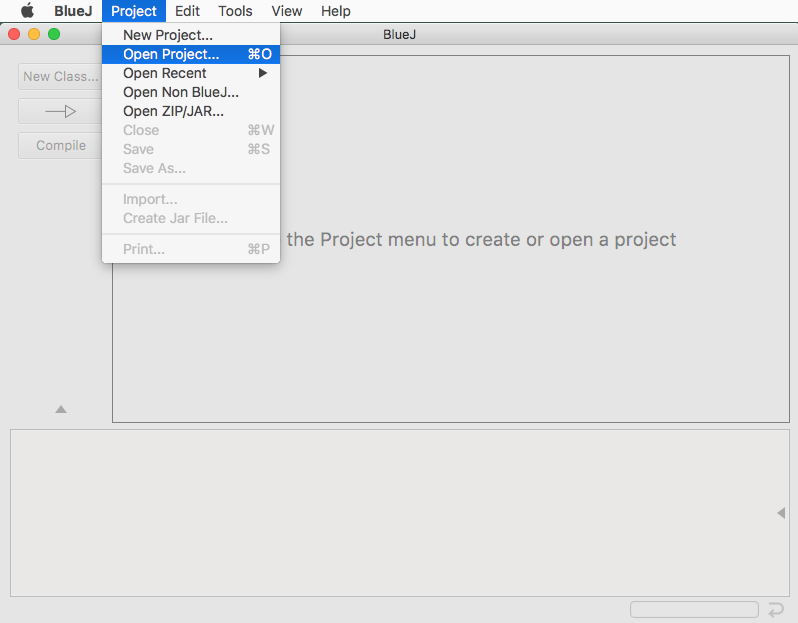
**Step 1 – Installation and Setting**

1. For Windows, download and install “BlueJ-windows-411-duke.zip”.

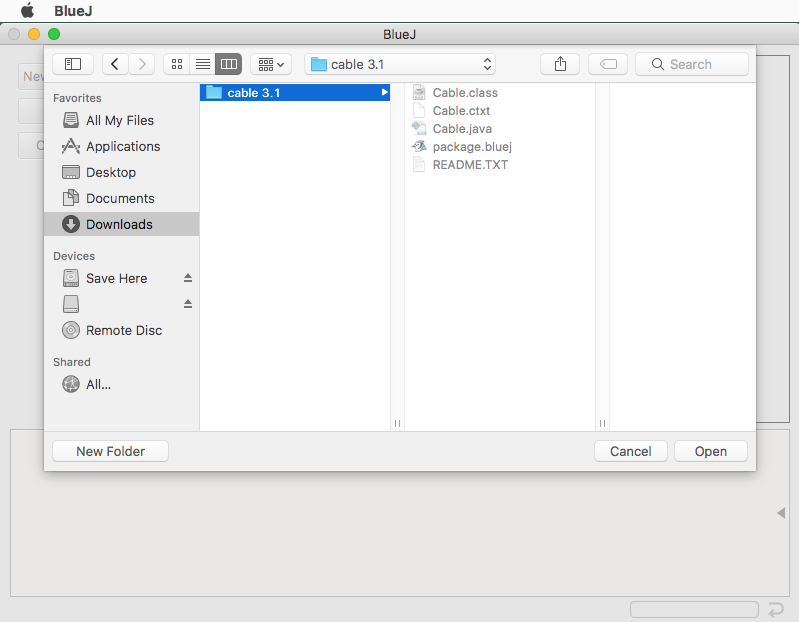
For Mac OS, download and install “BlueJ-mac-411-duke.msi”.

(Note: The following tutorial is based on the Mac OS. For Windows, the interface may look different, but the steps are the same.)

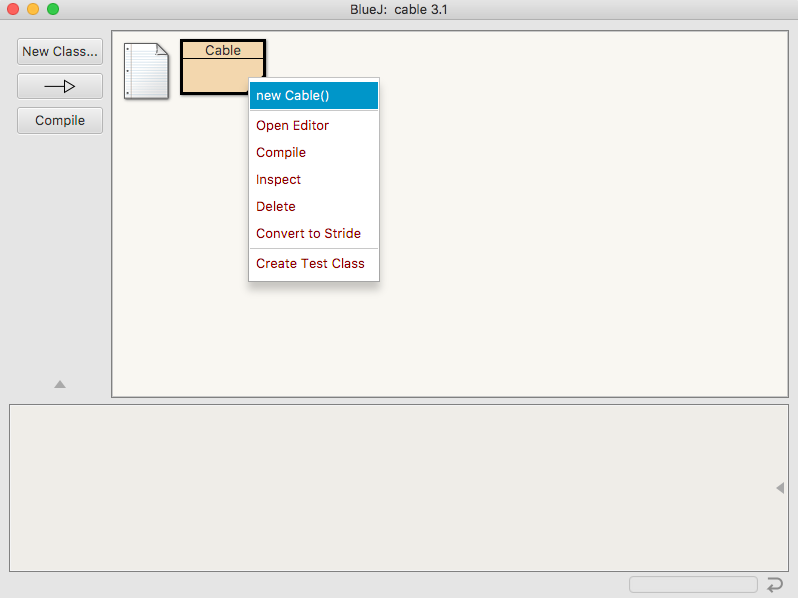
1. Download the folder “cable 3.1”.
2. Open BlueJ. Go to “Project” → “Open project”.



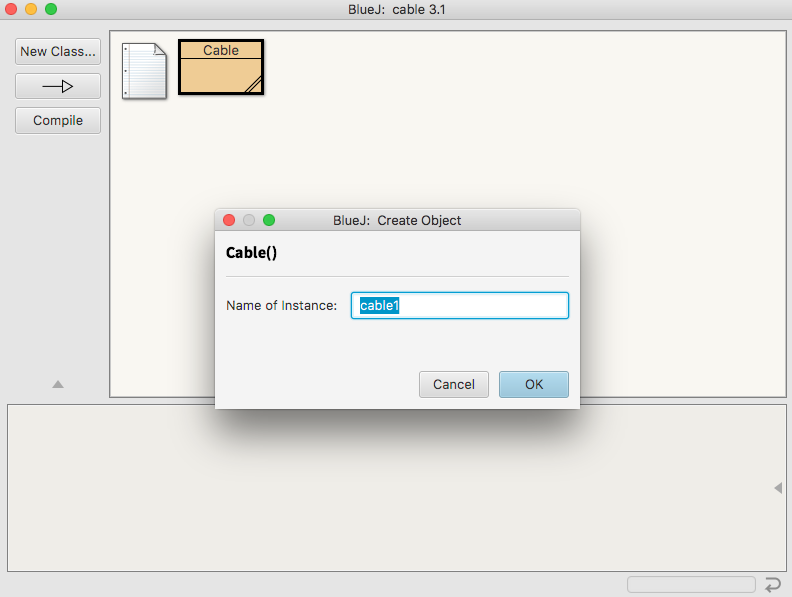
Open the folder “cable 3.1”, which we just downloaded.



