RK-A - Bug #1514

Removing key phrases which are starting with IN pos tag or a preposition

08/19/2021 12:49 PM - Nandini Bansal

Status:	Resolved	**Start date:**	08/20/2021
Priority:	Normal	**Due date:**	
Assignee:	Rohit Choudhary	**% Done:**	100%
Category:		**Estimated time:**	3.00 hours
Target version:		**Spent time:**	0.00 hour

Description

There are some cases of key phrases such as:

1. "within regular expressions" linked to "regular expressions"
2. "before python" linked to "before python initialization"
3. "since unicode strings" linked to "unicode strings"
4. "as a context manager" linked to "connection as a context manager"

These are bad looking keyphrases and should not be present. The common pattern followed by all these is first word with ADP/IN pos tag.

The changes are to be made in the **pos_filter_cands** method in **BR3_IR3_tagger.py**. We have to add a filter to discard key phrases which are not exactly as same as the header variant and starts with an ADP/IN word. For verifying whether the key phrase and header variant are equivalent, you can use the stemmed forms. Use **global_stem_dict** to access pre-calculated stemmed forms of some keyphrases and header variants.

Test the changes with the following books:
1. Python Whirlwind Tour.txt
2. Python Tutorial.txt
3. Python 3 - Library Reference.txt

URL:
https://edutestdev-240612.appspot.com/document/python-whirlwind-tour/m?documentURL=10054%2Fds9aug1528%2FWhirlwindTourOfPython%2F14-strings-and-regular-expressions-Special-characters-can-match-character-groups-94.html
The above URL has "within regular expressions" tagged as a purple link (or PL).

Subtasks:

Bug # 1520: Instead of removing entire key phrase starting with IN pos tag for all case...	**Closed**

History

#1 - 08/23/2021 11:51 AM - Nandini Bansal

- Assignee set to Rohit Choudhary

#2 - 10/20/2021 07:14 AM - Rohit Choudhary

- Status changed from New to Resolved