

**2017-chatterjee-07**

**Scraping and extracting articles on The Atlantic Magazine website**

---

報告者：尹 佩雯

# 目錄

01 動機

02 社群案例集

03 研究方法

04 驗證結果

05 心得建議

# 動機

- 課本是去利用Scrapy的爬蟲方式爬取超跑的網路論壇《Teamspeed》，從論壇中的內容去進行進行主題分類。並將分類的主題設定為20個類別，再分別印出每個類別中的十個代表性詞語。最後將每一個類別以長條圖的方式顯示詞語出現的頻率，以及這個主題所對應到的文章是哪些。
- 原本我想爬取的是BBC News的內容，因為在第一次的報告章節中有爬取過BBC的RSS的網站，但我覺得數量不夠多，而BBC News的網站新聞排列方式也不易讓我大量爬取內容。因此，我最後選擇了《大西洋雜誌》的網站。在網站中Latest的地方，它的排序方式為由新到舊，而這個網站並沒有一個最後一頁，會一直根據時間早晚持續增加，因此，我就使用Scrapy抓取了其中的600多篇的文章去進行主題分類，並且設定的主題類別為50，並一樣去印出每個類別中的代表性詞語、以長條圖的方式呈現詞語出現的頻率以及每個主題對應的文章是哪些。

# 社群案例集

- 《大西洋》（ 英語：The Atlantic ）是美國的一本雜誌及多平台出版物，創刊於1857年的麻薩諸塞波士頓，當時名為《大西洋月刊》（ 英語：The Atlantic Monthly ），是一本文學和文化評論雜誌，發表關於廢除當代政治事務中的奴隸制、教育及其他重大問題的著名作家評論。
- 本專案是爬取大西洋雜誌網站中的文章標題、時間、作者、文章連結以及簡介，共620筆資料。

# 社群案例集



Popular

Latest

*The Atlantic*

Sign In

Subscribe

## LATEST



### The Ugly Truth About the Beautiful Game

Globalization, commercialization, and competition killed the romance of soccer—creating the best competition in the world in the process.

**TOM MCTAGUE** MAY 28, 2022



### *The Atlantic* Daily: When You Get COVID Again and Again

Our writers tackle the mysteries of reinfection and the antiviral Paxlovid. Then: Why didn't the police act faster in the school shooting in Uvalde?

**KATHERINE J. WU** MAY 28, 2022



### 78 Minutes

# 社群案例集



Popular

Latest

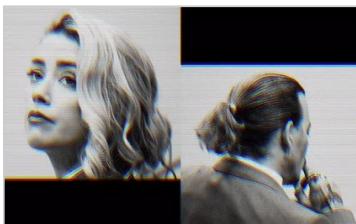
*The Atlantic*

Sign In

Subscribe



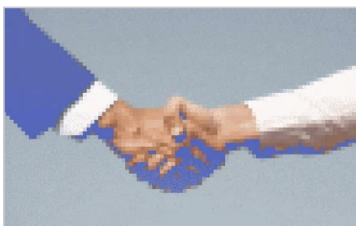
DAVID SIMS, SHIRLEY LI, AND MEGAN GARBER **MAY 27, 2022**



## The Amber Heard–Johnny Depp Trial Is Not a Joke

Why are so many people treating it like one?

MEGAN GARBER **MAY 27, 2022**



## How to Fix Twitter—And All of Social Media

The online-speech debate pretends that we must choose between absolute freedom and centralized control. Let's try something else.