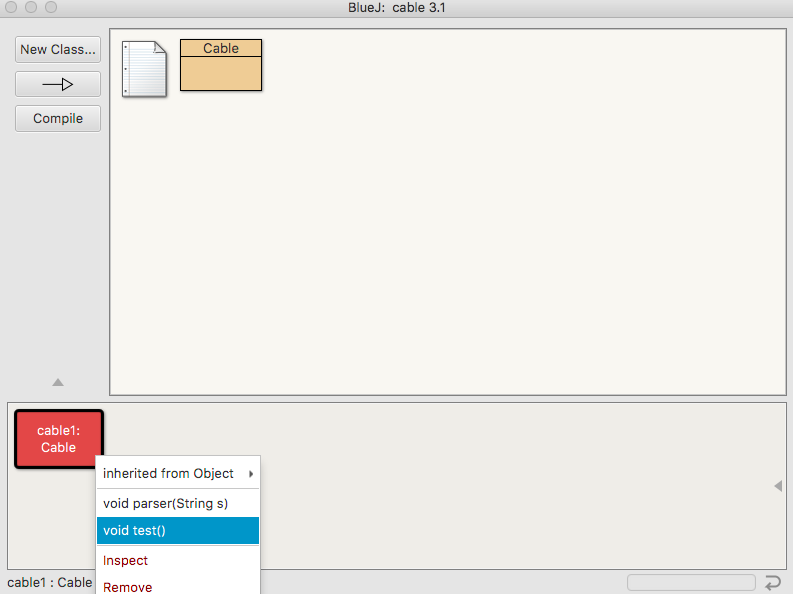
1. Now in the BlueJ panel, the project “cable” is added. Right click on it and choose “New Cable()”.



A window will pop up. Click “OK”.

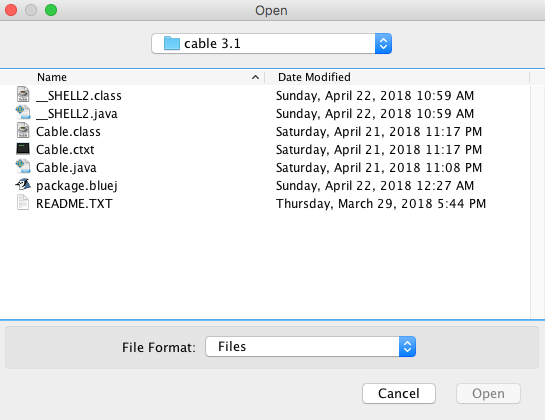


1. An object of Cable() is now created (the red rectangle). Right click on it and choose “void test()”.

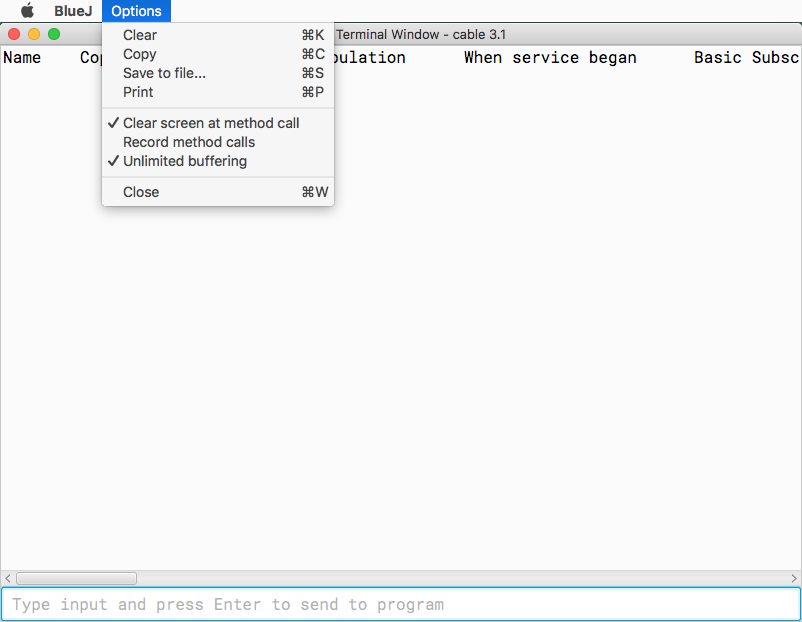


1. Two windows will pop up. Click “Cancel” for the one that prompts you to select a file.

(Note: This is how we will select a file and run the program later. For now, we are simply setting up, and after clicking “Cancel”, there will be an error message in the output window – nothing to worry about.)



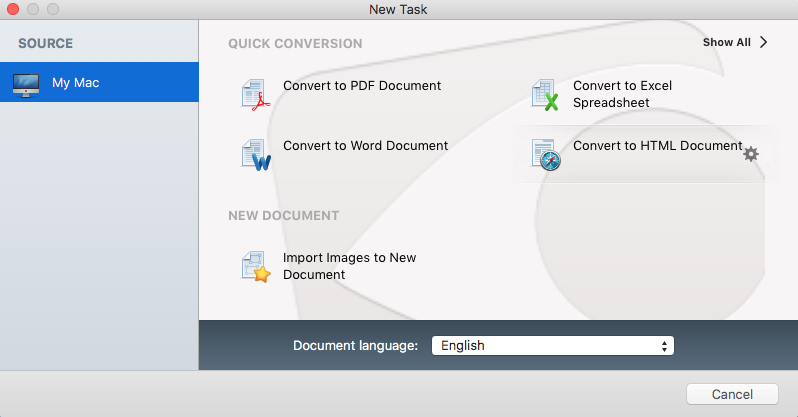
In the window “BlueJ: Terminal Window – cable 3.1”, go to “Options” → select “Clear screen at method call”, and “Unlimited buffering”.



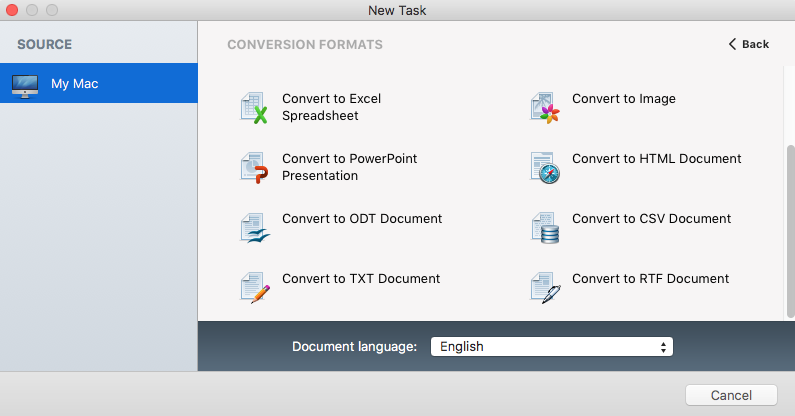
After that, close the window.

