Python 5

Dictionaries

Dictionaries are another iterable, like a string and list. They act like lists, with one big difference, instead of retrieving values with a numerical index they use strings. You can look items up in a python dictionary just like you do when you look up items in the English dictionary, the key, or a string.

For example, when you want to know what the definition of the word, 'onomatopoeia', you look it up using the word, 'onomatopoeia'. If you want to know the value, or in our exampple below, the sequence of TP53 you would use 'TP53' as a look up.

Dictionaries are a collection of key/value pairs. In Python, each key is separated from its value by a colon (:), the items are separated by commas, and the whole thing is enclosed in curly braces. An empty dictionary without any items is written with just two curly braces, like this: `{}`

Each key in a dictionary is unique, while values may not be. The values of a dictionary can be of any type, but the keys must be of an immutable data type such as strings, numbers, or tuples.

Data that is appropriate for dictionaries are two pieces of information that naturally go together, like gene name and sequence.

Key	Value
TP53	GATGGGATTGGGGTTTTCCCCTCCCATGTGCTCAAGACTGGCGCTAAAAGTTTTGAGCTTCTCAAAAGTC
BRCA1	GTACCTTGATTTCGTATTCTGAGAGGCTGCTGCTTAGCGGTAGCCCCTTGGTTTCCGTGGCAACGGAAAA

Creating a Dictionary

```
genes = { 'TP53' :
'GATGGGATTGGGGTTTTCCCCTCCCATGTGCTCAAGACTGGCGCTAAAAGTTTTGAGCTTCTCAAAAGTC' , 'BRCA1' :
'GTACCTTGATTTCGTATTCTGAGAGGCTGCTGCTTAGCGGTAGCCCCTTGGTTTCCGTGGCAACGGAAAA' }
```

Breaking up the key/value pairs over multiple lines make them easier to read.

```
genes = {
           'TP53' :
'GATGGGATTGGGGTTTTCCCCTCCCATGTGCTCAAGACTGGCGCTAAAAGTTTTGAGCTTCTCAAAAGTC' ,
           'BRCA1' :
'GTACCTTGATTTCGTATTCTGAGAGGCTGCTGCTTAGCGGTAGCCCCTTGGTTTCCGTGGCAACGGAAAA'
         }
```

Accessing Values in Dictionaries

To retrieve a single value in a dictionary use the value's key in this format `dict[key]`. This will return the value at the specified key.

```
>>> genes = { 'TP53' :
'GATGGGATTGGGGTTTTCCCCTCCCATGTGCTCAAGACTGGCGCTAAAAGTTTTGAGCTTCTCAAAAGTC' , 'BRCA1' :
'GTACCTTGATTTCGTATTCTGAGAGGCTGCTGCTTAGCGGTAGCCCCTTGGTTTCCGTGGCAACGGAAAA' }
>>>
>>> genes['TP53']
GATGGGATTGGGGTTTTCCCCTCCCATGTGCTCAAGACTGGCGCTAAAAGTTTTGAGCTTCTCAAAAGTC
```

> The sequence of the gene TP53 is stored as a value of the key 'TP53'. We can access the sequence by using the key in this format dict[key]

The value can be accessed and passed directly to a function or stored in a variable.

```
>>> print(genes['TP53'])
GATGGGATTGGGGTTTTCCCCTCCCATGTGCTCAAGACTGGCGCTAAAAGTTTTGAGCTTCTCAAAAGTC
>>>
>>> seq = genes['TP53']
>>> print(seq)
GATGGGATTGGGGTTTTCCCCTCCCATGTGCTCAAGACTGGCGCTAAAAGTTTTGAGCTTCTCAAAAGTC
```

Changing Values in a Dictionary

Individual values can be changed by using the key and the assignment operator.

```
>>> genes = { 'TP53' :
'GATGGGATTGGGGTTTTCCCCTCCCATGTGCTCAAGACTGGCGCTAAAAGTTTTGAGCTTCTCAAAAGTC' , 'BRCA1' :
'GTACCTTGATTTCGTATTCTGAGAGGCTGCTGCTTAGCGGTAGCCCCTTGGTTTCCGTGGCAACGGAAAA' }
>>>
>>> print(genes)
{'BRCA1': 'GTACCTTGATTTCGTATTCTGAGAGGCTGCTGCTTAGCGGTAGCCCCTTGGTTTCCGTGGCAACGGAAAA',
'TP53': 'GATGGGATTGGGGTTTTCCCCTCCCATGTGCTCAAGACTGGCGCTAAAAGTTTTGAGCTTCTCAAAAGTC'}
>>>
>>> genes['TP53'] = 'atg'
>>>
>>> print(genes)
{'BRCA1': 'GTACCTTGATTTCGTATTCTGAGAGGCTGCTGCTTAGCGGTAGCCCCTTGGTTTCCGTGGCAACGGAAAA',
'TP53': 'atg'}
```

> The contents of the dictionary have changed.