JARON LANIER **MAY 26, 2022**

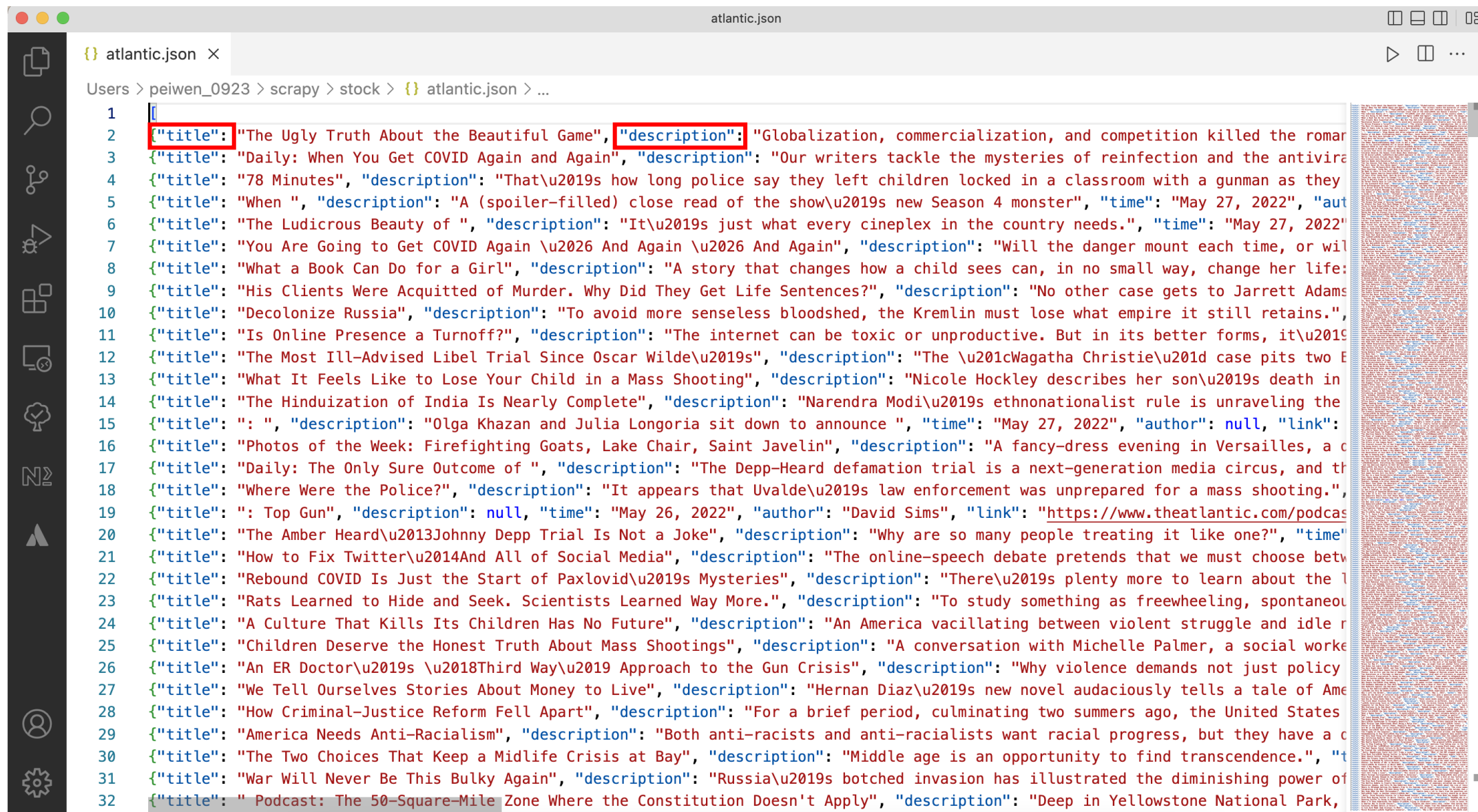
PREVIOUS

MORE STORIES

# 研究方法

- 使用的爬蟲方法為Scrapy，因為爬的網站不同，因此必須改寫全部的內容。而我爬取的內容有文章的標題、時間、作者、網址以及簡介。
- 因為文章全文的內容要付費，所以在主題分類的部分僅使用文章的標題以及簡介去進行分析。
- 在進行主題分類前，先使用NLTK的套件去將停用詞進行刪除，並且使用「punkt」及「averaged\_perceptron\_tagger」的套件去進行分詞及詞性的標注，並擷取「名詞」和「形容詞」的詞語進行分析。
- 主題分類使用的是python中的LDA套件，並使用CountVectorizer去進行特徵值的運算。
- 程式輸入的部分是dataframe、column\_name以及主題的數量，輸出的部分是每個主題的關鍵字以及主題對應的內容。

