Congratulations from this page you can download the

Loyola University of Delaware Identifier Splitting Oracle!

The oracle, found in the file loyola-udelaware-identifier-splitting-oracle (ludiso) consists of 2663 identifiers. These identifiers were split by volunteers using a web interface. In all 8522 splitting judgements were collected for 2731 identifiers. The data set was curated based on the sum of confidence associated with a particular split. In order to provide a single correct split in the oracle, the split receiving the highest confidence sum was used as the correct split. Of the 2731 identifers, 68 do not appear in the oracle because of ties. (These identifiers can be found in the raw data.) Details on the construction of the oracle can be found in "An empirical study of identifier splitting technique".

Each line of the oracle represents one of the 2663 remaining oracle identifiers and includes the following.

- the identifier's unique identification number,
- the original identifier as extracted from the source,
- the dominant language of the program that the id was extracted from,
- the program from which it was extracted,
- the identifier as shown to subjects where a '-' denotes a hard split,
- the number of splittings (always 1 in the oracle, but not in the raw data),
- the identifier as split by users,
- the number of confidences (1 - 5) and then each confidence.

The confidence was the user's assessment of their correctness using a scale from "a guess" (0) to "i'm sure" (2).

In addition to the oracle, the raw data can be downloaded here . The raw data includes all user splittings not just the one with the highest confidence sum. Finally, source code to read the oracle into a Java or C program can be found in the directory read-oracle . Filters in this code can be used to select subset of the raw data collected.

This project is supported by NSF grant CCF 0916081.