**Step 2 – Converting PDF to TXT with ABBYY Finereader**

Open the Finereader, and click “Show All”



Scroll down, and select “Convert to TXT Document”. Then, select the PDF file that we want to convert.

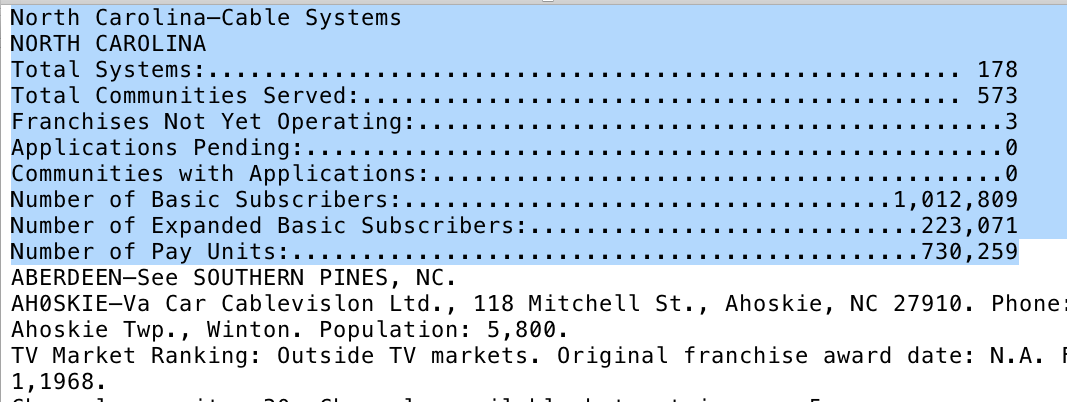


The conversion process may take a while. After it is done, it will guide you to export the TXT file.

**Step 3 ­– Clean up the TXT file**

OCR software is not 100% accurate, and certain elements in TXT file are crucial for the program to run smoothly. Thus, we have to manually fix the ones that are not correctly recognized.

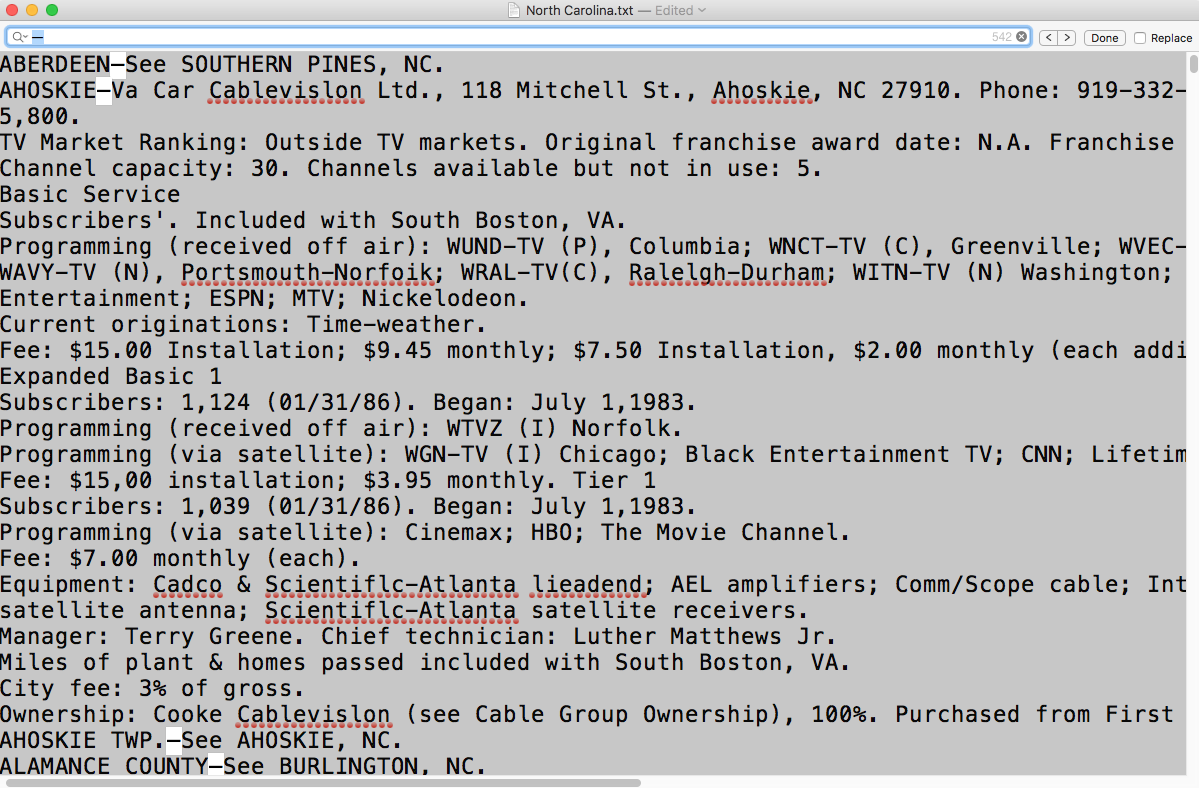
1. Some backgrounds: Each town name is followed by an em dash (—), which is different from the hyphens (-) that appear elsewhere (in phone number, programs, etc.) in this book. The algorithm for this program is entirely based on this difference. Unfortunately, the OCR software occasionally misreads a dash as a hyphen, and we need to manually fix them in the text editor. We also need to fix some other formatting issues along the way.
2. Open the TXT file and delete the highlighted part – the first word in this TXT file should be a town name.



