

## RK-A - Feature #1856

### Penalise the multi-word KP based on "Past" tense & vector similarity score

11/03/2021 08:39 AM - Nandini Bansal

|   |          |                        |            |
|---|----------|------------------------|------------|
| <b>Status:</b>  | Rejected | <b>Start date:</b>     | 11/03/2021 |
| <b>Priority:</b>  | Normal   | <b>Due date:</b>       |            |
| <b>Assignee:</b>  |          | <b>% Done:</b>         | 0%         |
| <b>Category:</b>  |          | <b>Estimated time:</b> | 2.50 hours |
| <b>Target version:</b>  | P1       | <b>Spent time:</b>     | 0.00 hour  |
| <b>Description</b>  |          |                        |            |
| <p>In filter_by_past_tense, there is a condition which checks whether the KP is multi-word. If it is, we don't check the past-tense condition and simply keep it in the final list of the KPs. We don't want to do that anymore.</p> <p>There are cases like "contained objects" matching with "container objects" in the C-API book that is passing through due to the above condition. Let us make changes such that if the KP is multi-word,</p> <ol style="list-style-type: none"><li>1) Check if there are any words with past tense in it. Let's call it A.</li><li>2) If there is, we need to find its stemmed equivalent in the header variant. Let's call it B.</li><li>3) Find the vector similarity score vec_sim between A &amp; B.</li><li>4) If vec_sim &lt; 0.374, add an additional penalty of 0.15 for the KP.</li></ol> <p>This will help us in reducing scores of KPs like "named attributes", "contained objects", "based class" which are actually quite horrible.</p> <p>Test this change with the C-API book and Lib Ref book.</p> |          |                        |            |

#### History

#1 - 11/03/2021 09:03 AM - Nandini Bansal

- Tracker changed from Support to Feature

#2 - 12/03/2021 10:17 AM - Nandini Bansal

- Status changed from New to Rejected