Opinion

Q:

出現跨頁漏抓的問題, 先提取PDF文本再進行抓取目標文本並標上頁碼, 解決正則表達式只能單頁處理的問題，跨頁的時候偶爾會出現不重要的字樣

已解決:

all\_text = re.sub(r'Page \d+ of \d+', '', all\_text)

all\_text = re.sub(r'\d+ Cal\..\*?;.\*?Cal\. LEXIS \d+,\s+\\*\\*\\*\\*\d+', '', all\_text)

Q:這是幾種可能會出現的:

147 Cal. Rptr. 410, \*410; 1978 Cal. App. LEXIS 1712, \*\*1

Cal. App. 3d 381, \*381; 162 Cal. Rptr. 487, \*\*487; 1980 Cal. App. LEXIS 1496, \*\*\*1

102 Cal. App. 3d 381, \*381; 162 Cal. Rptr. 487, \*\*487; 1980 Cal. App. LEXIS 1496, \*\*\*1

16 Cal. 3d 920, \*920; 549 P.2d 833, \*\*833; 130 Cal. Rptr. 1, \*\*\*1; 1976 Cal. LEXIS 268, \*\*\*\*1

41 Cal. 3d 460, \*460; 715 P.2d 564, \*\*564; 224 Cal. Rptr. 605, \*\*\*605; 1986 Cal. LEXIS 324, \*\*\*\*1

46 Cal. 3d 1147, \*1147; 762 P.2d 385, \*\*385; 252 Cal. Rptr. 221, \*\*\*221; 1988 Cal. LEXIS 243, \*\*\*\*1