Document 1: Input Data Format for PLATIN-UMS

Timotej Bernat, 2020

While PLATIN-UMS does most of the heavy lifting as far as chemically labelling and organizing data, there are still some simple and clearly-defined formatting rules for data inputs that must be obeyed in order for the program to function without errors; these are outlined here.

1. Each data file should be in the CSV (Comma Separated Value) file format. A spreadsheet can easily be made into a CSV within Excel by navigating to File --> Save As --> [Folder of Choice] --> Save As Type, selecting either “CSV (Comma-Delimited)” or “CSV (MS-DOS)”, and saving. These files types do not behave differently to ordinary spreadsheets when viewed in Excel, and thus may be comfortably used for data entry even prior to data processing/formatting.
2. Each row in the file should begin in Column A and should consist of two (and only two) parts; the name of the current instance and the accompanying spectral data. Each part has its own rules
   1. **NAMES:** the name of an instance should consist of:
      1. The name of the species the instance belongs to, written explicitly and following IUPAC conventions. It is strongly recommended that you capitalize species names (particularly if it consists of multiple words), though uncapitalized names will still be accepted
      2. Either a single hyphen or a single space
      3. The number of that instance, followed by any amount of whitespace.

|  |  |  |
| --- | --- | --- |
| Example Name | Acceptable? | Reason/Comments |
| Heptyl Acetate-35 | **✔** | This adheres to the rules outlined |
| Heptyl Acetate 35 | **✔** | You can opt for a space instead of a hyphen, as long as there’s only one |
| 1-Propanol 24 or  1-Propanol-24 | ✔ | IUPAC numbering is perfectly acceptable |
| Diethyl Ether 7 or  Diethyl Ether-7 | **✔** | Multi-word names are fine, as long as the other conditions are met |
| diethyl ether 7 | ⚠️ | Won’t break the program due to a special case for this, but you should feel great shame if you do this |
| Heptyl Acetate | ⚠️ | Won’t break the program, but outputs will overwrite one another if multiple instances are unnumbered |
| Acetylene 34 or Acetylene-34 | ⚠️ | Won’t break the program, but will be mislabeled as an alkene by IUPAC naming rules, use Ethyne-34 instead |
| MIBK 45 | ✖ | Species name says nothing about the chemical identity,  “Methyl-iBu-Ketone 45” would work instead |
| Chemical-36 | ✖ | Species name says nothing about the chemical identity, use a non-generic, IUPAC-recognized name instead |
| Heptyl Acetate----35 | ✖ | Too many hyphens, will fail to separate species |

* 1. **DATA:** the data should consist of a series of cells (as many as are required to accommo- date the spectrum data) starting in column B of the same row as the corresponding name. Each cell should contain one real number (can be whole or decimal, i.e. 0.3464, 1, and 1.0 are all valid entries). There are only a couple of constraints:

1. **IDENTIFICATION VECTORS ARE NOT NEEDED** and will in fact cause errors if included. A spectrum row with the final cell being [1,0,0,0,0] will throw an error when trying to read the data set that contains it. A spectrum row ending with a vector broken up among cells (i.e. 1 in one cell, 0 in the next, 0 in the next, etc.) will be interpreted as additional data points, which, while not error-inducing, may be detrimental to identification performance and data control.
2. **ALL ENTRIES ARE OF THE SAME LENGTH.** Regardless of application or configuration, all neural networks have an input layer of constant size into which tensors are fed. If each row of spectrum data contains a different number of data points, the program will throw an error upon attempting to train for this reason. Even if all your spectra are the same length, ensure there is no whitespace between the name and the first data point which might give the appearance of differing entry lengths.

If these few simple formatting rules are obeyed, PLATIN-UMS will accept your data and be able to label it and train over it immediately after providing your data, with no additional processing required. Additionally, existing datasets can be appended to at will, as long as all new row entries follow the rules outlined above.