RAG 文本切塊與檢索評估報告

HW Day3 — 學號 14112S2019

Q1：固定大小切塊和滑動視窗切塊的參數設定多少？

一、固定大小切塊（Fixed-size Chunking）

使用 LangChain 的

CharacterTextSplitter，以純字元數進行切割，不考慮語意邊界，可能在句子中間截斷。

參數	設定值	說明
chunk_size	300	每個切塊最多 300 個字元
chunk_overlap	0	切塊之間無重疊
separator	""（空字串）	不使用分隔符，純粹按字數切割

二、滑動視窗切塊（Sliding Window Chunking）

使用 LangChain 的 RecursiveCharacterTextSplitter，按照中文語意邊界進行遞迴切割，並透過 overlap 保留上下文。

參數	設定值	說明
chunk_size	300	每個切塊最多 300 個字元
chunk_overlap	100	相鄰切塊重疊 100 個字元
separators	"\n\n","\n","。","！","？","；","，",""	優先在段落、句號等語意邊界處切割

三、語意切塊（Semantic Chunking）

先按中文句號切分成句子，再透過 Embedding API（維度

4096）計算相鄰句子的餘弦相似度，當相似度低於門檻時斷開，形成語意段落。過長段落再以 RecursiveCharacterTextSplitter 細切。

參數	設定值	說明
similarity_threshold	0.5	相鄰句子相似度低於此值時斷開
max_chunk_length	500	超過 500 字的語意段落再細切
sub_chunk_size	400	細切時的 chunk_size
sub_chunk_overlap	50	細切時的 overlap

四、各方法切塊數量統計

資料檔案	字元數	固定大小	滑動視窗	語意切塊

data_01.txt	8,038	27	40	46
data_02.txt	2,189	8	11	10
data_03.txt	1,109	4	5	6
data_04.txt	2,178	8	11	16
data_05.txt	5,206	18	28	26
合計	18,720	65	95	104

Q2：哪一種切塊方法效果最好（平均值分數最高）？

一、各方法平均分數比較

切塊方法	平均分數	切塊總數	排名
固定大小（Fixed-size）	0.726049	65	第 3 名
滑動視窗（Sliding Window）	0.750550	95	第 2 名
語意切塊（Semantic）	0.760141	104	第 1 名

結論：語意切塊（Semantic Chunking）效果最好，平均分數 0.760141，高於滑動視窗 0.750550 和固定大小 0.726049。

二、20 題各方法分數明細

Q	題目摘要	固定大小	滑動視窗	語意切塊
1	校正流程分為哪兩種？	0.7028	0.6638	0.6874
2	衛福部長石崇良首波改革方向	0.6437	0.6633	0.7975
3	伊朗2025年大規模抗爭經濟與環境主因	0.7276	0.8030	0.7475
4	「復辟巴勒維王朝」口號背後民意	0.7993	0.8162	0.8428
5	何謂 OSINT？運作方式為何？	0.6482	0.7378	0.7331
6	衛星影像監測油槽可分析什麼情報？	0.7044	0.7842	0.7448
7	美國與海灣盟邦對軍事介入伊朗保留態度	0.7137	0.7376	0.7410
8	二戰後全球核試驗次數與地點特徵	0.7422	0.7891	0.7850
9	守望亞洲計畫功能與應用舉例	0.7902	0.7993	0.8417
10	健保兆元時代面臨哪些挑戰？	0.6618	0.6605	0.6824
11	伊朗2026年1月鎮壓的具體手段	0.7376	0.7472	0.7209
12	福衛九號的技術特色	0.7398	0.7612	0.8103
13	魯尼特圓頂目前面臨什麼危機？	0.6716	0.6887	0.6459
14	福衛八號自主研發的突破	0.7510	0.7693	0.7978
15	六大皆空是指哪六個科別？	0.7549	0.7573	0.7963
16	為何需自製衛星而非向美國購買？	0.7276	0.8227	0.7732
17	何謂「直美（Chokubi）」現象？	0.7752	0.7841	0.8264
18	鍶-90 與鈽-239 對人體的危害	0.6899	0.7045	0.6705

19	2026年為何是核武威脅關鍵時間點？	0.7682	0.7618	0.7910
20	福衛八號星系對國安監測能力提升	0.7713	0.7595	0.7675

三、各題最佳方法分佈統計

最佳方法	獲勝題數	佔比
語意切塊	10 題	50%
滑動視窗	8 題	40%
固定大小	2 題	10%

四、分析與討論

（1）語意切塊為何表現最佳？

語意切塊透過 Embedding
計算相鄰句子的餘弦相似度，在語意轉換處斷開，每個切塊內部語意連貫性最高。例如
Q2（衛福部改革方向）語意切塊得分 0.7975，遠高於固定大小的 0.6437，因為語意切塊將「改革方向」相關句子完整保留在同一切塊中。Q9（守望亞洲計畫）語意切塊得分
0.8417，也是三種方法中最高。

（2）滑動視窗為何優於固定大小？

滑動視窗使用中文語意邊界（句號、問號等）作為分隔符，並透過 100 字元的 overlap
保留上下文。不會在句子中間截斷，語意完整性較好。例如 Q16（為何需自製衛星）滑動視窗得分 0.8227 為該題最高，因為 overlap 恰好將前後文的關鍵論述連接起來。

（3）固定大小的限制

固定大小無 overlap、不考慮語意邊界，每 300
字元切一刀，導致許多切塊在句子中間斷開。優勢在於速度快、實作簡單，適合快速原型驗證，但在 RAG 檢索精度上不如另外兩種方法。

五、總結

在本次作業的 5 份資料、20 題檢索任務中，語意切塊以平均分數 0.760141 表現最佳，其次為滑動視窗（0.750550），固定大小（0.726049）最低。語意切塊的優勢在於能根據文本內容的語意變化自適應切割，保留完整的語意段落，使檢索結果更精準地匹配使用者問題。不過語意切塊需要額外的 Embedding API 呼叫，在處理速度與成本上需要權衡。在實際應用中，建議根據資料特性與效能需求選擇合適的切塊策略。