**Annotation Instructions Tutorial**

**General Instructions:**

A word cluster is represented in the form of a list of tokens in the first column followed by javalang generated automated labels and associated lines of code that the token belongs to. An LLM generated label is also provided for reference.

A screenshot of a computer

Description automatically generated

The annotation task consists of first looking at a group of code tokens (i.e., representing as a group or cluster) and answering the following questions.

1. Is the cluster or word group meaningful?
   1. **Yes:** if it represents a meaningful cluster.
   2. **No:** if it does not represent any meaningful cluster.
   3. **Don’t know or can’t judge:** if it does not have enough information to make a judgment. It is recommended to categorize the word groups using this label when the word group is not understandable at all.
2. **Lexical labels**: patterns related to naming of tokens in cluster. Ex: all tokens end with “obj” or all tokens have common substring
3. **Syntactic labels**: tokens that are a part of the abstract syntax tree (also provided using .label file.)
4. **Semantic labels:** patterns based on context in which tokens are provided. The context sentences have been provided within the annotation tool for reference.

**Multiple Concepts within a cluster:**

1. If the words can belong to multiple concepts, assign both concepts (for example, identifiers, numbers).
2. Label them in order of frequency. If order of frequency is unclear, put them down in a random order.
3. If there are more than three, end it with ‘etc.’ If there are too many, name it as miscellaneous if there is a theme,  for example miscellaneous identifiers, or leave the field blank.