Q: 發現最後一個HN通常不會有">"的內容

A: 所以將下界設一個情境: if 最後一個HN沒有 ">"內容, find 下一個包含">"內容的 HN

以下為PDF範例:

HN18 .....

.......

Antitrust & Trade Law > Clayton Act > Remedies > Injunctions

HN19[ ] Remedies, Damages

Under the Clayton Act, a plaintiff is entitled to an injunction to prevent threatened loss or damage by a violation of

the antitrust laws. 15 U.S.C.S. § 26 (1982). Injunctive relief is appropriate, even though the illegal practice has been

seasonally discontinued, unless it is clear that the practice will not recur.

Counsel: [\*\*1] Gael Mahony, Hill & Barlow, Boston, Massachusetts, Thomas J. Dacey, esq., for Plaintiff.

通常最後一個HN內容裡面不會有">"的內容, 且他們的格式不太統一, 有些HN結尾後面是接Counsel的小標題 有些是 Opinion的小標題, 所以統一用大寫的方式去區隔他們, 可以用一樣的方式修正正則表達式 , 確保每一個HN都有被檢測到有沒有">"的內容, 且不會去抓到除了這個範圍以外的內容

origin:

hn\_matches = re.findall(r'HN\d+\[|[^.]+>.\*?(?=HN\d+|$)|(?<=\>).\*?(?=Lawyers\' Counsel or Judges)', text, re.DOTALL)

新的發現: 也許是因為 Headnotes也有">"內容 導致抓錯

A:

V1:

hn\_matches = re.findall(r'HN\d+\[|[^.]+>.\*?(?=HN\d+|$)|(?<=\>)(?=\n[A-Z][^\s]\*\s|$)', text, re.DOTALL)

V2:

hn\_matches = re.findall(r'HN\d+\[[^\]]\*\]|\>[^\n]\*(?=\nHN\d+|$)', text, re.DOTALL)

抓對了,但是跨頁的問題出現

46 HN2[ Antitrust & Trade Law > Robinson-Patman Act > General Overview Antitrust & Trade Law > Regulated Practices > Price Discrimination > General Overview

47 HN4[ Civil Procedure > ... > Summary Judgment > Supporting Materials > General Overview Civil Procedure > Judgments > Summary Judgment > Partial Summary Judgment

中間還有一個HN3 跨頁了但沒抓到

Q: HN有跨頁沒抓到的問題

A:先將PDF裡面的文本取出, 再進行抓取分類, 最後再幫他們標上頁碼

Q: 發現CoreTerms \ Headnotes \ HN 三者會互相干擾

A:設立幫助函數 讓他們獨立進行

-----------------------------------------

Q: District p.287 , p303只有Core Terms

A: 後面有接Counsel所以不影響)

Q: 像是District p.8 只有Core Terms的情況 後面是接 Opinion

A: 可以用大寫的方式固定下界 因為內容都是小寫的, 下一個大寫的就是另一個小標題

設立一個情境為: 抓CoreTerms的內容直到遇到下一個大寫字元則停, 尋找下一個CoreTerms

這邊是我從PDF擷取一小段範例:

Core Terms

manufacturers, retailers, co-operation, style, dress, garments, copies, percent, piracy, prices, declaration, anti-trust,

pirating, regulations, ladies', evils, purposes, merchandise, registered, measures, restrain, by-laws, selling,

preliminary injunction, interstate commerce, original design, conditions, injunctive, originator, provisions

Opinion by: [\*\*1] BREWSTER

原本re：

core\_terms\_match = re.search(r'Core Terms\s+(.+?)(?=LexisNexis® Headnotes|Counsel|$)', text, re.S)

可以看到中間的內容為" manufacturers, ...... 到 ....,provisions " 然後我的意思是遇到像 Opinion 的 "O" 就停下來, 這樣就能完整抓到中間我要的Core Terms內容 Opinion是我不要的

ｖ１：

core\_terms\_match = re.search(r'Core Terms\s+(.+?)(?=[A-Z][^\s]\*\s|$)', text, re.S)

應注意Core Terms 裡面可能會有專有名詞大寫的地方

以District p303為例

Core Terms

gasoline, customers, rental, Deposition, products, rented, tying arrangement, defendants', anti trust law, summary

judgment, retail, tied product, car rental, antitrust, tank, buy

Counsel: [\*\*1] Celso E. Lopez, San Sebastian, Puerto Rico, and Carlos F. Lopex, Santurce, Puerto Rico, for

plaintiff

抓到rental,就停了 但是後面的內容是Core Terms的 但沒抓到, 定義大寫的部分 應該要在 結尾空白處後 也就是 ...buy 後

ｖ２：

core\_terms\_match = re.search(r'Core Terms\s+(.+?)(?=\n[A-Z][^\s]\*\s|$)', text, re.S)

