quốc gia nói tiếng Anh khác nhau.

Thu thập tất cả tiêu đề của các bài báo mà 6 trang tin trên đăng trong vòng 03 năm trở lại đây.

Tổ chức dataset theo cùng format với dataset tham khảo ở đây: <https://www.kaggle.com/rmisra/news-headlines-dataset-for-sarcasm-detection>

Các nhóm được hợp tác với nhau để làm bài tập này. Tuy nhiên mỗi trang tin không được có quá 4 nhóm cùng chọn. Các nhóm phải ghi rõ các danh sách 6 trang tin mình đã chọn trong comment.

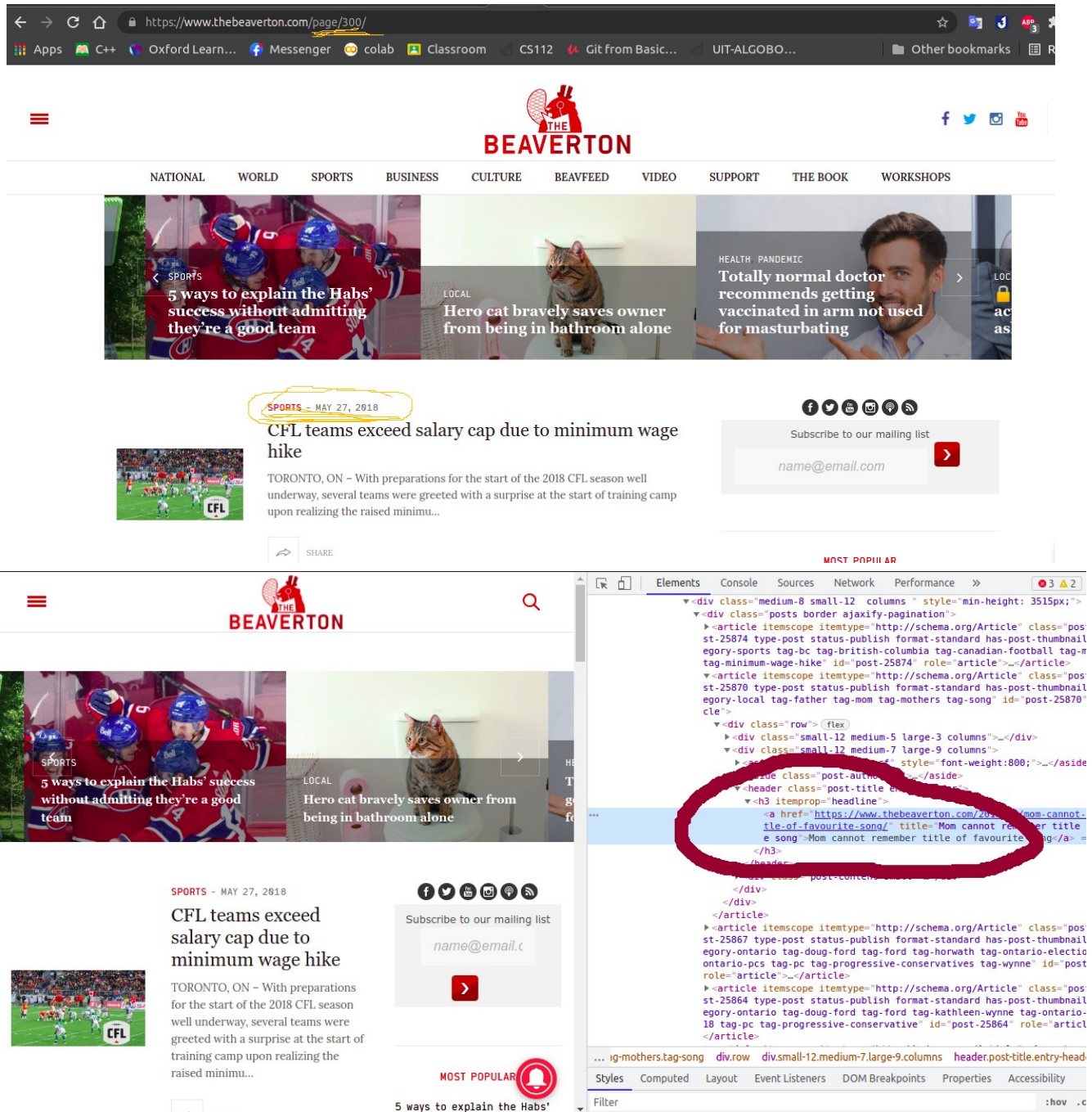
2. Định dạng:

