Duplicate Note (2017/12/6)

2017年12月6日　　上午 11:21

Duplicate 處理程序

整理好duplicate group, 每一個 group包含**兩個**或**兩個以上**(發生串接)的 resource ids, 且**只有一個 *rid代表*, 其餘都是要 *rid被取代***;
/* 若發生group內沒有*rid代表*, 請記錄這些rids, 我將回報UCSD */

For each group
For each *rid被取代*

// 新增mention資訊
一一檢視 *rid被取代* 出現的pmid
若該pmid沒有出現在*rid代表*的pmid集合內, 新增該筆mention紀錄

// 新增 resource co-mention的資訊
一一檢視與 *rid被取代* co-mention的 *rid跟被取代co-mention*
若 *rid跟被取代co-mention* == *rid代表*
// 不動作 應該能在*rid代表*的co-mention resources內找到*rid被取代*

若 *rid跟被取代co-mention* 不曾出現在 *rid代表*的co-mention
將 所有*rid跟被取代co-mention*與*rid被取代* 共同出現的pmid 新增成*rid代表* 與*rid跟被取代co-mention*共同出現的pmid (留意co-mention count也要 update)

若 *rid跟被取代co-mention* 曾出現在 *rid代表*的co-mention
將*rid跟被取代co-mention*與*rid被取代*下未曾出現在 *rid代表* 與*rid跟被取代co-mention*的pmid 進行新增 (留意co-mention count也要 update)

End for each *rid被取代*

/* 上半部只有新增 mention, co-mention資料,
尚未將多餘的資料清除,
所以最後須清掉所有*rid被取代* 的mention, co-mention資料
此外, 若*rid代表*內有跟*rid被取代*有co-mention pmids, 也請一併清除
注意: 一定要處理完所有*rid被取代*的新增動作後再來清理
不然可能會清不乾淨
如 SCR_1是代表id, SCR_2 and SCR_3是被取代id
若先清掉SCR_2後再來處理 SCR_3的新增,
則有可能SCR_3有跟SCR_2 co-mention而不小心又新增了SCR_2 與 SCR_1的co-mention,
造成資料沒清乾淨 */
Do duplicate data cleaning

End for each group