# Text Retrieval for biomimicry function identification in a corpus of biology papers

CS-410 Fall 2021

Progress Report

Tasks completed:

1. Evaluation measures have been identified. For this project we will compare text classification approaches based on MAP, gMAP, and average precision by topic. These measures were selected because our task is to present a user with a ranked list of documents based on their relevancy to a particular topic.
2. Class names have been identified. For this project we will only focus on the 10 level1 labels.
   1. physically\_assemble\_or\_disassemble
   2. protect\_from\_harm
   3. sense\_send\_or\_process\_information
   4. chemically\_modify\_or\_change\_energy\_state
   5. maintain\_structural\_integrity
   6. attach
   7. move
   8. process\_resources
   9. sustain\_ecological\_community
   10. change\_size\_or\_color
3. Dataset has been compiled and loaded into github
4. A Jupyter notebook that provides the template for our experiments has been created that performs the following:
   1. Loads the data from github.
   2. Preprocesses the data to remove papers with no titles or level1 labels
   3. Runs each paper through a pretrained multi-label text classification model (bart-large-mnli) that outputs scores for each label for each paper
   4. MAP, gMAP, average precision and PR curves for each topic are computed by comparing the model’s predictions against the ground truth.

Remaining Tasks:

1. Try different hypothesis templates for bart-large-mnli
2. Try different class names
3. Save evaluation measures to disk so they can be compared later.
4. Train a supervised multi-label text classification model such as SciBERT for comparison.
5. Try out other text classification approaches such as LOTClass for comparison

Challenges:

1. Colab “are you still there” timeouts