Other assignment operators can also be used to change a value of a dictionary key.

```
>>> genes = { 'TP53' :
'GATGGGATTGGGGTTTTCCCCTCCCATGTGCTCAAGACTGGCGCTAAAAGTTTTGAGCTTCTCAAAAGTC' , 'BRCA1' :
'GTACCTTGATTTCGTATTCTGAGAGGCTGCTGCTTAGCGGTAGCCCCTTGGTTTCCGTGGCAACGGAAAA' }
>>>
>>> genes['TP53'] +=
'TAGAGCCACCGTCCAGGGAGCAGGTAGCTGCTGGGCTCCGGGGACACTTTGCGTTCGGGCTGGGAGCGTG'
>>>
>>> print(genes)
{'BRCA1': 'GTACCTTGATTTCGTATTCTGAGAGGCTGCTGCTTAGCGGTAGCCCCTTGGTTTCCGTGGCAACGGAAAA',
'TP53':
'GATGGGATTGGGGTTTTCCCCTCCCATGTGCTCAAGACTGGCGCTAAAAGTTTTGAGCTTCTCAAAAGTCTAGAGCCACCGTCCAG
GGAGCAGGTAGCTGCTGGGCTCCGGGGACACTTTGCGTTCGGGCTGGGAGCGTG'}
```

> Here we have used the '+=' concatenation assignment operator. This is equivalent to `genes['TP53'] = genes['TP53'] + 'TAGAGCCACCGTCCAGGGAGCAGGTAGCTGCTGGGCTCCGGGGACACTTTGCGTTCGGGCTGGGAGCGTG'`.

Accessing Each Dictionary Key/Value

Since a dictionary is an iterable object, we can iterate through its contents.

A for loop can be used to retrieve each key of a dictionary one a time:

```
>>> for gene in genes:
...     print(gene)
...
TP53
BRCA1
```

Once you have the key you can retrieve the value:

```
>>> for gene in genes:
...     seq = genes[gene]
...     print(gene, seq[0:10])
...
TP53 GATGGGATTG
BRCA1 GTACCTTGAT
```

Building a Dictionary one Key/Value at a Time

Building a dictionary one key/value at a time is akin to what we just saw when we change a key's value. Normally you won't do this. We'll talk about ways to build a dictionary from a file in a later lecture.

```
>>> genes = {}
>>> print(genes)
{}
>>> genes['Brca1'] =
'GTACCTTGATTTCGTATTCTGAGAGGCTGCTGCTTAGCGGTAGCCCCTTGGTTTCCGTGGCAACGGAAAA'
>>> genes['TP53'] =
'GATGGGATTGGGGTTTTCCCCTCCCATGTGCTCAAGACTGGCGCTAAAAGTTTTGAGCTTCTCAAAAGTC'
>>> print(genes)
{'Brca1': 'GTACCTTGATTTCGTATTCTGAGAGGCTGCTGCTTAGCGGTAGCCCCTTGGTTTCCGTGGCAACGGAAAA',
'TP53': 'GATGGGATTGGGGTTTTCCCCTCCCATGTGCTCAAGACTGGCGCTAAAAGTTTTGAGCTTCTCAAAAGTC'}
```

> We start by creating an empty dictionary. Then we add each key/value pair using the same syntax as when we change a value.
> dict[key] = new_value

Checking That Dictionary Keys Exist

Python generates an error (NameError) if you try to access a key that does not exist.

```
>>> print(genes['HDAC'])
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'HDAC' is not defined
```

Dictionary Operators

Operator	Description
`in`	`key in dict` returns True if the key exists in the dictionary
`not in`	`key not in dict` returns True if the key does not exist in the dictionary

Because Python generates a NameError if you try to use a key that doesn't exist in the dictionary, you need to check whether a key exists before trying to use it.

The best way to check whether a key exists is to use `in`

```
>>> gene = 'TP53'
>>> if gene in genes:
...     print('found')
...
found
>>>
>>> if gene in genes:
...     print(genes[gene])
...
GATGGGATTGGGGTTTTCCCCTCCCATGTGCTCAAGACTGGCGCTAAAAGTTTTGAGCTTCTCAAAAGTC
>>>
```

Building a Dictionary one Key/Value at a Time using a loop

Now we have all the tools to build a dictionary one key/value using a for loop. This is how you will be building dictionaries more often in real life.

