Assignment 4 - Project Tips

For the function get_list_of_university towns, the data is in the text file *university_towns.txt* is in the following format:

StateOne[edit]

RegionOne (UniversityOne)[1]

RegionTwo (UniversityTwo)

RegionThree (UniversityThree)[2]

The question asks for this data to be cleaned up and returned in a DataFrame of this format:

	State	RegionName
0	StateOne	RegionOne
1	StateOne	RegionTwo
2	StateOne	RegionThree

So for example if you have:

Alabama[edit]

Auburn (Auburn University)[1]

Florence (University of North Alabama)

Jacksonville (Jacksonville State University)[2]

This needs to be converted into the dataframe:

	State	RegionName
0	Alabama	Auburn
1	Alabama	Florence
2	Alabama	Jacksonville

Some tips for cleaning the data:

1. For "State", removing characters from "[" to the end.

2. For "RegionName", when applicable, removing every character from " (" to the end.

3. Depending on how you read the data, you may need to remove newline character '\n'.

There may be a few RegionNames in the list that don't match up well to actual cities, but don't worry about doing anything special for these entries for the purposes of this assignment. Punctuation at the end of the RegionName not encapsulated in the parenthesis does not need to be removed.

An example:

The RegionName given to us by the text file:

The Five College Region of Western Massachusetts:

Should be unmodified:

	State	RegionName
184	Massachusetts	The Five College Region of Western Massachusetts:

Good luck on the project!