

# BÁO CÁO BÀI TẬP COLAB 4

## CRAWL DATA

-Lớp: Máy học - CS114.L22.KHCL

-Giảng viên hướng dẫn:

- PGS.TS. Lê Đình Duy

- Ths. Phạm Nguyễn Trường An

-Nhóm thực hiện:

- Thái Trần Khánh Nguyên - 19520188

- Nguyễn Khánh Như - 19520209

- Đoàn Nguyễn Nhật Quang - 19520235

### 1. Phân tích bài toán:

#### Đề bài:

- 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](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.
- **Bài nộp bao gồm:**
  - 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.

#### Phân tích:

- Sau khi tham khảo [link](#) thầy gửi thì nhóm em nhận thấy dữ liệu cần crawl từ các trang web là:

- is\_sarcastic: 1 nếu là báo châm biếm còn lại là 0

-headline: tiêu đề của bài báo

-article\_link: link dẫn đến bài báo

-Sau khi tham khảo các web châm biếm và các web báo chính thống thì nhóm em đã chọn ra 6 trang web để crawl dữ liệu:

- Trang báo châm biếm:

+ [The Babelon Bee](#)

+ [The Beaverton](#)

+ [The Chaser](#)

- Trang báo chính thống:

+ [The Guardian](#)

+ [NBC News](#)

+ [The Sun](#)

## 2. Phân tích các trang web cần crawl:

Sau khi tìm hiểu trên website [The Babylon Bee](https://www.babylonbee.com/) nhóm em có một vài nhận xét như sau:



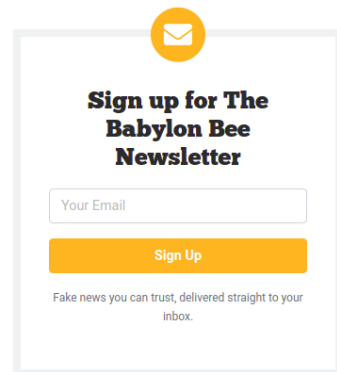
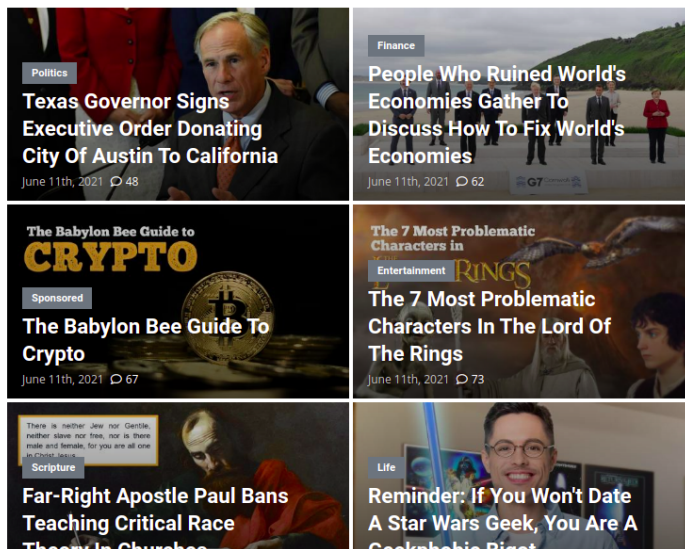
### Latest Videos



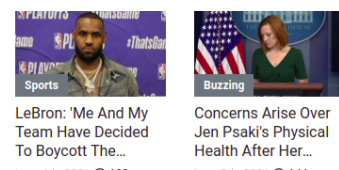
- Để có thể xem được danh sách bài báo đã được đăng thì website này có cung cấp đường dẫn có dạng :  
“[babylonbee.com/news?page=](https://babylonbee.com/news?page=1)” + số thứ tự của trang cần xem danh sách  
Ví dụ muốn xem danh sách các bài báo ở trang số 1 thì đường link sẽ là:  
“<https://babylonbee.com/news?page=1>”
- Và các bài báo ở đây được sắp xếp theo dạng newest:

Home > Categories > All

### Latest News



### Must Read



- Điều nhóm em cần quan tâm đến là headline của bài báo, article link và thời gian đăng để khớp với yêu cầu các bài báo trong vòng 3 năm gần đây. Do đây là trang web báo châm biếm nên hiển nhiên `is_sarcastic = 1`

- Sau khi xem thử file html của web thì nhóm em cũng đã phát hiện ở thẻ article-card có lưu tất cả những thông tin mà chúng em cần