Here we are going to count and store nucleotide counts:

```
#!/usr/bin/env python3

# create a new empty dictionary
nt_count={}

# loop example from loops lecture
dna = 'GTACCTTGATTTCGTATTCTGAGAGGCTGCTGCTTAGCGGTAGCCCCTTGGTTTCCGTGGCAACGGAAAA'
for nt in dna:

  # is this nt in our dictionary?
  if nt in nt_count:
    # if it is, lets increment our count
    previous_count = nt_count[nt]
    new_count = previous_count + 1
    nt_count[nt] = new_count
  else:
    # if not, lets add this nt to our dictionary and make count = 1
    nt_count[nt] = 1;

print(nt_count)
```

> {'G': 20, 'T': 21, 'A': 13, 'C': 16}

What is another way we could increment our count?

```
nt_count={}

dna = 'GTACCTTGATTTCGTATTCTGAGAGGCTGCTGCTTAGCGGTAGCCCCTTGGTTTCCGTGGCAACGGAAAA'
for nt in dna:
  if nt in nt_count:
    nt_count[nt] += 1
  else:
    nt_count[nt] = 1;

print(nt_count)
```

> remember that `count=count+1` is the same as `count+=1`

Sorting Dictionary Keys

If you want to print the contents of a dictionary, you should sort the keys then iterate over the keys with a for loop. Why do you want to sort the keys?

```
for gene_key in sorted(genes): # python allows you to use this shortcut in a for loop
                               # you don't have to write genes.keys() in a for loop
                               # to iterate over the keys
  print(gene_key, '=>' , genes[gene_key])
```

> This will print keys in the same order every time you run your script. Dictionaries are unordered, so without sorting, you'll get a different order every time you run the script, which could be confusing.

Dictionary Functions

Function	Description
`len(dict)`	returns the total number of key/value pairs
`str(dict)`	returns a string representation of the dictionary
`type(variable)`	Returns the type or class of the variable passed to the function. If the variable is dictionary, then it would return a dictionary type.

These functions work on several other data types too!

Dictionary Methods

Method	Description
`dict.clear()`	Removes all elements of dictionary dict
`dict.copy()`	Returns a shallow copy of dictionary dict. Shallow vs. deep copying only matters in multidimensional data structures.
`dict.fromkeys(seq,value)`	Create a new dictionary with keys from seq (Python sequence type) and values set to value.
`dict.items()`	Returns a list of (key, value) tuple pairs
`dict.pop(key)`	Removes the key:value pair and returns the value
`dict.keys()`	Returns list of keys
`dict.get(key, default = None)`	get value from dict[key], use default if not present
`dict.setdefault(key, default = None)`	Similar to get(), but will set dict[key] = default if key is not already in dict
`dict.update(dict2)`	Adds dictionary dict2's key-values pairs to dict
`dict.values()`	Returns list of dictionary dict's values

Link to Python 5 Problem Set

Sets

A set is another Python data type. It is essentially a dictionary with keys but no values.

- A set is unordered
- A set is a collection of data with no duplicate elements.
- Common uses include looking for differences and eliminating duplicates in data sets.

Curly braces `{}` or the `set()` function can be used to create sets.

> Note: to create an empty set you have to use `set()`, not `{}` the latter creates an empty dictionary.

```
>>> basket = {'apple', 'orange', 'apple', 'pear', 'orange', 'banana'}
>>> print(basket)
{'orange', 'banana', 'pear', 'apple'}
```

> Look, duplicates have been removed

Test to see if an value is in the set

```
>>> 'orange' in basket
True
>>> 'crabgrass' in basket
False
```

> The in operator works the same with sets as it does with lists and dictionaries