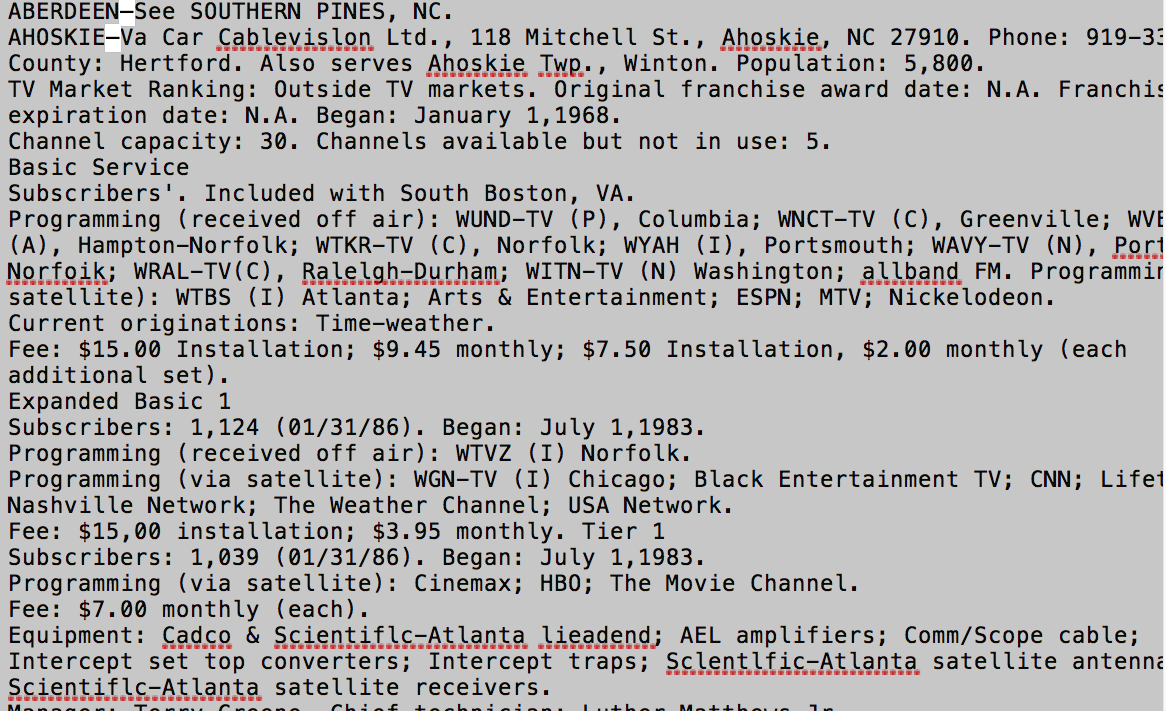
1. In the Mac OS text editor, dashes and hyphens look almost the same.

In order to check if dashes are correctly recognized, we will use the search function provided by the text editor (“Command + F” in Mac OS, *or* “Ctrl + F” in Windows), and search for the dashes.

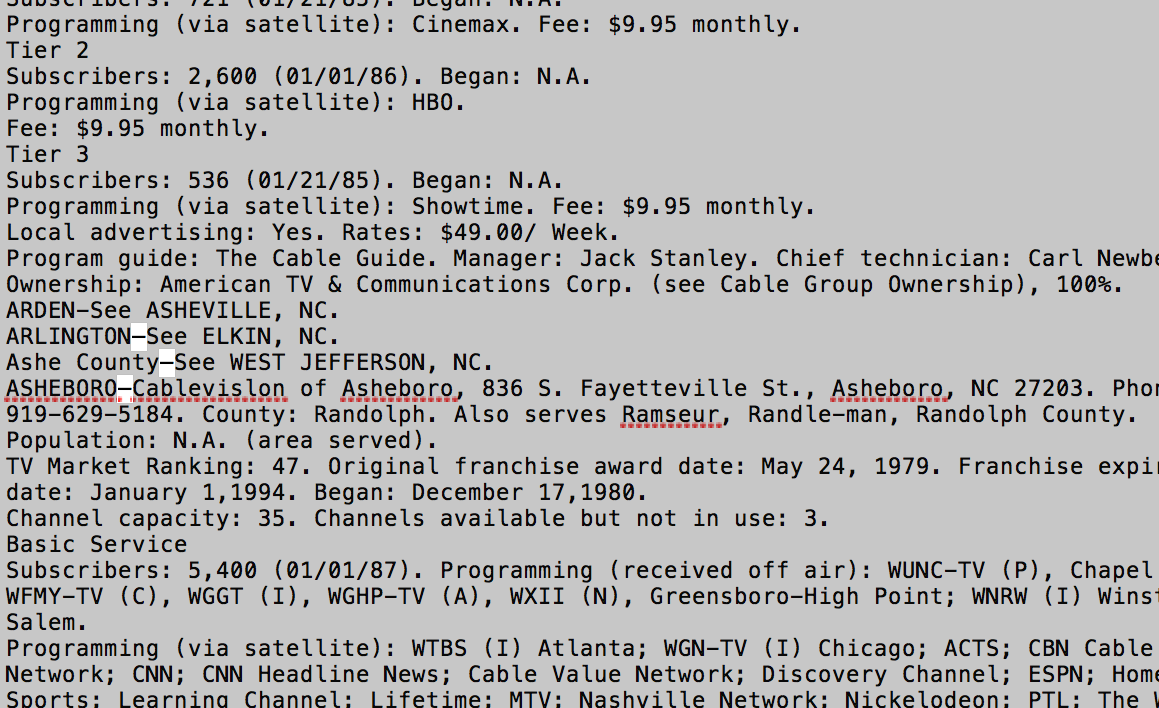
(Note: To type a dash - “Shift + alt + -” in Mac OS)



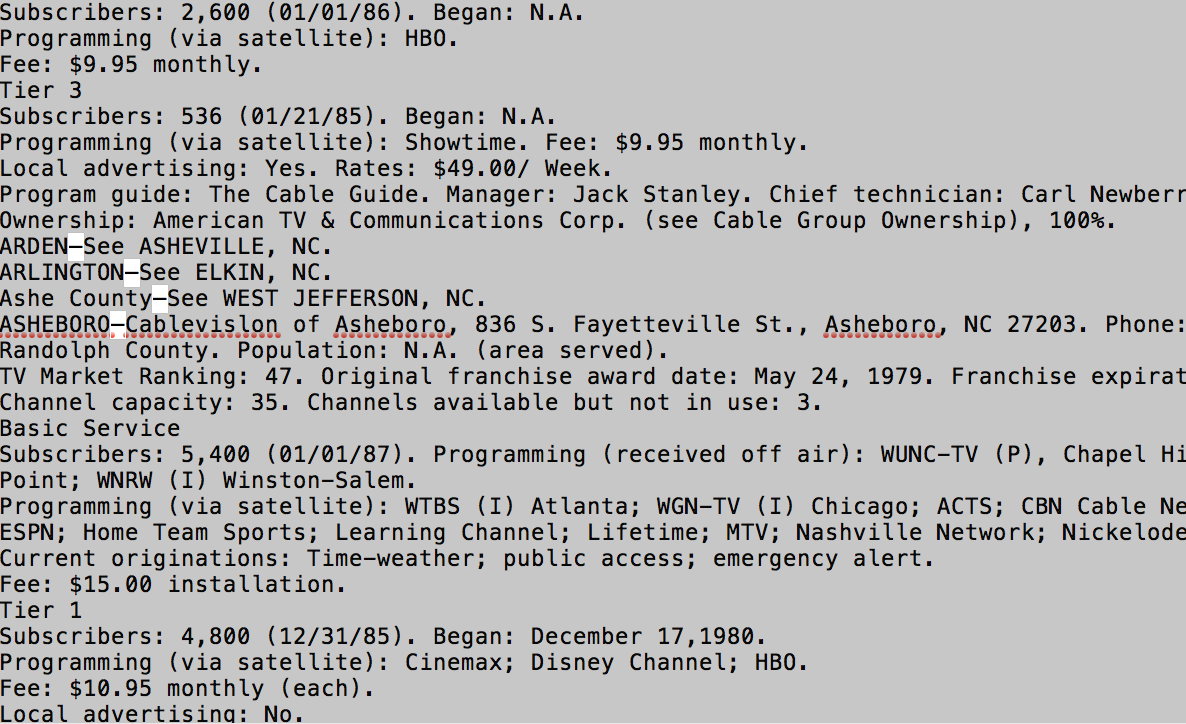
1. The search function will highlight all the dashes.