# 驗證結果



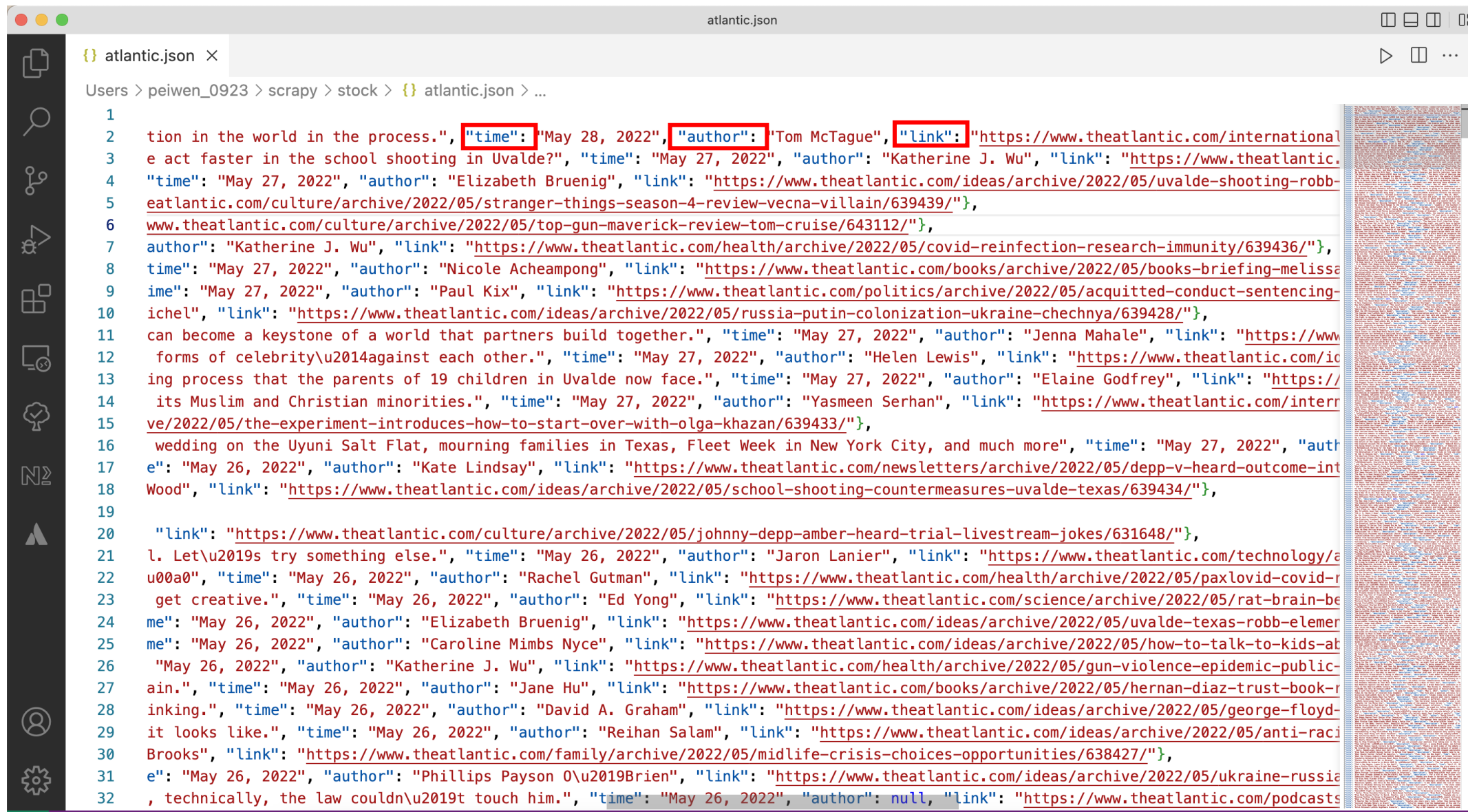
```
atlantic.json

Users > peiwen_0923 > scrapy > stock > {} atlantic.json > ...

1 [{"title": "The Ugly Truth About the Beautiful Game", "description": "Globalization, commercialization, and competition killed the roman
2 {"title": "Daily: When You Get COVID Again and Again", "description": "Our writers tackle the mysteries of reinfection and the antivira
3 {"title": "78 Minutes", "description": "That\u2019s how long police say they left children locked in a classroom with a gunman as they
4 {"title": "When ", "description": "A (spoiler-filled) close read of the show\u2019s new Season 4 monster", "time": "May 27, 2022", "aut
5 {"title": "The Ludicrous Beauty of ", "description": "It\u2019s just what every cineplex in the country needs.", "time": "May 27, 2022"
6 {"title": "You Are Going to Get COVID Again \u2026 And Again \u2026 And Again", "description": "Will the danger mount each time, or wil
7 {"title": "What a Book Can Do for a Girl", "description": "A story that changes how a child sees can, in no small way, change her life:
8 {"title": "His Clients Were Acquitted of Murder. Why Did They Get Life Sentences?", "description": "No other case gets to Jarrett Adams
9 {"title": "Decolonize Russia", "description": "To avoid more senseless bloodshed, the Kremlin must lose what empire it still retains.",
10 {"title": "Is Online Presence a Turnoff?", "description": "The internet can be toxic or unproductive. But in its better forms, it\u2019s
11 {"title": "The Most Ill-Advised Libel Trial Since Oscar Wilde\u2019s", "description": "The \u201cWagatha Christie\u201d case pits two E
12 {"title": "What It Feels Like to Lose Your Child in a Mass Shooting", "description": "Nicole Hockley describes her son\u2019s death in
13 {"title": "The Hinduization of India Is Nearly Complete", "description": "Narendra Modi\u2019s ethnonationalist rule is unraveling the
14 {"title": "": " ", "description": "Olga Khazan and Julia Longoria sit down to announce ", "time": "May 27, 2022", "author": null, "link":
15 {"title": "Photos of the Week: Firefighting Goats, Lake Chair, Saint Javelin", "description": "A fancy-dress evening in Versailles, a c
