Module 10: Python Regular Expressions and Dictionaries

Module Overview

In this module you'll use Python regular expressions to parse `/scratch/go-basic.obo` and put the fields in a dictionary with GO id as the key. The GO records in this file are multi-line, so you'll need to use a record separator other than the newline character. Unlike Perl, Python doesn't allow you to change the record separator for the `readline()` method, so you'll read the whole file in with the `read()` method and then use a regular expression to split the file into records. Since you want to find all the records with your regular expression, you'll need to use:

```
re.findall(r"your regex here",goFile, re.DOTALL)
```

which returns a list of matches. `re.DOTALL` tells Python to match across line breaks with `.*`.

Required Reading

- *Python for Biologists* Chapter 7
- *Python for Biologists* Chapter *8*

SwissProt Parser

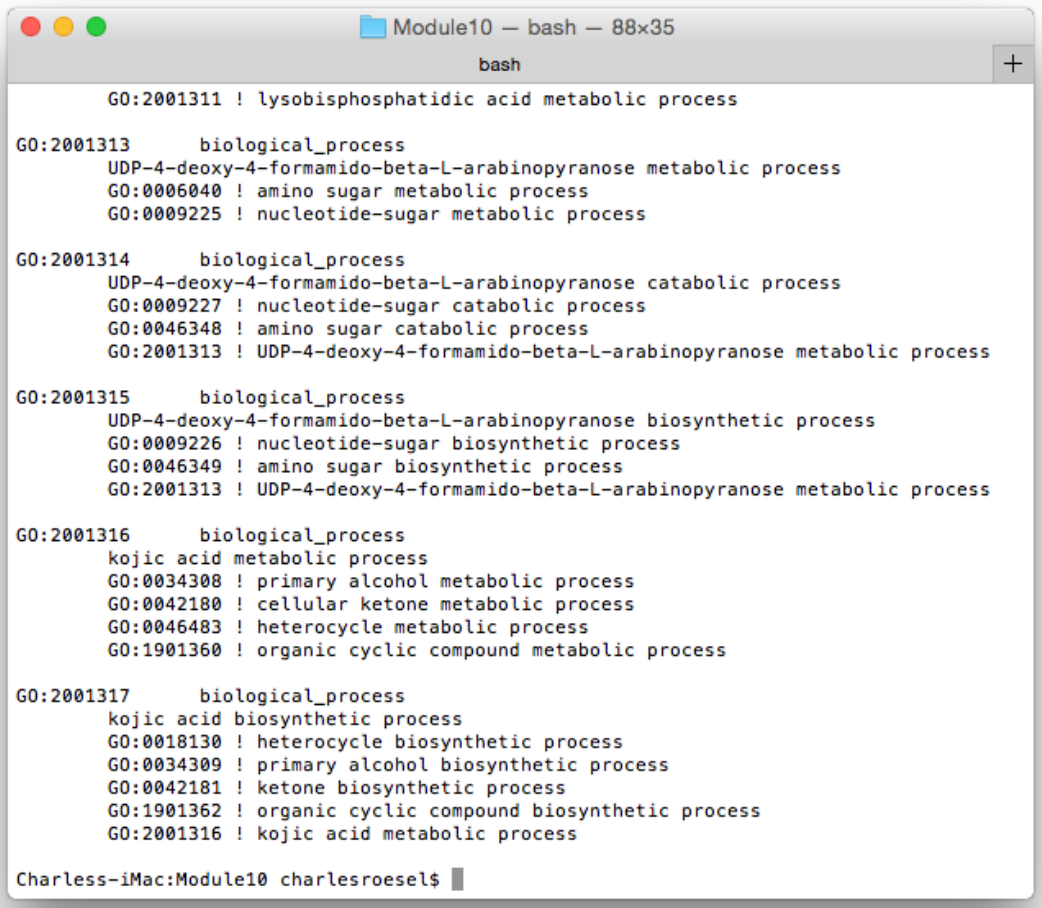
The code shown below parses a SwissProt file. SwissProt records are multi-line, so it's very similar to what you need to do to parse go-basic.obo.

```python
import re

def parse_record(record):
    accession = re.search(r"AC\s+(.*?)\n",record, re.DOTALL)
    dates = re.findall(r"DT\s+(.*?)\n",record, re.DOTALL)
    if(accession):
        print (accession.group(1))
        for date in dates:
            print(date)

def split_records(file):
    sp_file = open(file)
    sp_records = sp_file.read()
    sp_split_records = re.findall(r"(ID.*?)\/\/",sp_records, re.DOTALL)
    for sp_record in sp_split_records:
        parse_record(record=sp_record)
    sp_file.close()

split_records(file="/scratch/SampleDataFiles/example.sp")
```

Assignment

Complete as many exercises from the book as necessary to understand the concepts. These will not be graded. The graded part of the assignment is to use regular expressions to parse `/scratch/go-basic.obo` and put the results in a dictionary. Your program should be written for Python3 and named `~/BINF6200/Module10/parseGoInfo.py`.

- Parse the GO id, name, namespace, and is_a values for each term.
- Create a string with namespace on the first line followed by a line for name, and one line per is_a.
- Put the string as the value in a dictionary where go_id is the key.
- Iterate over the keys in the dictionary, printing go_id followed by a tab, then the string containing the name, namespace, and is_a values.
- Create a function for splitting the file into records, and a function for splitting the records into fields.
- Your output should look something like this:



```
Module10 — bash — 88×35
bash

        GO:2001311 ! lysobisphosphatidic acid metabolic process

GO:2001313      biological_process
        UDP-4-deoxy-4-formamido-beta-L-arabinopyranose metabolic process
        GO:0006040 ! amino sugar metabolic process
        GO:0009225 ! nucleotide-sugar metabolic process

GO:2001314      biological_process
        UDP-4-deoxy-4-formamido-beta-L-arabinopyranose catabolic process
        GO:0009227 ! nucleotide-sugar catabolic process
        GO:0046348 ! amino sugar catabolic process
        GO:2001313 ! UDP-4-deoxy-4-formamido-beta-L-arabinopyranose metabolic process

GO:2001315      biological_process
        UDP-4-deoxy-4-formamido-beta-L-arabinopyranose biosynthetic process
        GO:0009226 ! nucleotide-sugar biosynthetic process
        GO:0046349 ! amino sugar biosynthetic process
        GO:2001313 ! UDP-4-deoxy-4-formamido-beta-L-arabinopyranose metabolic process

GO:2001316      biological_process
        kojic acid metabolic process
        GO:0034308 ! primary alcohol metabolic process
        GO:0042180 ! cellular ketone metabolic process
        GO:0046483 ! heterocycle metabolic process
        GO:1901360 ! organic cyclic compound metabolic process

GO:2001317      biological_process
        kojic acid biosynthetic process
        GO:0018130 ! heterocycle biosynthetic process
        GO:0034309 ! primary alcohol biosynthetic process
        GO:0042181 ! ketone biosynthetic process
        GO:1901362 ! organic cyclic compound biosynthetic process
        GO:2001316 ! kojic acid metabolic process

Charless-iMac:Module10 charlesroesel$
```