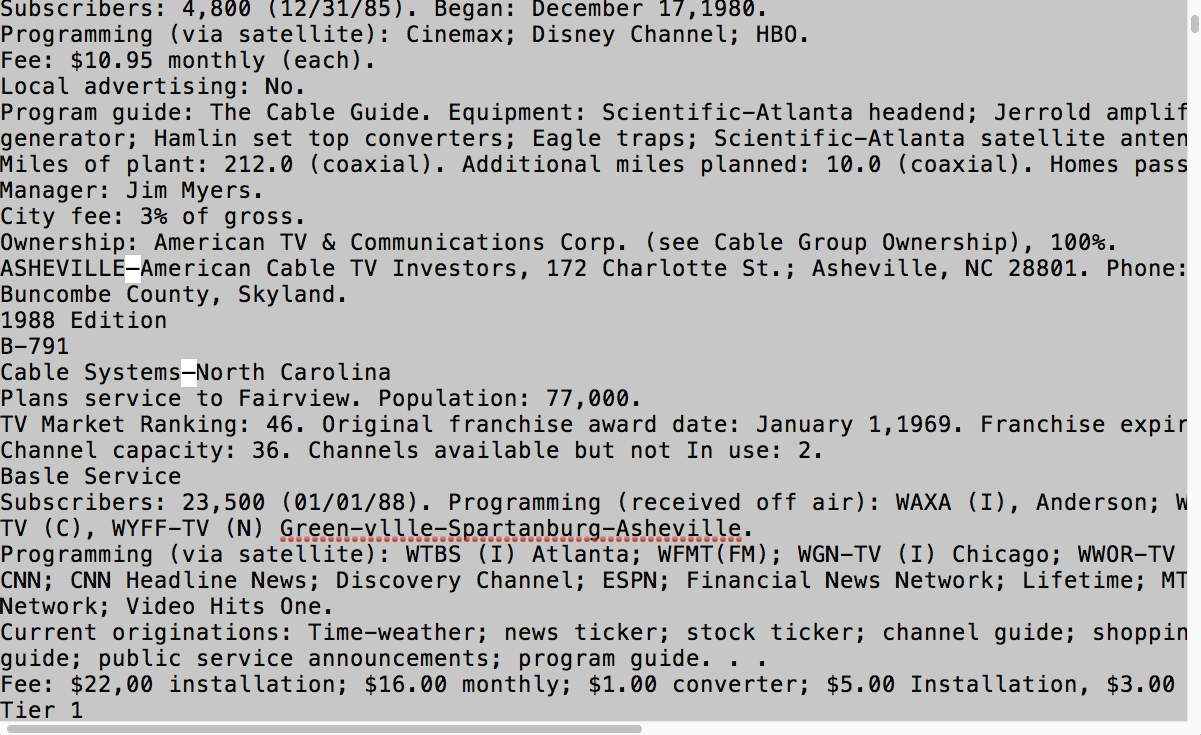
When we see a town name followed by a dash that is not highlighted (it is actually a hyphen), we need to replace it with a long em dash.



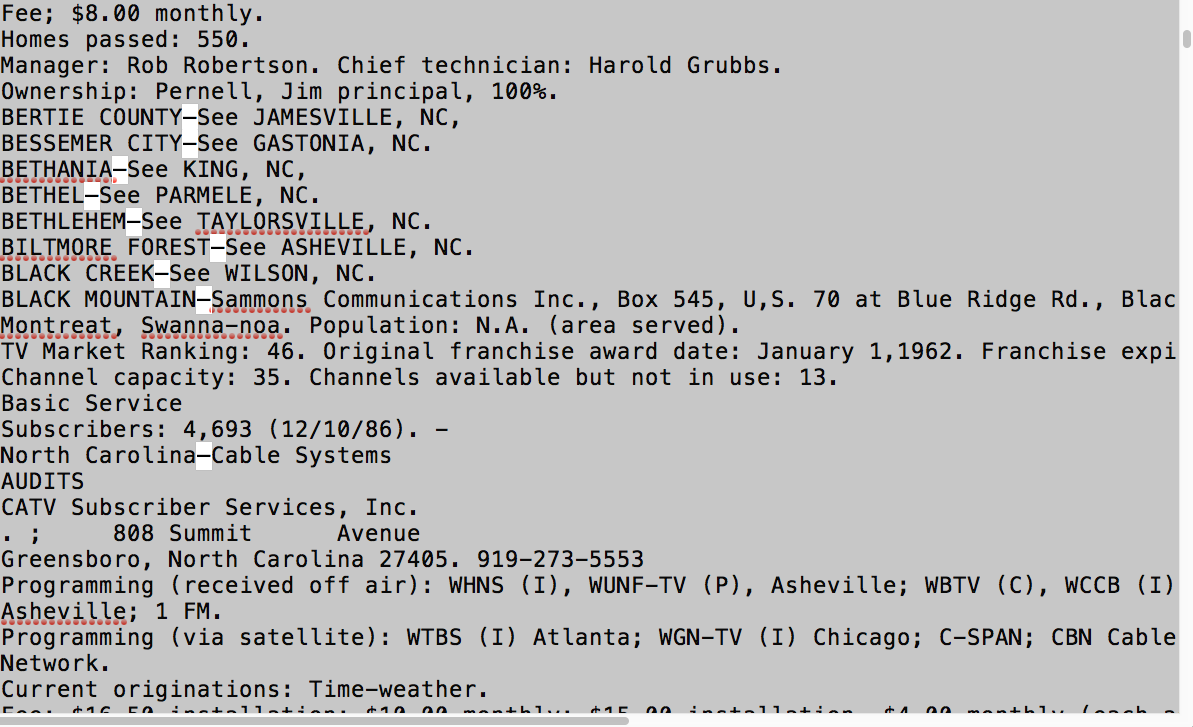
After replacing it, it will be highlighted.