Union, intersection, difference and symmetric difference can be done with sets

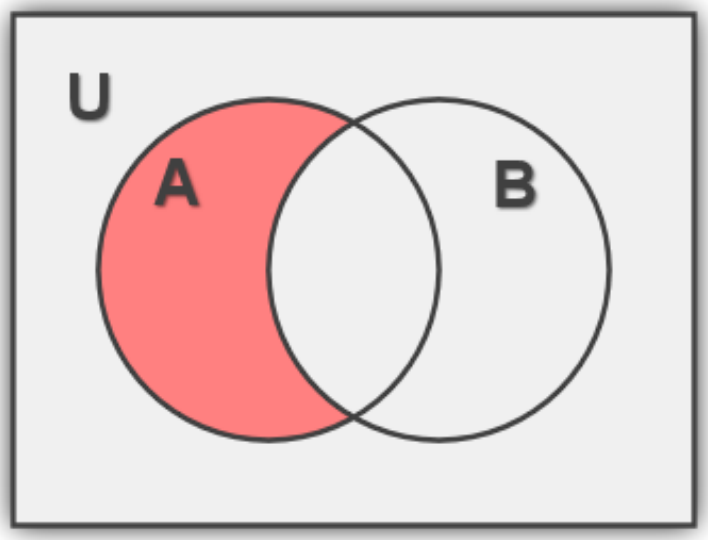
```
>>> a = set('abracadabra')
>>> b = set('alacazam')
>>> a
{'a', 'r', 'b', 'c', 'd'}
```

> Sets contain unique elements, therefore, even if duplicate elements are provided they will be removed.

Set Operators

Difference

The difference between two sets are the elements that are unique to the set to the left of the `-` operator, with duplicates removed.

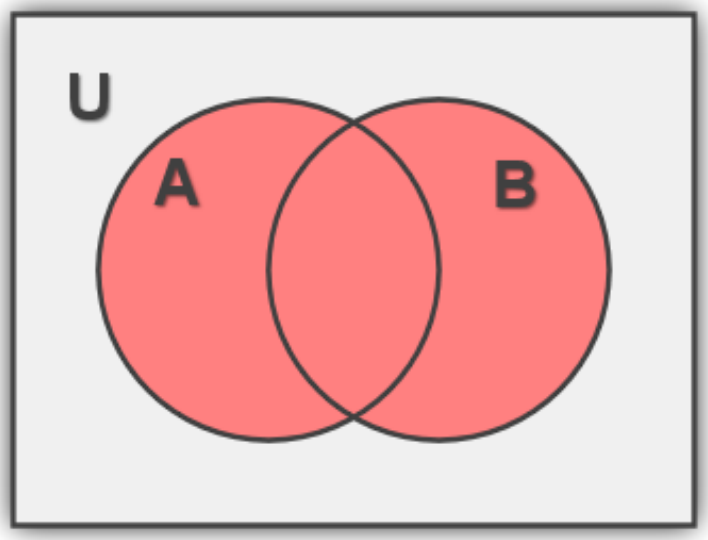



```
>>> a = set('abracadabra')
>>> b = set('alacazam')
>>> a - b
{'r', 'd', 'b'}
```

> This results the letters that are in a but not in b

Union

The union between two sets is a sequence of the all the elements of the first and second sets combined, with duplicates removed.

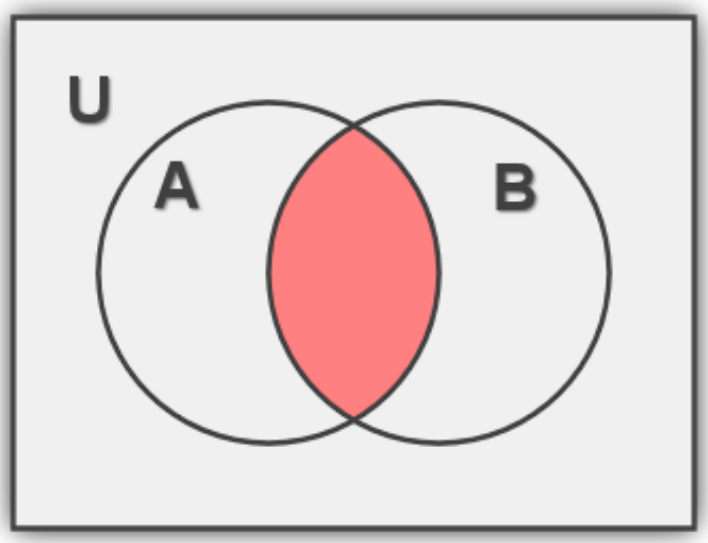


```
>>> a = set('abracadabra')
>>> b = set('alacazam')
>>> a | b
{'a', 'c', 'r', 'd', 'b', 'm', 'z', 'l'}
```

> This returns letters that are in a or b both

Intersection

The intersection between two sets is a sequence of the elements which are in both sets, with duplicates removed.

