# Problem formulation

Our project is a hierarchical multi-class classification problem, with 3 major classes, and 6 minor classes corresponding to one of the major classes. For our evaluation metric, we use macro-F1 score, which is the average F1 score across all classes for a given dataset.

# Metric explanation

The F1 score can be interpreted as a harmonic mean of precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is: F1 = 2 \* (precision \* recall) / (precision + recall). Since our classification task is multiclass, we take the average of each class without weighting. This metric does not take label imbalance into account.

# Evaluation script

To use the evaluation script, import the `evaluation.py` file, and then run `evaluation.get\_f1\_score()`, passing in the gold sequence and the model’s prediction. This function returns the macro F1 score. Additionally, we provide a helper function `evaluation.train\_and\_evaluate()`, that trains a model from scratch, and optionally plots the confusion matrix. The `key` parameter should be either “class” or “implicit\_class”, for the major or minor classification problems, respectively.

# Reference:

[^1] Wikipedia F1 score: <https://en.wikipedia.org/wiki/F-score>

[^2] scikit learn function: <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html>