16 {"title": "Daily: The Only Sure Outcome of ", "description": "The Depp-Heard defamation trial is a next-generation media circus, and th
17 {"title": "Where Were the Police?", "description": "It appears that Uvalde\u2019s law enforcement was unprepared for a mass shooting.",
18 {"title": "": Top Gun", "description": null, "time": "May 26, 2022", "author": "David Sims", "link": "https://www.theatlantic.com/podcas
19 {"title": "The Amber Heard\u2013Johnny Depp Trial Is Not a Joke", "description": "Why are so many people treating it like one?", "time"
20 {"title": "How to Fix Twitter\u2014And All of Social Media", "description": "The online-speech debate pretends that we must choose betw
21 {"title": "Rebound COVID Is Just the Start of Paxlovid\u2019s Mysteries", "description": "There\u2019s plenty more to learn about the l
22 {"title": "Rats Learned to Hide and Seek. Scientists Learned Way More.", "description": "To study something as freewheeling, spontaneou
23 {"title": "A Culture That Kills Its Children Has No Future", "description": "An America vacillating between violent struggle and idle r
24 {"title": "Children Deserve the Honest Truth About Mass Shootings", "description": "A conversation with Michelle Palmer, a social worke
25 {"title": "An ER Doctor\u2019s \u2018Third Way\u2019 Approach to the Gun Crisis", "description": "Why violence demands not just policy
26 {"title": "We Tell Ourselves Stories About Money to Live", "description": "Hernan Diaz\u2019s new novel audaciously tells a tale of Ame
27 {"title": "How Criminal-Justice Reform Fell Apart", "description": "For a brief period, culminating two summers ago, the United States
28 {"title": "America Needs Anti-Racism", "description": "Both anti-racists and anti-racialists want racial progress, but they have a c
29 {"title": "The Two Choices That Keep a Midlife Crisis at Bay", "description": "Middle age is an opportunity to find transcendence.", "t
30 {"title": "War Will Never Be This Bulky Again", "description": "Russia\u2019s botched invasion has illustrated the diminishing power of
31 {"title": "Podcast: The 50-Square-Mile Zone Where the Constitution Doesn't Apply", "description": "Deep in Yellowstone National Park,
```



