# Dataset Training Collection

## API Mention Definition

We limit our scope of focus to Java Programming Language. In other words, our model will only classify API mention based on Java.

The definition of our API mention is fairly simple:

* Class.Method()
* Object.Method()
* Method()

To sum it up, we consider every method call to be API mention. The Class can either be native Java class or user-defined class.

In our dataset training collection from StackOverflow, we select posts that meet our API mention requirements above. Additionally, those API mentions must fulfil these formatting requirements as well:

* inline
* surrounded in <code></code> tag
* not in a <pre></pre> tag
* not in a <blockquote></blockquote> tag

We do not define any regulations for the number of parameters and non-existing parentheses as long as it is a valid method name in the given context.

## Collecting Training Data

We selected our training data from the data collected with StackExchange API. Based on our API mention definition, we displayed all the JSON response in a more readable form (using index.html), and manually selected and extracted the post that contains the API mention (with extract.py).

# POS Tagging

For this task, we randomly selected 10 questions from the whole dataset and saved it as a .txt file. By using Natural Language Toolkit (nltk) Python Library, we apply word\_tokenize function by nltk to the selected sentences (.txt file) line by line and saved the result in the form of .txt file as well.