The screenshot shows a web browser with a code editor open, displaying a BeautifulSoup script. The script is designed to fetch the latest news from the Babylon Bee website. The output of the script is a list of news items, each with a title, image path, and primary category.

```

[1] 1 import requests
    2 from bs4 import BeautifulSoup as Soup

[2] 1 url = 'https://babylonbee.com/news?page=1'
    2 request = requests.get(url)
    3 soupSite = Soup(request.text, 'html')

[3] 1 soupSite
    <a href="/>Home</a> > >
    <a href="https://babylonbee.com/news/categories">Categories</a> > >
    All
    <div class="plan-title registration-title">Latest News</div>
    <div class="row mb-2 gutter-4">
    <div class="col-sm-6">
    <article-card :comment_count="48" :image_path="https://media.babylonbee.com/thumbs/article-8832-1-thumb.jpg" :is_premium="0" :path="/news/texas-governor-signs-executive-order-donati
    /article-card>
    </div>
    <div class="col-sm-6">
    <article-card :comment_count="63" :image_path="https://media.babylonbee.com/thumbs/article-8831-1-thumb.jpg" :is_premium="0" :path="/news/people-who-ruined-worlds-economies-gather-t
    /article-card>
    </div>
    <div class="col-sm-6">
    <article-card :comment_count="67" :image_path="https://media.babylonbee.com/thumbs/article-8820-2-thumb.jpg" :is_premium="0" :path="/news/the-babylon-bee-guide-to-bitcoin" :primary
    /article-card>
    </div>
    <div class="col-sm-6">
    <article-card :comment_count="73" :image_path="https://media.babylonbee.com/thumbs/article-8828-2-thumb.jpg" :is_premium="0" :path="/news/7-problematic-characters-in-the-lord-of-the
    /article-card>
    </div>
    <div class="col-sm-6">
    <article-card :comment_count="89" :image_path="https://media.babylonbee.com/thumbs/article-8827-3-thumb.jpg" :is_premium="0" :path="/news/far-right-apostle-paul-bans-teaching-critic
    /article-card>
    </div>
    <div class="col-sm-6">
    <article-card :comment_count="76" :image_path="https://media.babylonbee.com/thumbs/article-8823-3-thumb.jpg" :is_premium="0" :path="/news/reminder-if-you-wont-date-a-star-wars-nerd-
    /article-card>
    </div>
    <div class="col-sm-6">
    <article-card :comment_count="198" :image_path="https://media.babylonbee.com/thumbs/article-8824-3-thumb.jpg" :is_premium="0" :path="/news/toobin-gets-off-easy" :primary_category="
    /article-card>
    </div>
  
```

The output shows a list of news items, each with a title, image path, and primary category. The items are:

- Latest News
- Categories
- Home
- Article 1: Texas Governor Signs Executive Order Donating...
- Article 2: People Who Ruined World's Economies Gather...
- Article 3: The Babylon Bee Guide to Bitcoin
- Article 4: 7 Problematic Characters in The Lord of the...
- Article 5: Far-Right Apostle Paul Bans Teaching Critic...
- Article 6: Reminder If You Won't Date a Star Wars Nerd...
- Article 7: Toobin Gets Off Easy

```
[1] 1 import requests
    2 from bs4 import BeautifulSoup as Soup

[2] 1 url = 'https://babylonbee.com/news?page=1'
    2 request = requests.get(url)
    3 soupSite = Soup(request.text, 'html')

1 print(soupSite.find_all("article-card"))

<article-card :comment count="48" :image path="https://media.babylonbee.com/thumbs/article-8832-1-thumb.jpg" :is premium="0" :path="/news/texas-governor-signs-executive-order-donatin
</article-card>
<article-card :comment count="63" :image path="https://media.babylonbee.com/thumbs/article-8832-1-thumb.jpg" :is premium="0" :path="/news/people-who-ruined-worlds-eco
</article-card>
<article-card :comment count="67" :image path="https://media.babylonbee.com/thumbs/article-8832-1-thumb.jpg" :is premium="0" :path="/news/the-babylon-bee-guide-to-bli
</article-card>
<article-card :comment count="73" :image path="https://media.babylonbee.com/thumbs/article-8878-2-thumb.jpg" :is premium="0" :path="/news/7-problematic-characters-in
</article-card>
<article-card :comment count="89" :image path="https://media.babylonbee.com/thumbs/article-8827-1-thumb.jpg" :is premium="0" :path="/news/far-right-apostle-paul-bans-
</article-card>
<article-card :comment count="76" :image path="https://media.babylonbee.com/thumbs/article-8874-1-thumb.jpg" :is premium="0" :path="/news/reminder-if-you-wont-date-a-
</article-card>
<article-card :comment count="198" :image path="https://media.babylonbee.com/thumb/article-8851-1-thumb.jpg" :is premium="0" :path="/news/toobin-gets-off-easy" :pri
</article-card>
<article-card :comment count="75" :image path="https://media.babylonbee.com/thumbs/article-8824-1-thumb.jpg" :is premium="0" :path="/news/osteen-an-attack-on-me-is-an
</article-card>
<article-card :comment count="63" :image path="https://media.babylonbee.com/thumbs/article-8822-2-thumb.jpg" :is premium="0" :path="/news/csn-enforced-to-reset-days-sin
</article-card>
<article-card :comment count="75" :image path="https://media.babylonbee.com/thumbs/article-8888-1-thumb.jpg" :is premium="0" :path="/news/if-you-dont-use-the-n-word-y
</article-card>
<article-card :comment count="68" :image path="https://media.babylonbee.com/thumb/article-8819-1-thumb.jpg" :is premium="0" :path="/news/pelosi-asks-for-clarity-on
</article-card>
<article-card :comment count="94" :image path="https://media.babylonbee.com/thumbs/article-8817-1-thumb.jpg" :is premium="0" :path="/news/new-york-times-relocates-off
</article-card>
<article-card :comment count="82" :image path="https://media.babylonbee.com/thumbs/article-8814-2-thumb.jpg" :is premium="0" :path="/news/biden-deploys-us-military-to
</article-card>
<article-card :comment count="88" :image path="https://media.babylonbee.com/thumbs/article-8813-2-thumb.jpg" :is premium="0" :path="/news/new-york-times-unwinds-claims-16
</article-card>
<article-card :comment count="134" :image path="https://media.babylonbee.com/thumb/article-8810-1-thumb.jpg" :is premium="0" :path="/news/10-reasons-homeschooling-is
</article-card>
<article-card :comment count="109" :image path="https://media.babylonbee.com/thumbs/article-8869-1-thumb.jpg" :is premium="0" :path="/news/i-do-not-study-science-1-am
</article-card>
<article-card :comment count="57" :image path="https://media.babylonbee.com/thumbs/article-8864-1-thumb.jpg" :is premium="0" :path="/news/half-black-man-ordered-to-pa
</article-card>
<article-card :comment count="80" :image path="https://media.babylonbee.com/thumbs/article-8865-1-thumb.jpg" :is premium="0" :path="/news/experts-unsure-where-joe-bid
</article-card>
<article-card :comment count="84" :image path="https://media.babylonbee.com/thumbs/article-8867-1-thumb.jpg" :is premium="0" :path="/news/blues-clues-to-be-guest-host
</article-card>
<article-card :comment count="66" :image path="https://media.babylonbee.com/thumbs/article-8776-1-thumb.jpg" :is premium="0" :path="/news/exclusive-we-have-acquired-j
```