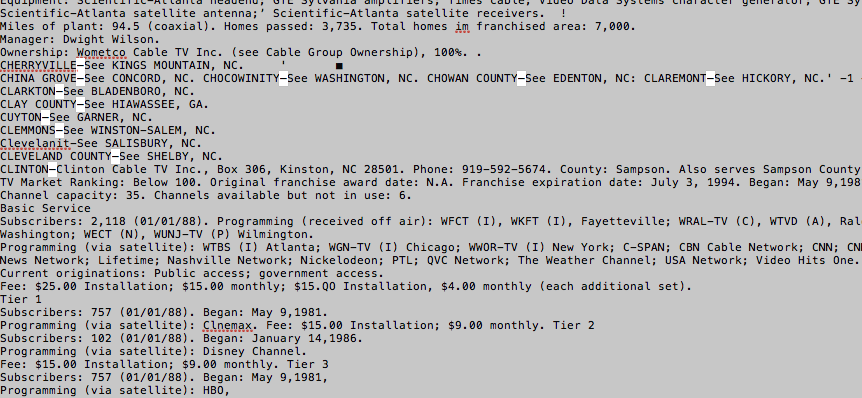
1. Continue the process. Since we are “eyeballing” the mistakes, it is likely that we will miss some of them. That is alright. When we run the program later, the program will run into an error whenever there is a formatting issue, and we will have to come back to the TXT file to fix it. The more mistakes we identify now, the easier it will be for us to run the program later.
2. There are some other formatting issues that we need to fix along the way:
   1. There are dashes included in the headers of the book. We need to delete them. (i.e. The only place we want a dash to appear is after a town name). We can simply delete everything in the red rectangle.



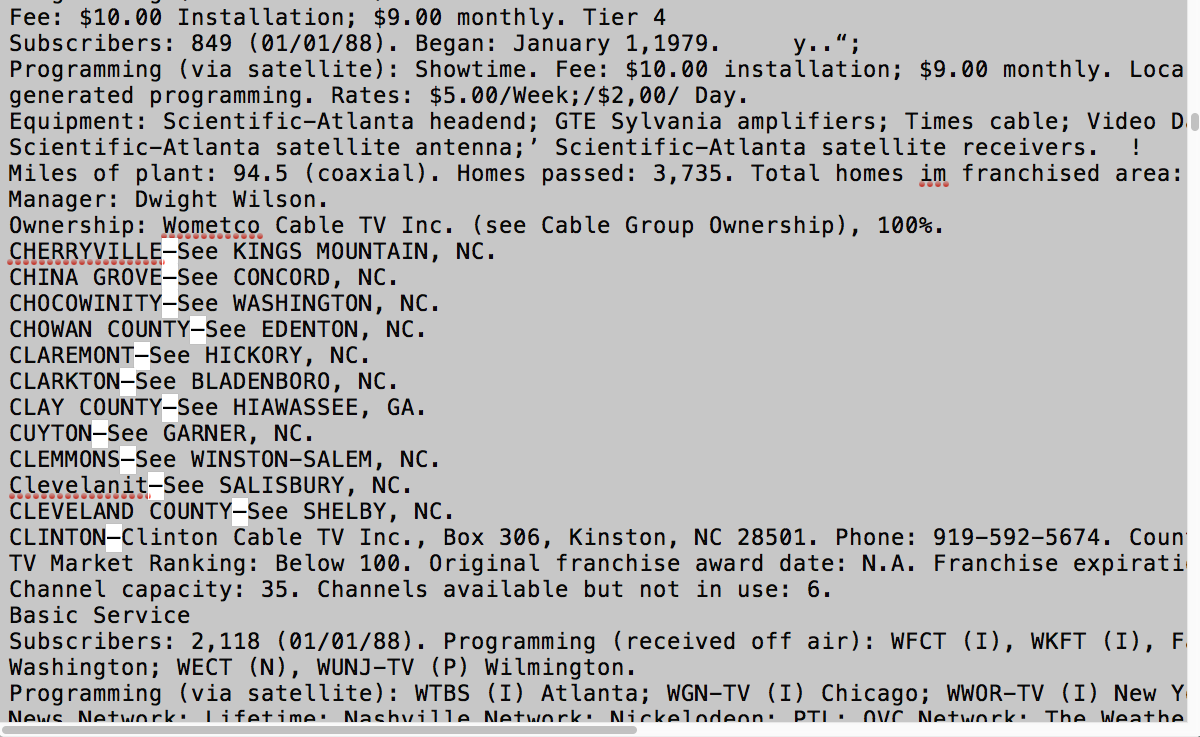
* 1. The OCR may misread a period (.) as a comma (,) at the end of a record, especially if it is copied. The period (.) at the end is also crucial for the algorithm. Again, it is hard to find these mistakes, and the program will complain if we miss them. However, keep an eye on them along the process will make our job easier later.



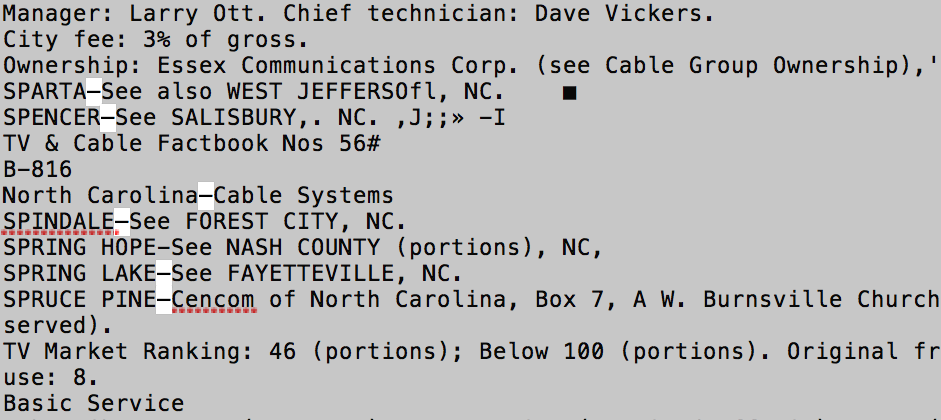
* 1. Some records (especially the copied ones), are clustered on one single line. This should not be a problem for the program, but this is usually a sign that the OCR is doing a poor job at this part. As we see in this example there are multiple problems.



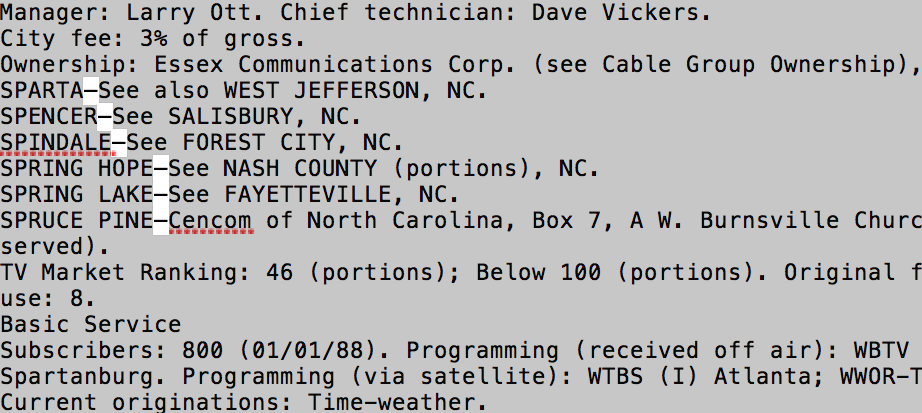
I would suggest separating them and starting a new line for each town.



