DATASET SITES:

* <https://ocw.mit.edu/courses/mathematics/18-700-linear-algebra-fall-2013/study-materials/>
* <https://ocw.mit.edu/courses/chemistry/5-61-physical-chemistry-fall-2013/lecture-notes/>
* <https://ocw.mit.edu/courses/physics/8-02-physics-ii-electricity-and-magnetism-spring-2007/readings/>
* <https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-006-introduction-to-algorithms-fall-2011/lecture-notes/>

Feature Weights for vicinity function candidates:

I manually created a dataset of 100 sentences with labels, extracted the 3 features:

1. Presence of words: below, above, following, preceeding, etc. (0/1)

2. Number of digits

3. Number of operators.

The limit for vicinity(upper bound) seems to be safe at 6.

70-30 for testing. Gives 80% accuracy with L2 normalization. 83% with scaling.

Weights without scaling: Obtained weights are: [[0.09378368 0.37030374 0.15054267]].

Weights with scaling: [[0.08150946 0.16566993 0.04817939]]

I get this with sklearn's scaling. I plugged in some random values like [0 8 4] etc, they work fine with this. Threshold for the LR function is 0.5 of course.

Excel sheets shared(GT of key concepts extracted and results of every extraction phase + Def, APP extraction GT):

* <https://docs.google.com/spreadsheets/d/1RJwwvapZoQz9gsBOE0yvk-k6Vi_MO_-6fu2Tg0E3mEY/edit?usp=sharing>
* <https://docs.google.com/spreadsheets/d/1qLmwSNg6E7mNg6Ma8hYYlUd8hNPQn9_UQxPheSqzPws/edit?usp=sharing>
* <https://docs.google.com/spreadsheets/d/1PuAazgBqLjt6kReqhgc1sS-a0GZY2m8hb3SDFakjmwQ/edit?usp=sharing>
* <https://docs.google.com/spreadsheets/d/1CpiqNRdnP5DQhotDist9bHKBVp93MI4XzLjb4zW97wU/edit?usp=sharing>

Final Results of KCE after incorporating StackExchange Pruning:

<https://docs.google.com/spreadsheets/d/1U8NUDVvGVhkuwSb_4b7vj_5gbFqbeRuLxe8iru3It-8/edit#gid=0>