-Chúng em xây dựng thêm các hàm để có thể lấy được dữ liệu mình cần:

```
# Dùng để gửi request đến trang web với tham số là số thứ tự trang cần thu thập
def sendRequest(page):
    #Tạo đường liên kết đến trang cần thu thập dữ liệu
    url = "https://babylonbee.com/news?page=" + str(page)
    #Gửi request đến trang đó
    request = requests.get(url)
    #BeautifulSoup dùng để phân tích dữ liệu html thành dữ liệu cây để chúng ta dễ dàng thao tác sau này
    soupSite = Soup(request.text, 'html.parser')
    return soupSite
```

```
def getData(soupSite, data):
    count = 0
    for soup in soupSite:
        soup = str(soup)

        #check Time
        a = soup.find(':published_on=')
        b = soup.find('\\" ', a, len(soup))
        if (2019 <= int(soup[b-4:b])):
            # Lấy link
            posArticleLinkStart = soup.find(":path")
            posArticleLinkEnd = soup.find('\\" ', posArticleLinkStart, len(soup))
            ArticleLink = "https://babylonbee.com/" +
soup[posArticleLinkStart+8:posArticleLinkEnd]

            # Lấy Title
            posTitleStart = soup.find(':title=') + 8
            posTitleEnd = soup.find('>', posTitleStart, len(soup))
            Title = soup[posTitleStart:posTitleEnd].replace("&quot;",
""").replace('"', "").replace("'", "'")

            #Thêm phần vừa get được vào data chính
            data.append([Title, ArticleLink, 1])
            count += 1
    return count
```

```
# Hàm dùng để xuất file json theo yêu cầu của đề bài
def writeJson(data, fileName):
    with open(fileName+'.json', 'w') as f:
        for i in data:
            f.write('{"is_sarcastic": ' + str(i[2]) + ', "headline": "' + i[0] + '",
"article_link": "' + i[1] + '"}\n')
```

```
#Thêm thư viện tqdm để theo dõi quá trình crawl dữ liệu
def crawlDataBabylonBee():
    data = []
    count = 0

    print("___ CRAWL DATA FROM BABYLON BEE ___")
    for page in tqdm(range(1, 355)):
        soupSite = sendRequest(page).find_all("article-card")
        count += getData(soupSite, data)
    writeJson(data, 'BabylonBee')

    print('Completed...')
    print("Crawled:", count, "ArticleLinks")
```

- Các trang web khác cũng tương tự như vậy nhưng thông tin sẽ lưu tên khác nhau, code chi tiết em đã nộp trong file notebook kèm theo.
- Tuy nhiên trang web [The Sun](#) có tính năng chặn request ẩn danh nên nhóm em đã thêm headers vào để có thể crawl dữ liệu

### 3.Kết quả thu được:

Sau khi crawl hoàn tất 6 trang web trên thì nhóm em thu được kết quả:

Trang web	Label	Số tập dữ liệu crawl được
<a href="#">The Babelon Bee</a>	1	4423
<a href="#">The Beaverton</a>	1	2333
<a href="#">The Chaser</a>	1	1434
<a href="#">The Guardian</a>	0	1313
<a href="#">NBC News</a>	0	49609
<a href="#">The Sun</a>	0	35922
<b>Tổng cộng:</b>		95034