* 1. Also, try to clean up things that look weird and unnecessary. You can almost delete anything that is not part of the information of our interest. For example:



After cleaning up (Also deleted the header):



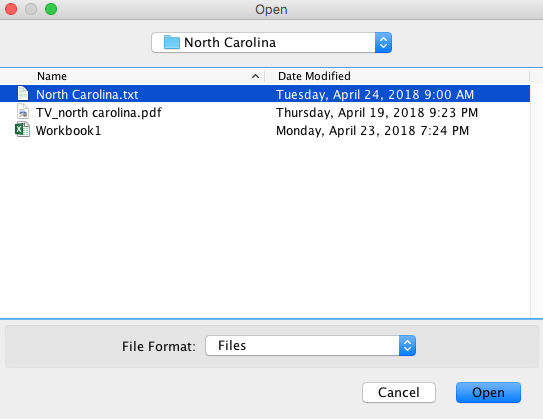
**Step 3 – Running the program**

1. If you still have BlueJ opened, repeat 5 in **Step 1**.

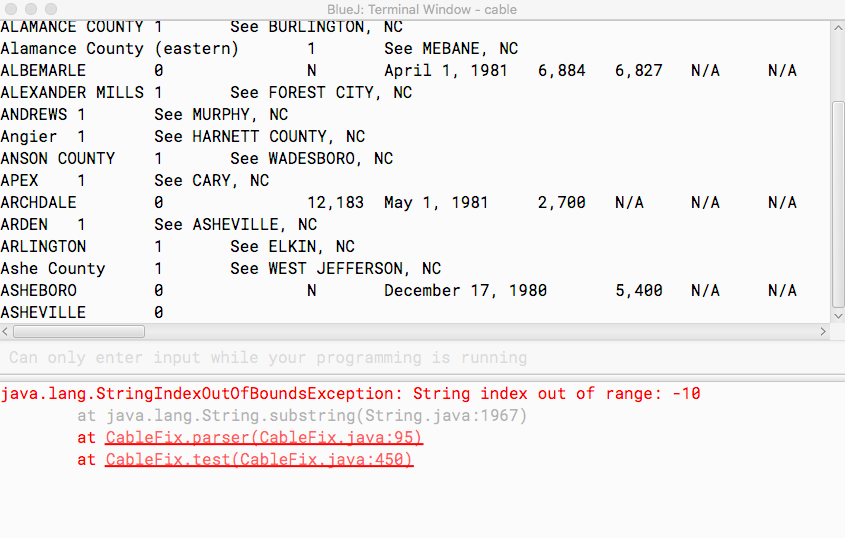
Otherwise, Open BlueJ and repeat 4 and 5 in **Step 1**.

1. Now, instead of clicking cancel, we will select the TXT file that we just cleaned up.

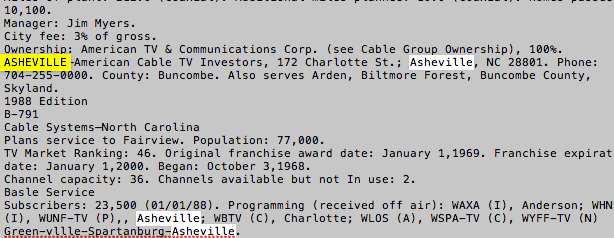
And the program will start running.



1. Almost for sure, we will run into error(s). This is because we have missed some formatting mistakes in the TXT file.

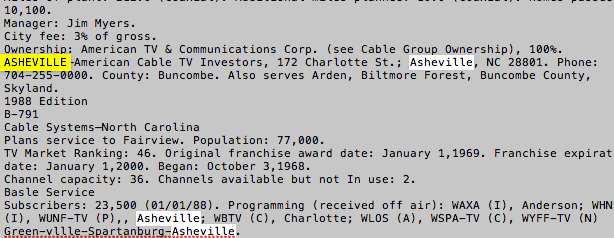


1. Ignore the error message. What we will do is to search for the town where it stops in the TXT file, and figure out what goes wrong. In this example, we will search for “ASHEVILLE” in the TXT file.



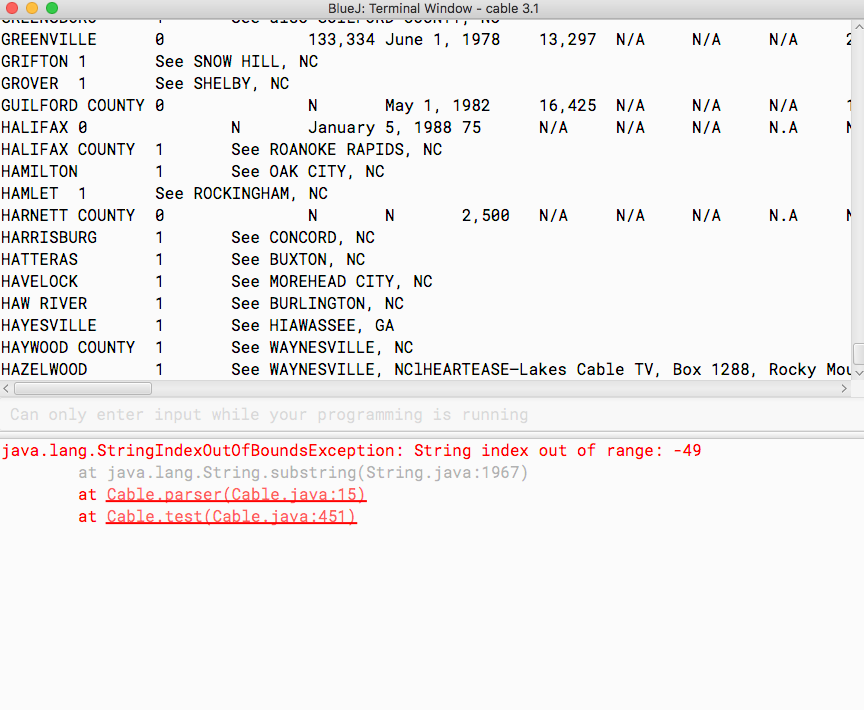
It turns out that we forgot to delete header part.

