**12-04-2023**

1. Can you place dataset inside the gdrive so that I can also execute the code
2. Formula for evaluation of submission
3. Labels should not be concatenated and an then find the frequency
4. Go through some the notebooks in kaggle

**Formula for evaluation of submission**

* Predictions that more accurately match the precise words used to identify the dataset within the publication will score higher.
* Predictions should be cleaned using the clean\_text function from the [Evaluation page](https://www.kaggle.com/c/coleridgeinitiative-show-us-the-data/overview/evaluation) to ensure proper matching.
* Submissions are evaluated on a [Jaccard-based](https://en.wikipedia.org/wiki/Jaccard_index) [FBeta](https://en.wikipedia.org/wiki/F-score" \t "_blank) score between predicted texts and ground truth texts(Actual text), with Beta = 0.5 (F0.5 score)
* The Jaccard-based F Beta score is a performance metric that combines the Jaccard index and the F Beta score.
* The Jaccard index is a measure of similarity between two sets.
* you can calculate the Jaccard score between two strings using the following function:

def jaccard(str1, str2):

a = set(str1.lower().split())

b = set(str2.lower().split())

c = a.intersection(b)

return float(len(c)) / (len(a) + len(b) - len(c))

* F-beta score is a weighted harmonic mean of precision and recall
* F0.5 score = (1 + 0.5^2) \* (precision \* recall) / (0.5^2 \* precision + recall)
* The final winners of the competition will be determined based on their performance on a separate, private test set of articles.

**Additional information**

Clean\_text function is not an inbuild function

def clean\_text(txt):

return re.sub('[^A-Za-z0-9]+', ' ', str(txt).lower())