# 驗證結果

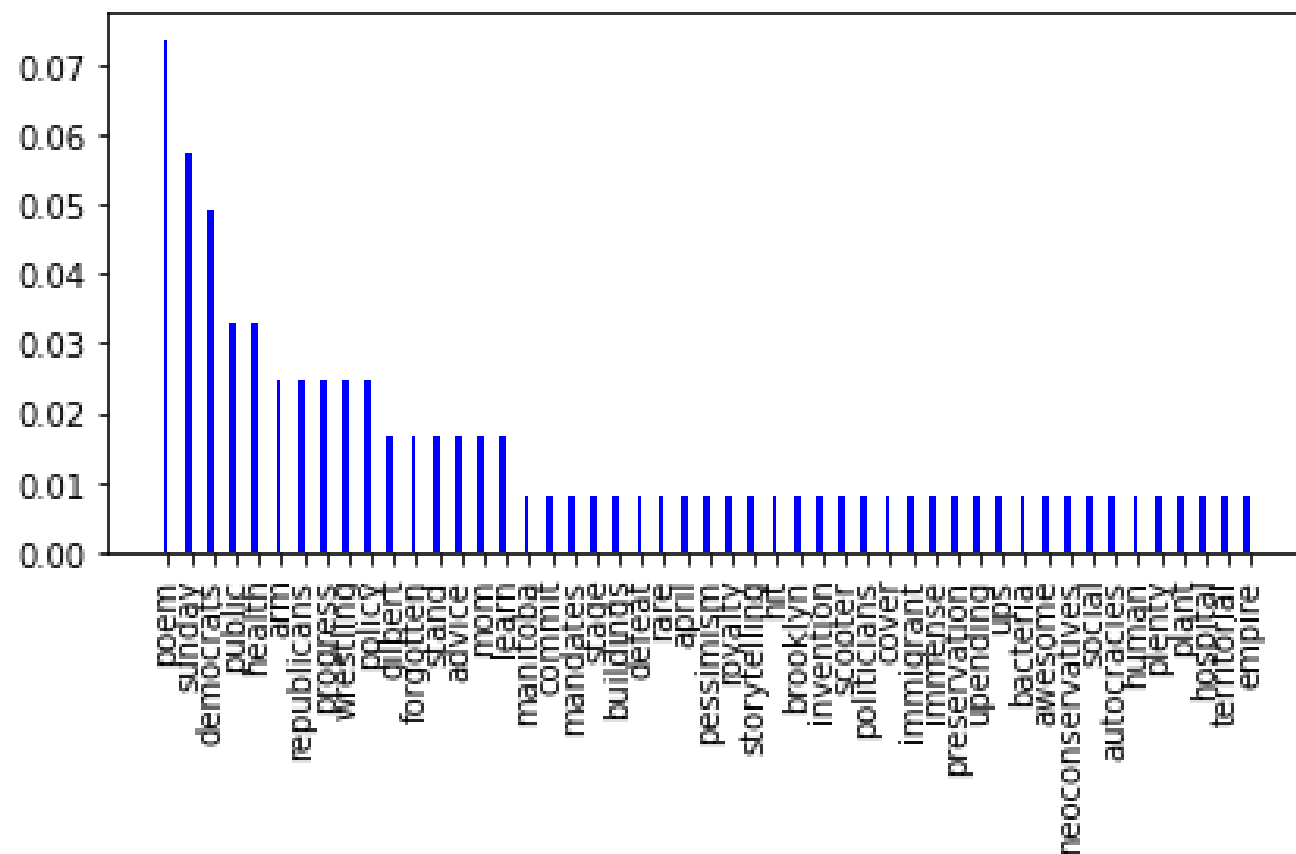


```
atlantic.json
Users > peiwen_0923 > scrapy > stock > {} atlantic.json > ...
1
2 tion in the world in the process.", "time": "May 28, 2022", "author": "Tom McTague", "link": "https://www.theatlantic.com/international
3 e act faster in the school shooting in Uvalde?", "time": "May 27, 2022", "author": "Katherine J. Wu", "link": "https://www.theatlantic.
4 "time": "May 27, 2022", "author": "Elizabeth Bruenig", "link": "https://www.theatlantic.com/ideas/archive/2022/05/uvalde-shooting-robb-
5 eatlantic.com/culture/archive/2022/05/stranger-things-season-4-review-vecna-villain/639439/"},
6 www.theatlantic.com/culture/archive/2022/05/top-gun-maverick-review-tom-cruise/643112/"},
7 author": "Katherine J. Wu", "link": "https://www.theatlantic.com/health/archive/2022/05/covid-reinfection-research-immunity/639436/"},
8 time": "May 27, 2022", "author": "Nicole Acheampong", "link": "https://www.theatlantic.com/books/archive/2022/05/books-briefing-melissa
9 ime": "May 27, 2022", "author": "Paul Kix", "link": "https://www.theatlantic.com/politics/archive/2022/05/acquitted-conduct-sentencing-
10 ichel", "link": "https://www.theatlantic.com/ideas/archive/2022/05/russia-putin-colonization-ukraine-chechnya/639428/"},
11 can become a keystone of a world that partners build together.", "time": "May 27, 2022", "author": "Jenna Mahale", "link": "https://www
12 forms of celebrity\u2014against each other.", "time": "May 27, 2022", "author": "Helen Lewis", "link": "https://www.theatlantic.com/ic
13 ing process that the parents of 19 children in Uvalde now face.", "time": "May 27, 2022", "author": "Elaine Godfrey", "link": "https://
14 its Muslim and Christian minorities.", "time": "May 27, 2022", "author": "Yasmeen Serhan", "link": "https://www.theatlantic.com/interr
15 ve/2022/05/the-experiment-introduces-how-to-start-over-with-olga-khazan/639433/"},
16 wedding on the Uyuni Salt Flat, mourning families in Texas, Fleet Week in New York City, and much more", "time": "May 27, 2022", "auth
