Kestrel Taxonomy Finder

Version 0.1

Shawn.M.Rupp@asu.edu

Contents

Introduction.....2
Installation.....2
Getting Started...3
Scripts.....4

Introduction

Kestrel is a program for resolving species' common names and synonyms with "official" scientific names and extracting taxonomies from internet databases.

Dependencies:
Python3
Cython
NLTK
BeautifulSoup4
Active internet connection

Installation

Cython
Kestrel utilize Cython to compile python code into C and drastically improve performance. Cython can be installed from the pypi repository or via Miniconda (it is installed by default with the full Anaconda package).

> To install with Miniconda:
> `conda install cython`

NLTK
Kestrel uses python's Natural Language Processing Toolkit to differentiate between common and scientific names in its input. To install on any Debian-based Linux platform, enter the following into a terminal:

```
sudo pip install -U nltk
```

Kestrel comes with it's own training dataset, so you do not need to download any additional data from NLTK.

BeautifulSoup4
Kestrel the BeautifulSoup module, and the lxml parser, to parse hmtl and xml pages.

```
apt-get install python3-bs4
apt-get install python-lxml
```

Kestrel
Download the git repository, change into the directory, and build the Cython scripts.

```
git clone https://github.com/icwells/Kestrel.git
cd Kestrel/
./install.sh
```

Getting Started

EOL API Key
Kestrel queries the Encyclopedia of Life heavily, so you will need to generate an api key. To do so, create and EOL profile if you do not already have one. Sign in, go to you profile, and click “edit my profile.” Next, generate an api key in the box on the right side. Once you generate the api key, copy it and paste it into example-API.txt on the same line as “EOL=”. Finally, change the file name to API.txt (otherwise it will be erased if you run “git pull”).

Running Kestrel
Assuming you have a functional internet connection, you are ready to run the program. To run it, change into Kestrel/bin/ and enter

```
python kestrel.py {options} –c <column number> –i <input file> –o
<output file>
```

into a terminal (detailed command description in the Scripts section). The -c flag signifies the column number of the input file where the names the names to be searched are located. This is a 0-based integer, so if the names are in the first column specify “-c 0”. The program will extract unique names from this column. It will only search for a specific name once to avoid burdening any servers. Kestrel starts by reading names from any existing output files before reading names from the input file. If the given output file already exists, it will append output to this file which allows you to resume if the program is interrupted.

If neither the –common nor the –scientific flag is given, it will then use the species names in commonNames.csv.gz to train a feature classifier with NLTK. It will then use this classifier to determine whether a input name is a common or scientific name. If either of the above flags are given, it will skip this step.

Kestrel will query EOL, GBIF, and Wikipedia in succession, but will quit once it finds a taxonomy that includes a genus entry and has no more than 1 missing value (the vast majority of results will have all 7 seven fields). Scientific names are first queried against GBIF, then EOL, and lastly Wikipedia. If no matches are found, the name is written to the kestrelMisses.txt file. Common names are queried against EOL and Wikipedia. If no matches are found and the name has more than one word, the program will remove the first word before repeating the search loop. It will continue until a match is found or there is only one word left. If there is no match at this point, the name is recorded in the misses file.

Output
For names with an identified taxonomy, output will be written to the given output file in csv format. The query name will be in the first column, followed by a columns for each taxonomy field. The last column contains the url the taxonomy was extracted from. This identifies the source of the data and allows it to be accessed again. Any names for which a taxonomy is not identified will be written to the kestrelMisses.txt file, which will be located in the same directory as the ouput file.

Scripts

kestrel.py
This is the only executable script in the package and will take input from a given column of a given file and search GBIF, EOL, and Wikipedia for taxonomy information.

Usage:
python kestrel.py {options} -c <column number> -i <input file> -o <output file>

-h, --help	show this help message and exit
-v	Prints version info and exits.
--common	Indicates that input contains only common names.
--scientific	Indicates that input contains only scientific names.
-i I	Path to input file.
-o O	Path to output file.
-c C	Column containing species names (integer starting from 0).
-t T	Number of threads for identifying taxa (default = 1).