Lab: Parsing Json

Part II: last time we wrote a `.py` file that searched for an AA sequence in a .json file and returned it

```
P = ParsingJsonExtractAA()
```

```
P.ExtractAA()
```

```
method: ratio

index and match:

411    sequenceCwu
dtype: object
------------------------------


------------------------------

method: partial_ratio

index and match:

411              sequenceCwu
412    sequenceListNewRules
413    sequenceListOldRules
414       sequencesListText
dtype: object
------------------------------


------------------------------

method: token_sort_ratio

index and match:

411    sequenceCwu
dtype: object
------------------------------


------------------------------

method: token_set_ratio

index and match:

411    sequenceCwu
dtype: object
------------------------------


------------------------------
```

Inde:	Type	Size	Value
0	str	16	QVQLQESGPGLVKPSQ
1	str	11	GSISSGGYYWS
2	str	14	WIRQHPGKGLEWIG
3	str	16	SIYYSGSTYYNPSLKS
4	str	16	RVTISVDTSKNQFSLK
5	str	11	AREGYHSGMDV
6	str	11	WGQGTTVTVSS
7	str	16	QVQLQESGPGLVKPSQ
8	str	16	LKSRVTISVDTSKNQF
9	str	16	DIQMTQSPSSLSASVG
10	str	11	QASQDISNYLN
11	str	15	WYQQKPGKAPKLLIY
12	str	7	DASNLAT
13	str	16	GVPSRFSGSGSGTDFT

Lab: Parsing Json

Part II: Now we want to return a summary containing all the information about the sequence **(both: DNA and AA)**

for key = *'sequenceListNewRules'*

```
Out[8]:
{'type': 'AA',
 'Sequence': 'WGQGTTVTVSS',
 'ID': '10711PRT',
 'Info': 'Artificial SequenceVH FR4 of Antibodies SIRPAB-1 to SIRPAB-5, and SIRPAB-17 to SIRPAB-21 7T'}
```

Summary of all the sequences found:

```
Sequence at location 197 could not be identified!
Check manually!
Sequence at location 200 could not be identified!
Check manually!
Among 225 records
187 AA sequences and 36 DNA sequences were found.
```

```
{'type': 'DNA',
 'Sequence': ['CAGGTGCAGC',
  'TGCAGGAGTC',
  'GGGCCCAGGA',
  'CTGGTGAAGC',
  'CTTCACAGAC',
  'CCTGTCCCTC',
  '60ACCTGTACTG',
  'TCTCTGGTGG',
  'CTCCATCAGC',
  'AGTGGTGGTT',
  'ACTACTGGAG',
  'CTGGATCCGC',
  '120CAGCACCCAG',
  'GGAAGGGCCT',
  'GGAGTGGATT',
  'GGGTCAATCT',
  'ATTACAGTGG',
  'GAGCACCTAC',
  '180TACAACCCGT',
  'CCCTCAAGAG',
  'TCGAGTTACC',
  'ATATCAGTAG',
  'ACACGTCTAA',
  'GAACCAGTTC',
  '240TCCCTGAAGC',
  'TGAGTTCTGT',
  'GACCGCCGCA',
  'GACACGGCGG',
  'TGTACTACTG',
  'CGCCAGAGAG',
  '300GGATACCACT',
  'CAGGAATGGA',
  'CGTATGGGGC',
  'CAGGGAACAA',
  'CTGTCACCGT',
  'CTCCTCA'],
 'ID': '108357DNA',
 'Info': 'Artificial SequenceVH Nucleotide Sequences of Antibody SIRPAB-1 8c'}
```

Lab: Parsing Json

Part II:

```
Sequence at location 197 could not be identified!
Check manually!
Sequence at location 200 could not be identified!
Check manually!
Among 225 records
187 AA sequences and 36 DNA sequences were found.
```

```
Out[8]:
{'type': 'AA',
 'Sequence': 'WGQGTTVTVSS',
 'ID': '10711PRT',
 'Info': 'Artificial SequenceVH FR4 of Antibodies SIRPAB-1 to SIRPAB-5, and SIRPAB-17 to SIRPAB-21 7T'}
```

- work in groups of 4 – 5 during Lab
- submit code (individual copy each student) by Friday, 13th
- if code returns correct information: 25pts
- hint I: start with `ParsingJsonExtractSeqs_Students.py`
- hint II: protein sequences are indicated with 'PRT' and DNA sequences with 'DNA'
- hint III: try to understand what `Seq_regex = re.compile(r'(\d+(?:PRT|DNA))')` does when applied to Seq
- hint IV: explore what `re.search` does, especially `group()` and `span()`
- optional: make the code work for `key = 'sequenceListOldRules'` too **(5 extra points)**