17 e": "May 26, 2022", "author": "Kate Lindsay", "link": "https://www.theatlantic.com/newsletters/archive/2022/05/depp-v-heard-outcome-int
18 Wood", "link": "https://www.theatlantic.com/ideas/archive/2022/05/school-shooting-countermeasures-uvalde-texas/639434/"},
19
20 "link": "https://www.theatlantic.com/culture/archive/2022/05/johnny-depp-amber-heard-trial-livestream-jokes/631648/"},
21 l. Let\u2019s try something else.", "time": "May 26, 2022", "author": "Jaron Lanier", "link": "https://www.theatlantic.com/technology/e
22 u00a0", "time": "May 26, 2022", "author": "Rachel Gutman", "link": "https://www.theatlantic.com/health/archive/2022/05/paxlovid-covid-r
23 get creative.", "time": "May 26, 2022", "author": "Ed Yong", "link": "https://www.theatlantic.com/science/archive/2022/05/rat-brain-be
24 me": "May 26, 2022", "author": "Elizabeth Bruenig", "link": "https://www.theatlantic.com/ideas/archive/2022/05/uvalde-texas-robb-elemer
25 me": "May 26, 2022", "author": "Caroline Mimbs Nyce", "link": "https://www.theatlantic.com/ideas/archive/2022/05/how-to-talk-to-kids-at
26 "May 26, 2022", "author": "Katherine J. Wu", "link": "https://www.theatlantic.com/health/archive/2022/05/gun-violence-epidemic-public-
27 ain.", "time": "May 26, 2022", "author": "Jane Hu", "link": "https://www.theatlantic.com/books/archive/2022/05/hernan-diaz-trust-book-r
28 inking.", "time": "May 26, 2022", "author": "David A. Graham", "link": "https://www.theatlantic.com/ideas/archive/2022/05/george-floyd-
29 it looks like.", "time": "May 26, 2022", "author": "Reihan Salam", "link": "https://www.theatlantic.com/ideas/archive/2022/05/anti-raci
30 Brooks", "link": "https://www.theatlantic.com/family/archive/2022/05/midlife-crisis-choices-opportunities/638427/"},
31 e": "May 26, 2022", "author": "Phillips Payson O\u2019Brien", "link": "https://www.theatlantic.com/ideas/archive/2022/05/ukraine-russia
32 , technically, the law couldn\u2019t touch him.", "time": "May 26, 2022", "author": null, "link": "https://www.theatlantic.com/podcasts
```

