RK-A - Task #1726

Handling cases of bad header variants like "representation"

10/11/2021 05:42 AM - Nandini Bansal

Status:	Resolved	**Start date:**	10/13/2021
Priority:	Normal	**Due date:**	
Assignee:		**% Done:**	0%
Category:		**Estimated time:**	5.00 hours
Target version:	P1	**Spent time:**	0.00 hour

Description

In BR3_IR3_tagger.py, we have a function called **variations_in_common_section_words** that strips all the common words (extracted from the dataset) from the beginning and end of the header variant to generate new header variants. While most of the header variants generated are good, there are some bad cases like "representation" which do not necessarily result in good KPs. We need to eliminate these cases.

To do so, we can add a CW filter of 4K words. If the header variant generated is a single word variant and it lies within the 4K CW list, we save it with ignore_flag = True.

Let us first of all look at header variants that will be deleted using this filter for the Library Reference book and analyze if it's okay for us to lose them. If they look good, we can remove them and test the changes on the full book.

Subtasks:

Bug # 1743: Checking singular and plural forms of the tmp_var from variations_in_common...	**Resolved**
Bug # 1744: Calculating the fullness_ratio of the header variants to decide a threshold...	**Resolved**

History

#1 - 10/13/2021 06:12 AM - Nandini Bansal

- Estimated time changed from 5.00 h to 48.00 h

Estimate time increased as we are stuck with some cases that are difficult to manage

#2 - 10/13/2021 06:36 AM - Nandini Bansal

- Estimated time changed from 48.00 h to 5.00 h

#3 - 10/13/2021 12:21 PM - Anonymous

- Status changed from New to In Progress

#4 - 10/28/2021 05:55 AM - Anonymous

- Status changed from In Progress to Resolved