- Notebook chứa các code cần thiết để crawl dữ liệu.
- Báo cáo kết quả quá trình thu thập dữ liệu.
- File json theo đúng format.

```
1 {
2   "root": [
3     {
4       "is_sarcastic": 1,
5       "headline": "ud83dludd12 Man who lu201cloves readinglu201d actually just likes falling asleep while holding a book",
6       "article_link": "https://www.thebeaverton.com/2021/06/man-who-loves-reading-actually-just-likes-falling-asleep-while-holding-a-book/"
7     },
8     {
9       "is_sarcastic": 1,
10      "headline": "Feminism Win! This Sex Doll has a mouth",
11      "article_link": "https://www.thebeaverton.com/2021/06/feminism-win-this-sex-doll-has-a-mouth/"
12    }
13  ]
14 }
```

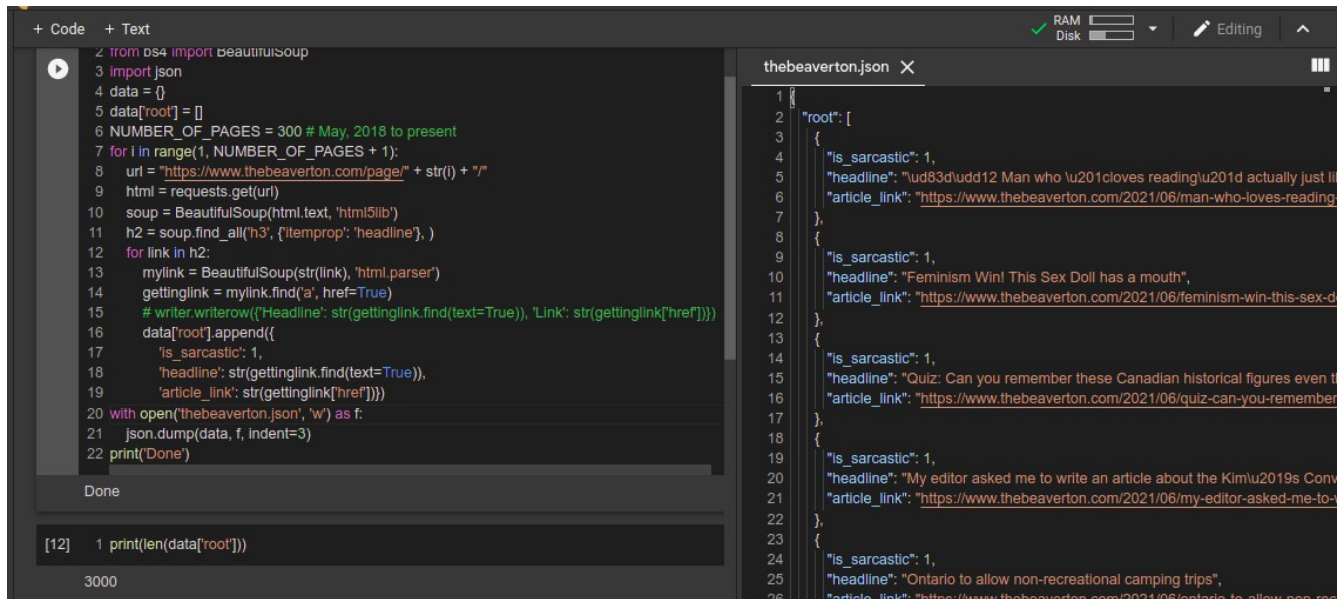
3. Phân tích kịch bản để thu thập dữ liệu:

Chúng em nhận thấy trang <https://www.thebeaverton.com/> mỗi trang có 10 headline và được sắp xếp theo thứ tự mới nhất (theo thời gian đăng bài). Phía sau url có dạng url + "/pages/" + số trang. Điều này giúp chúng em nghĩ đến vòng lặp cho chạy từ đầu đến cuối trang để lấy dữ liệu từ trang web. Nhưng để lấy những bài cách đây không quá 3 năm thì chúng em đã cho chạy đến page thứ 300 (trong hình dưới)



Sau khi kiểm tra mã nguồn của trang thì chúng em rút ra điểm chung là sẽ có thẻ **h3** với **itemprop = 'headline'** và **a href** để có thể lấy headline và link.

Dưới đây là đoạn code thu thập headline từ trang <https://www.thebeaverton.com/>



```
+ Code + Text
2 from os4 import BeautifulSoup
3 import json
4 data = {}
5 data['root'] = []
6 NUMBER_OF_PAGES = 300 # May, 2018 to present
7 for i in range(1, NUMBER_OF_PAGES + 1):
8     url = "https://www.thebeaverton.com/page/" + str(i) + "/"
9     html = requests.get(url)
10    soup = BeautifulSoup(html.text, 'html5lib')
11    h2 = soup.find_all('h3', {'itemprop': 'headline'}, )
12    for link in h2:
13        mylink = BeautifulSoup(str(link), 'html.parser')
14        gettinglink = mylink.find('a', href=True)
15        # writer.writerow(("Headline": str(gettinglink.find(text=True)), "Link": str(gettinglink["href"])))
16        data['root'].append({
17            'is_sarcastic': 1,
18            'headline': str(gettinglink.find(text=True)),
19            'article_link': str(gettinglink["href"])
20        })
21    with open("thebeaverton.json", 'w') as f:
22        json.dump(data, f, indent=3)
23    print("Done")

[12] 1 print(len(data['root']))

3000

thebeaverton.json
1 {
2   "root": [
3     {
4       "is_sarcastic": 1,
5       "headline": "ud83d\udd12 Man who \u201cloves reading\u201d actually just li
6       "article_link": "https://www.thebeaverton.com/2021/06/man-who-loves-reading-
7     },
8     {
9       "is_sarcastic": 1,
10      "headline": "Feminism Win! This Sex Doll has a mouth",
11      "article_link": "https://www.thebeaverton.com/2021/06/feminism-win-this-sex-d
12    },
13    {
14      "is_sarcastic": 1,
15      "headline": "Quiz: Can you remember these Canadian historical figures even t
16      "article_link": "https://www.thebeaverton.com/2021/06/quiz-can-you-remember
17    },
18    {
19      "is_sarcastic": 1,
20      "headline": "My editor asked me to write an article about the Kim\u2019s Con
21      "article_link": "https://www.thebeaverton.com/2021/06/my-editor-asked-me-to-
22    },
23    {
24      "is_sarcastic": 1,
25      "headline": "Ontario to allow non-recreational camping trips",
26      "article_link": "https://www.thebeaverton.com/2021/06/ontario-to-allow-non-re-
```

Các trang còn lại cũng tương tự như trên, nhưng cần kiểm tra mã nguồn của trang và điều chỉnh lại code một chút.

4. Kết quả:

Nhóm em đã thu thập dữ liệu từ 06 trang và nhận được kết quả như sau:

1. <https://www.thebeaverton.com/> : 3000 headlines
2. <https://clickhole.com/> : 2509 headlines
3. <https://www.thepoke.co.uk/> : 12160 headlines
Tổng: 17699 headlines (is_sarcasti = 1)
4. <https://www.nbcnews.com/> : 49609 headlines
5. <https://www.economist.com/asia> : 6600 headlines
6. <http://www.thecivilian.co.nz/> : 600 headlines
Tổng: 56809 headlines (is_sarcasti = 0)

Tổng cộng: 74508 headlines