# 驗證結果

- 總共分為50類，舉其中兩個類別來當作驗證結果的觀察：

Topic 0 :



# 驗證結果

Topic 0:

24 : An ER Doctor's 'Third Way' Approach to the Gun Crisis Why violence demands not just policy solutions, but public-health ones

38 : At the Graveyard With Anne A **poem** for Wednesday

134 : Why Interview an Autocrat? Readers respond to our April 2022 cover story and more.

153 : Desire A **poem** for Sunday

155 : How Public Health Failed America The U.S. clearly failed to heed expert advice, but there's plenty of blame to go around.

196 : The Forgotten Stage of Human Progress Invention is easily overrated, and implementation is often underrated.

204 : The Essential Sophie Gilbert Reading List A staff writer at

222 : In the Hospital Rooms of My Country A **poem** for Sunday

278 : What Historic Preservation Is Doing to American Cities Laws meant to safeguard great buildings and neighborhoods can also present an obstacle to social progress.

286 : Why I Left the Garden A **poem** for Sunday

353 : Father Andrews A **poem** by Thomas Lynch, published in

381 : 10 Practical Ways to Improve Happiness For when you need advice that goes beyond "Be Danish"

384 : What Masks Off on Public Transit Means for the Pandemic Why is this mandate different from all other mandates?

391 : Two Nights in the Brooklyn Arm-Wrestling Scene You may think you could be good at arm wrestling, but you would probably be really bad at arm wrestling.

414 : First **Poem** After Parting None

447 : Gilbert Gottfried Was More Than Just a Funny Voice The late comedian helped move stand-up beyond the realm of the merely observational and create space for the absurd.

474 : The Necklace A **poem** for Sunday, published in

544 : Mysterium Lunae A **poem** for Sunday

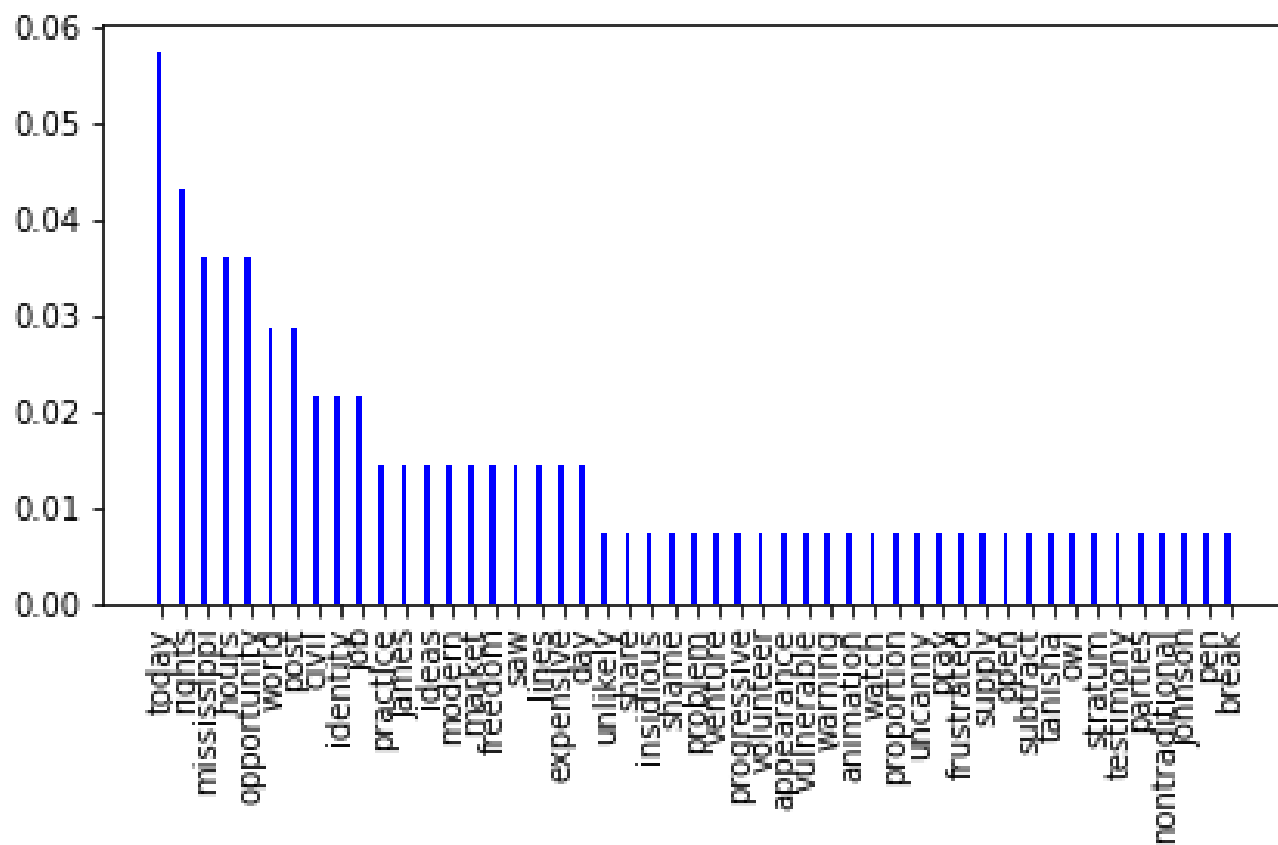
551 : The 'Putin Is Bad, But' Republicans At a recent conference, some members of the right identified foreign-policy hawks and neoconservatives, rather than Democrats, as their biggest enemies.

