RK-A - Feature #1593

Eliminate certain header variants generated from variations_in_common_section_words

09/01/2021 01:38 PM - Nandini Bansal

Status:	Rejected	**Start date:**	09/01/2021
Priority:	Normal	**Due date:**	
Assignee:		**% Done:**	0%
Category:		**Estimated time:**	4.00 hours
Target version:	P1	**Spent time:**	0.00 hour

Description

For cases where two-word header_var is entirely made up of NOUNs/PROPNs and a single word tmp_variant is generated and exists within 20K common words, do not save the header variant. These changes are to be made within the variations_in_common_section_words function of BR3_IR3_tagger.py.

This is a testing change and there is no surety of good results. To evaluate the impact of changes, we need to re-generate the master_cands.pkl, run the entire annotation, save the similar_docs.csv and perform the KPI exercise on it. Based on the impact reflected in the KPI stats and manual inspection of annotated text files, we will decide whether this change is desirable or not.

Datasets to test with:
1. Whirlwind Book
2. Library Reference
3. Tutorial Book

NOTE: While checking whether tmp_variant is present within 20K CW, you need to check both singular and plural forms of the KP.

History

#1 - 11/01/2021 05:25 AM - Nandini Bansal

- Status changed from New to Rejected