Python CSV Parsing

Write a Python program called "blastomatic.py" that takes a BLAST hits file (-outfmt 6, tab-delimited format) as a single positional argument and a named "–annotations" argument that is an annotations file that gives genus and species information for a given sequence ID. Check that both are actually files and die ' "XXX" is not a file' if they are not. Iterate over the BLAST hits and use the sequence ID (`saccver`) to lookup the sequence in the annotations file so that you can print out the seq ID and the percent identity (`pident`) from the hits file along with the `genus` and `species` from the annotations file.

As a BLAST tab-delimited file does not include headers, it would be helpful for you to read `blastn -help` to find what they are:

```
When not provided, the default value is:
'qaccver saccver pident length mismatch gapopen qstart qend sstart send
evalue bitscore', which is equivalent to the keyword 'std'
```

You have two "hits" files that look like this (using my "blast6chk" alias, cf notes):

```
$ blast6chk hits1.tab
// ****** Record 1 ****** //
qseqid   : NR_125480.1
sseqid   : bfb6f5dfb4d0ef0842be8f5df6c86459
pident   : 99.567
length   : 231
mismatch : 1
gapopen  : 0
qstart   : 728
qend     : 958
sstart   : 1
send     : 231
evalue   : 3.93e-118
bitscore : 422
```

The providede "centroids.csv" annotation file looks like this:

```
$ tabchk.py centroids.csv
// ****** Record 1 ****** //
centroid : e5d49c0803f04032b482a1ee836e18ab
domain   : Bacteria
kingdom  : Proteobacteria
phylum   : Alphaproteobacteria
class    : Rhodospirillales
order    : AEGEAN-169 marine group
genus    : uncultured bacterium
species  : uncultured bacterium
```

When looking up the genus and species, print 'NA' when no useable value is

present. For any sequence that cannot be found in the annotations file, print `Cannot find seq "XXX" in lookup` to STDERR.

Accept an optional "–out" argument that is the name of an output file to which to write the STDOUT of the program. If not provided, you will print to STDOUT.

The output should be tab-delimited with the fields "seq_id," "pident," "genus," and "species."

```
$ tabchk.py out
// ****** Record 1 ****** //
seq_id  : 229584169f4724188010dcfc36f2c933
pident  : 90.526
genus   : NA
species : NA
```

Expected Behavior

```
$ ./blastomatic.py -h
usage: blastomatic.py [-h] [-a FILE] [-o FILE] FILE

Annotate BLAST output

positional arguments:
  FILE                  BLAST output (-outfmt 6)

optional arguments:
  -h, --help            show this help message and exit
  -a FILE, --annotations FILE
                        Annotation file (default: )
  -o FILE, --outfile FILE
                        Output file (default: )
$ ./blastomatic.py -a foo bar
"bar" is not a file
$ ./blastomatic.py -a centroids.csv foo
"foo" is not a file
$ ./blastomatic.py -a centroids.csv hits1.tab -o out 2>&1 | head
Cannot find seq "875518c5d2436c94f50924425cb37f42" in lookup
Cannot find seq "2e5eeadcccb672a3410ddc6a8ff9ceee" in lookup
Cannot find seq "e16e05492dbcdbeb1de332614d5d002d" in lookup
Cannot find seq "39491c3b0dce84b718a274eafff3915c" in lookup
Cannot find seq "f42d5121911f169e12fd4c6bac1977f3" in lookup
Cannot find seq "1caa4b8dabc32ca88ce99513239e0a45" in lookup
Cannot find seq "e064229aac7487f068c9b8abf4a741e0" in lookup
Cannot find seq "661c26e0a8ac2956e6ba5b52dcaf11f2" in lookup
```

```
Cannot find seq "b4cd45a37eefcc49e5e9e153dffa783d" in lookup
Cannot find seq "197b74f559ec647315375dd5588792f3" in lookup
$ ./blastomatic.py -a centroids.csv hits1.tab 2>err | head | column -t
seq_id                              pident  genus            species
bfb6f5dfb4d0ef0842be8f5df6c86459  99.567  Prochlorococcus  MIT9313  NA
0dab11245fb6fe800362cdc20953d0f6  98.701  Prochlorococcus  MIT9313  Ambiguous_taxa
9c2271504f3393684fd1ed93d1d1a9ab  98.701  Prochlorococcus  MIT9313  Ambiguous_taxa
26cbd1b8b6fcd255774f4f79be2f259c  98.701  Prochlorococcus  MIT9313  NA
6192b152a8c84ff13fe6a7dced9c9357  98.268  Prochlorococcus  MIT9313  NA
61d060a46dadd0fbcdb099bbf4a36221  98.268  Prochlorococcus  MIT9313  NA
6da08abcdd74ae66dd2ef4112384faa5  98.268  Prochlorococcus  MIT9313  Ambiguous_taxa
50d394faf698e238e9bd05b251499cee  97.835  Prochlorococcus  MIT9313  NA
1642658999590e25a39926d281dea501  96.537  Synechococcus    CC9902   NA
```

Test Suite

A passing test suite looks like the following:

```
$ make test
python3 -m pytest -v test.py
============================= test session starts ==============================
platform darwin -- Python 3.6.8, pytest-4.2.0, py-1.7.0, pluggy-0.8.1 -- /anaconda3/bin/pyth
cachedir: .pytest_cache
rootdir: /Users/kyclark/work/worked_examples/07-csv, inifile:
plugins: remotedata-0.3.1, openfiles-0.3.2, doctestplus-0.2.0, arraydiff-0.3
collected 4 items

test.py::test_usage PASSED                                               [ 25%]
test.py::test_bad_input PASSED                                           [ 50%]
test.py::test_good_input1 PASSED                                         [ 75%]
test.py::test_good_input2 PASSED                                         [100%]

=========================== 4 passed in 0.42 seconds ===========================
```