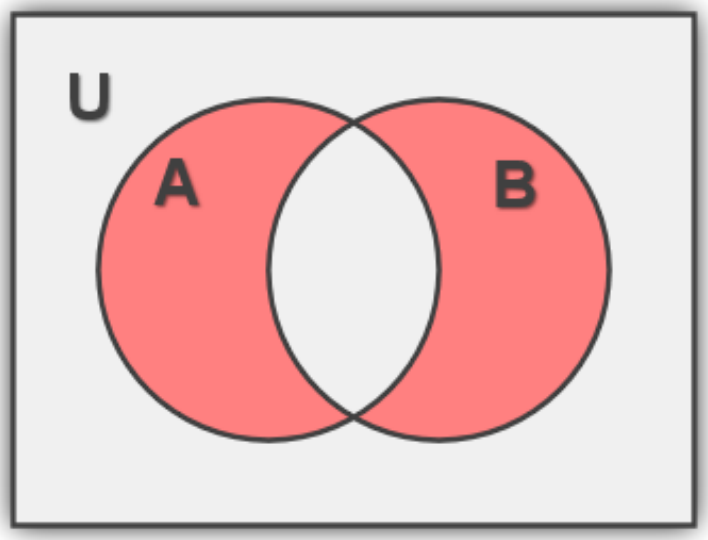



```
>>> a = set('abracadabra')
>>> b = set('alacazam')
>>> a & b
{'a', 'c'}
```

> This returns letters that are in both a and b

Symmetric Difference

The symmetric difference is the elements that are only in the first set plus the elements that are only in the second set, with duplicates removed.




```
>>> a = set('abracadabra')
>>> b = set('alacazam')
>>> a ^ b
{'r', 'd', 'b', 'm', 'z', 'l'}
```

> This returns the letters that are in a or b but not in both (also known as exclusive or)

Set Functions

Function	Description
`all()`	returns True if all elements of the set are true (or if the set is empty).
`any()`	returns True if any element of the set is true. If the set is empty, return False.
`enumerate()`	returns an enumerate object. It contains the index and value of all the items of set as a pair.
`len()`	returns the number of items in the set.
`max()`	returns the largest item in the set.
`min()`	returns the smallest item in the set.
`sorted()`	returns a new sorted list from elements in the set (does not alter the original set).
`sum()`	returns the sum of all elements in the set.

Set Methods

Method	Description
`set.add(new)`	adds a new element
`set.clear()`	remove all elements
`set.copy()`	returns a shallow copy of a set
`set.difference(set2)`	returns the difference of set and set2
`set.difference_update(set2)`	removes all elements of another set from this set
`set.discard(element)`	removes an element from set if it is found in set. (Do nothing if the element is not in set)
`set.intersection(sets)`	return the intersection of set and the other provided sets
`set.intersection_update(sets)`	updates set with the intersection of set and the other provided sets
`set.isdisjoint(set2)`	returns True if set and set2 have no intersection
`set.issubset(set2)`	returns True if set2 contains set
`set.issuperset(set2)`	returns True if set contains set2
`set.pop()`	removes and returns an arbitrary element of set.
`set.remove(element)`	removes element from a set.
`set.symmetric_difference(set2)`	returns the symmetric difference of set and set2
`set.symmetric_difference_update(set2)`	updates set with the symmetric difference of set and set2
`set.union(sets)`	returns the union of set and the other provided sets
`set.update(set2)`	update set with the union of set and set2

Build a dictionary of NT counts using a set and loops

Let us put a twist on our nt count script. Let's use a set to find all the unique nts, then use the string `count()` method to count the nucleotide instead if incrementing the count as we did earlier.

Code:

```
#!/usr/bin/env python3

# create a new empty dictionary
nt_count = {}

# get a set of unique characters in our DNA string

dna = 'GTACCNTTGATTTCGTATTCTGAGAGGCTGCTGCTTAGCGGTAGCCCCTTGGTTTCCGTGGCAACGGAAAA'
unique = set(dna)

print('unique nt: ', unique) ## {'C', 'A', 'G', 'T', 'N'}

# iterate through each unique nucleotide
for nt in unique:
  # count the number of this unique nt in dna
  count = dna.count(nt)

  # add our count to our dict
  nt_count[nt] = count


print('nt count:', nt_count)
```

Output:

```
unique nt:  {'N', 'C', 'T', 'G', 'A'}
nt count: {'G': 20, 'T': 21, 'A': 13, 'C': 16, 'N': 1}
```

> We have the count for all NTs even ones we might not expect.

`set` problemset questions are combined into Python 6: File I/O problemset.
Link to Python 6 Problem Set
Question 1-5 are all about Sets.
Question 6+ are a combination of File I/O and sets.