560 : Who Moved My Seed? A rare animal found a rare plant. Then, it seems, the two teamed up.

613 : Goodbye Letter to My Lover's Wife A **poem** for Sunday

# 驗證結果

Topic 2 :



# 驗證結果

Topic 2:

59 : What Is Life Like When We Subtract Work From It? Sabbaticals can give people an invaluable opportunity to rest and reflect on their identity beyond their job.

106 : The U.S. Housing Market Has Peaked But no, we're not headed for anything even close to 2008.

107 : Podcast: Fighting to Remember Mississippi Burning At the height of the Freedom Summer, the Ku Klux Klan killed three civil-rights workers in Philadelphia, Mississippi. Now the reporter Ko Bragg searches for memories in a town that would rather forget.

129 : The Problem With Wills A striking proportion of Americans doesn't have one. Nontraditional families are left uniquely vulnerable.

255 : What Alito Got Right The Court's job is not to determine which rights we

265 : The Day the Civil-Rights Movement Changed What my father saw in Mississippi

298 : Let Coach Kennedy Pray In

379 : 'Workcations' Aren't an Escape. They're Practice. Think of them as an opportunity to break bad work-life-balance habits.

426 : The Post-villain Era of Animation Today's Disney movies aren't like the ones you grew up with. That isn't a bad thing.

445 : Today Your Phone Became a Police Radio Roughly four hours after an unusual push-alert dragnet, Frank James was captured. Did policing just change?

503 : Burying a Burning The killing of three civil-rights workers in Neshoba County, Mississippi, in 1964 changed America. But today, if you want to know what happened here, you need to know who to ask.

575 : Needed an Update The original is "a masterpiece ... but it's also a bit of a tough watch today."

589 : What Trump Is Hiding Seven hours and 37 minutes of missing phone records on January 6 suggest consciousness of guilt.

# 驗證結果

- 可能的應用：

雖然在大西洋雜誌的網站已經有針對文章做一個分類，像是政治、科技、商業等非常多不同的分類，而這次專案用的文章是依照時間去做排序的，進而將文章以內容去做區分，但我認為這個網站可以再針對大的分類再去做細部的分類，像是政治類的文章可以再去依照國家去做區分等，將每一個分類都利用LDA的主題分類去做更細部的分類，讓使用者在瀏覽網站時，可以更迅速的瀏覽到想觀看的內容。



# 心得建議

- 雖然這次期末專案跟在報告這個章節是差不多的概念，先用爬蟲去爬取網站的內容，再根據內容去用LDA進行主題分類。但上次爬的網站除了使用到scrapy，還額外使用到request的套件去進行爬蟲，所以在爬取大量的資料時，需要花的時間比較長。這次的期末專案只有單純使用到scrapy去進行，我發現這次爬的資料量比上次多很多，但爬取的速度也非常的快，很快就把幾百筆的資料爬完了。
- LDA的主題建模部分，之前其實也有接觸過，但之前的做法是自己使用文章進行模型的訓練，並沒有直接使用過python內部的套件去做使用，所以也是透過這堂課報告的機會，讓我可以了解到原來有其他方法可以使用，也相對更方便一些。
- 在準備這堂課的報告時，雖然有些部分之前有接觸過一點點，但是之前並不知道原來還有很多方法可以用，像是爬蟲，我之前只有使用過request去對網站進行大量爬取，常常要等待的時間就會非常久，

# 心得建議

並不知道還有其他更快速的方法可以使用，在修改程式碼的時候，也更熟悉了一些python 的套件，像是NLTK也是過去沒有接觸過的，在三次的報告中都有利用到這個套件，我覺得NLTK的套件非常方便，除了可以下載一些資料集，也可以去對句子進行斷句或是擷取特定詞性等，在自然語言處理上有非常多可以使用的東西。

- 這堂課的三次報告我都是選擇了與自然語言處理相關的章節，透過一些文字的分析，讓我對於文字處理的部分有更深入的了解，也讓我之後會想對自然語言處理這塊做更近一步的探索。