Q:

Core Terms 的部分若逗號後面為大寫英文字母可能就停了 但後面還有內容

這是產出CSV:

157 Core Terms purchasers, overcharges, joinder, indirect, parties, damages, multiple liabilities, antitrust, chain of distribution,

這是原PDF 157頁 Core Terms該有的內容:

Core Terms

purchasers, overcharges, joinder, indirect, parties, damages, multiple liabilities, antitrust, chain of distribution,

Cartwright Act, duplicative, federal action, industrial, joined, substantial risk, indirectly, fraudulent concealment,

intermediate, conspiracy, consumers, potential plaintiff, treble damages, common fund, claimants, suits, statute of

limitations, anti trust law, petitioners', alleges, chain

這是目前的正則表達式:

process\_section(r'Core Terms\s+(.+?)(?=\n[A-Z][^\s]\*\s|$)', "Core Terms")

V1:

process\_section(r'Core Terms\s+([\s\S]+?)(?=\n[A-Z][^\s]\*\s|$)', "Core Terms")

V2:

process\_section(r'Core Terms\s+([\s\S]+?)(?=\n[A-Z]+\s)', "Core Terms")

Q:

若逗號後面換行接大寫字母內容就會被視為結尾無法抓到後面的內容

Q:

然後遇到像是 "LexisNexis® Headnotes" or "Counsel" 這樣大寫開頭的小標題 就代表結束 尋找下一個Core Terms內容

結尾的正則表達式需表達

若結尾為空白字符且下一個字串為大寫字母開頭的小標題表示結束

process\_section(r'Core Terms\s+([\s\S]+?)(?=(?:,?\s\*\n[A-Z][^\s]\*\s|$))', "Core Terms")

V3:

process\_section(r'Core Terms\s+([\s\S]+?)(?=(?:\s\*\n[A-Z]+\s)|(?:\s\*$))', "Core Terms")

V4:

process\_section(r'Core Terms\s+([\s\S]+?)(?=\n[A-Z][^\s]\*(?:\s[A-Z][^\s]\*)\*[:\n])', "Core Terms")

Q:

會擷取到Opinion by :

這是幾種後面會出現的小標題的形式 : "LexisNexis® Headnotes" or "Opinion by: [\*\*1] BREWSTER" or "Counsel: [\*1]" or "Judges: [\*1] WILLIAM H. FOLLETT, J.

" or "Opinion"

V5:

在正則表達式後面多加一個結尾遇到Opinion by: 的例外

process\_section(r'Core Terms\s+([\s\S]+?)(?=\n(?:[A-Z][^\s]\*(?:\s[A-Z][^\s]\*)\*[:\n]|Opinion by:))', "Core Terms")

--------------------------------------------------------------------------------------------------------------

加入計數器看產出的CSV中 Core Terms , Headnotes , HN1 有幾個跟PDF有沒有相符合

--------------------------------------------------------------------------------------------------------------

Circuit : 24 HN2 , 29HN3 抓到31 Headnotes的內容

51 HN14[ Antitrust & Trade Law > Regulated Practices > Trade Practices & Unfair Competition > General Overview

56 HN1[ Antitrust & Trade Law > Regulated Practices > Trade Practices & Unfair Competition > General Overview

像是 51頁的HN14 內容為 56頁的 Headnotes內容 也就是 HN1所在的頁數 , 其他也是一樣 都會抓到下一段Headnotes的內容 但那應該是歸類在Headnotes 的 Section 不應該跑來 HN Label的Section

hn\_matches = re.findall(r'HN\d+\[.\*?\].\*?(?=HN\d+|\Z)', all\_text, re.DOTALL)

V1:

hn\_matches = re.findall(r'HN\d+\[(.\*?)\](.\*?)((?=HN\d+\[)|(?=HN\d+$)|\Z)', all\_text, re.DOTALL)

多出來的問題 , 會在下一個段落開始前 抓到下一個段落Headnotes的內容分類進去當前HN部分

已解決:

hn\_matches = re.findall(r'HN\d+\[.\*?\].+?(?=HN\d+|LexisNexis® Headnotes|\Z)', all\_text, re.DOTALL)

加入Headnotes 防止Headnotes的內容被誤以為是HN的內容