3/27/2024

The datasets include the 1989 rare diseases (rows) that were found from 10,000 projects, which are the latest ones obtained through the RePORTER API.

**Criteria:**

The goal is to determine whether the focus of a project is on a identified rarer disease or not:

1. If the primary focus of the context is about the GARD name, the answer is 'True'.
2. Whether the keyword is matched or not (column F), the text can still be relevant to the identified rare disease.
3. If the text is about the study of a mechanism, we still considered it for further analysis.
4. It doesn’t matter whether the text is about humans or animals. If it is about the GARD name, we considered it as 'True'.

**Semantic similarity:**

We applied a fine-tuned pretrained transformer model to each project. The semantic similarity score is between 0 and 1. To calculate semantic similarity, we used the definition\* of the GARD disease instead of the GARD name itself.

1. **Column I**: Here, we calculate semantic similarity between the definition of the GARD name and sentences (part of the abstract\*\*) that includes either the GARD name or one of its synonyms (column N).
2. **Column J**: Here, we calculate semantic similarity between the definition of the GARD name and the whole abstract.

*\*In case there is no GARD name definition, we use the GARD name for calculating semantic similarity.*

*\*\* If the project does not have an abstract, we use the title instead for calculating the semantic similarity.*

**Confidence score (Column H):**

1. If we find the GARD name/its synonyms in the abstract, I give a **score of 20** to it.
2. If we find the GARD name/its synonyms in the abstract:

* If it is in the sentence with "the goal of" or 'aim', we consider the **weight of 7** for the number of repetitions of the word.
* If it is in the first sentence, we consider the **weight of 5** for the number of repetitions of the word.
* If it is in the sentence with future tense, we consider the **weight of 3** for the number of repetitions of the word.
* If it is in the sentence with past tense, we consider the **weight of 1** for the number of repetitions of the word.

For example, if there are two terms in the sentence with future tense, 1 term in the sentence with "the goal of", and 3 terms in the sentence with past tense, the confidence score is: 23 + 7 + 31 = 16.

1. If we find the GARD name/its synonyms in the public health relevance statement, we double the weights compared to those of abstracts:

* If it is in the sentence with "the goal of" or 'aim', we consider the **weight of 14** for the number of repetitions of the word, and so on.

**Expected evaluation:**

1. In the 'evaluation' column (column K), please consider 'True' if the focus of text is about the identified rare disease, and 'False' if it is not related to the identified rare disease.
2. Please put any comments you have in the 'comments' column (column L).
3. We expect that relevant results have higher confidence scores and higher semantic similarity, and vice versa. If you notice cases that go against this situation, you may report it as a "Bad CONF\_SCORE" or "Bad SEM\_SIM", or you may add a comment about that.