

BINF 6111/ITCS 8111: Lists and dictionaries

Question ONE: Which of the following expressions or methods return values, do work through side effects, or do both?

	Description	Example invocation	Side Effects?	Return Value?
<i>Example</i>	Taking a slice from a list.	<pre>>>> lst = [1,2,3] >>> lst[0:2]</pre>	No	Yes
(a)	The <code>append</code> method belonging to a list.	<pre>>>> lst = ['a',2,3] >>> lst.append('foo')</pre>		
(b)	The <code>sort</code> method belonging to a list.	<pre>>>> lst.sort()</pre>		
(c)	<code>print</code>	<pre>>>> print "foo"</pre>		
(d)	The <code>rstrip</code> method belonging to a string	<pre>>>> 'foo'.rstrip()</pre>		

Question TWO: List comprehensions.

1. Write a list comprehension expression that computes the square root and then adds 5 to a sequence of integers from 1 to 100, inclusive. (*Hint:* The `math` module has a function `sqrt`. You can use `import` to load the `math` module into your evaluation environment.)
2. Write an expression that subtracts one from all even numbers from 2 to 100, inclusive. (*Hint:* `range` can have three parameters).
3. Write an expression using `==` which tests that the expression you wrote for part 2 returns the correct answer.
4. Create the following list of lists: `[[1, 2], [2, 3], [3, 4], [4, 5]]`

Question THREE: Research how the built-in method `cmp` compares strings. Consider: Does it use the length of the strings? Does it consider the order each strings' first letter or letters appear in the alphabet? Explain how `cmp` compares strings in plain language, using examples if necessary. Then test your answer using some example lists.

Question FOUR: Download the python program file `read_gene2pubmed.py` from the course Web site. Then, download `gene2pubmed.gz` from the class Web site.

Read the documentation (README file) describing `gene2pubmed.gz`, which comes from <ftp://ftp.ncbi.nih.gov/gene>.

Use your knowledge of Unix to extract all lines referring to human genes and save them to a file. Then, use the functions in `read_gene2pubmed.py` to build a dictionary in which each key is an Entrez Gene id for *human genes only* and values are lists of PubMed article identifiers that were associated with their respective keys in the file `gene2pubmed.gz`.

Tip: To save typing, import the module `read_gene2pubmed` into your python environment using the following command:

```
>>> import read_gene2pubmed as rg
```

And then, invoke its methods like so:

```
>>> rg.do_something()
```

Recall that `import` will execute without an error if the module file (`read_gene2pubmed.py`) is in the same directory where you invoked the python interpreter *or* if it resides in one of the directories specified by your `PYTHONPATH` environment variable.

Now, answer the following questions. **Hint:** These will be easy to answer using combinations of `filter`, `max`, `len`, and the built-in dictionary method `keys`.

1. How many human genes were represented in the file?
2. How many human genes have more than 100 article references, if any?
3. Which human gene has the largest number of article references? (If several tie for top place, list them all.) How many articles were there?

Question FIVE: Answer the same questions as above but for (a) *Arabidopsis thaliana* Col-0 (b) *Drosophila melanogaster* and (c) the house mouse. You can obtain their taxonomy ids from the Taxonomy database at NCBI.