(The header contains an em dash, and we only want it to appear after a town name)

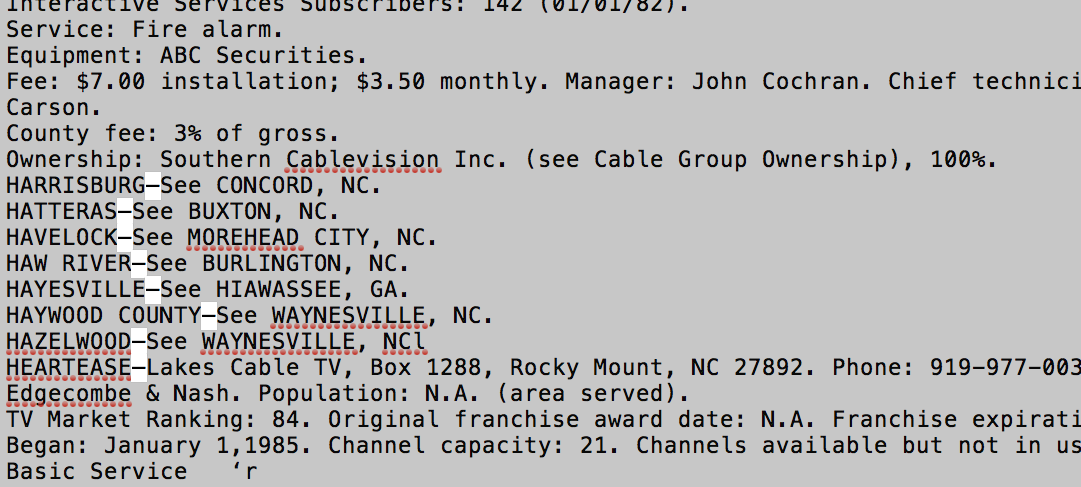


1. After we delete it and save the TXT file, we will run the program again. (repeat 5 in step 1)

Unsurprisingly, we run in to an error again. However, this time it stops at a different town, which means that the previous problem is fixed, and this is a new formatting issue in the TXT file. We will go back to the TXT file again and see what the problem is.



The “l” after “NC” should be a period (.) instead.



Here are some common formatting issues you might run into during this process:

1. The period at the end of each record is missing or misread as something else, such as a comma (,).
2. The long em dash after a town name is misread as a short hyphen.
3. A long em dash appears somewhere else other than after a town name.
4. Continue this process until the program does not generate any error message.

**Step 5 – Cleaning up in Excel**

1. Select all in the output window, and copy-paste it to an Excel spreadsheet.

Some shortcuts:

Select All – Command + A (Mac OS) *or* Ctrl + A (Windows)

Copy – Command + C (Mac OS) *or* Ctrl + C (Windows)

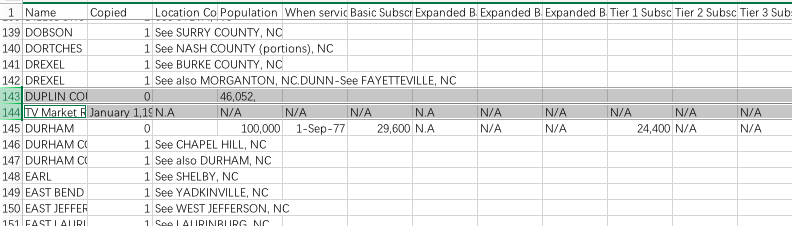
Paste – Command + V (Mac OS) or Ctrl + V (Windows)

1. At this stage, there still exist some formatting issues which are undetected by the program. However, we can identify them easily in a spreadsheet.

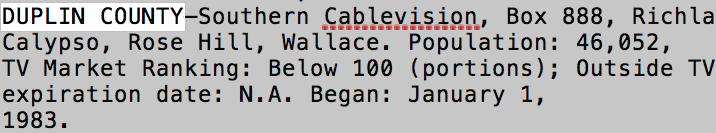
Depending on what kind of issues they are, we can either fix them directly in Excel, or fix them in the TXT file and run it again later when everything is fixed.

Here are some common issues that you may observe:

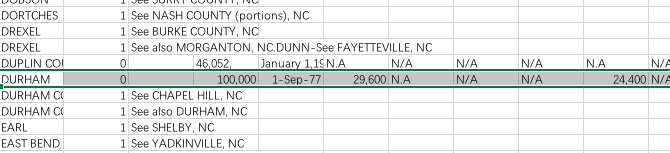
* 1. A new line with unrelated words or information included.



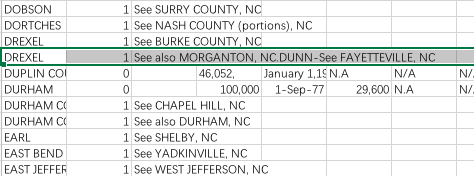
This happens because in the TXT file, the population number should be followed by a period(.) instead of a comma (,)



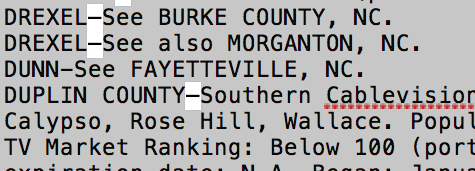
To fix this, we can either fix the TXT file and re-run it later, or, since the rest of the information should be accurate, simply move the second line above and delete the unnecessary part.



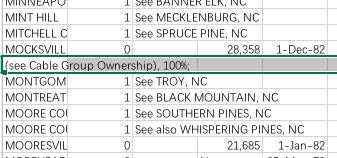
* 1. Two towns on the same line.



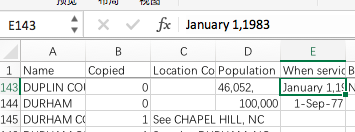
Again, this should be one of those “punctuation problems”. In this example, the issue is caused by the short hyphen.



* 1. An extra line that is irrelevant – just delete it.



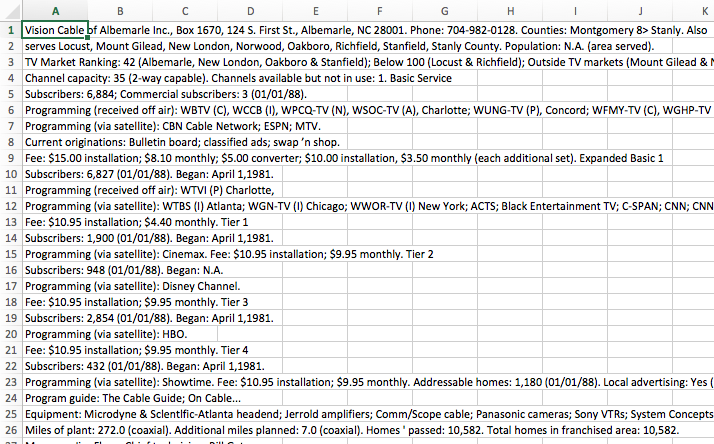
* 1. Spelling mistakes, mostly caused by the OCR software.



Other common OCR spelling problems:

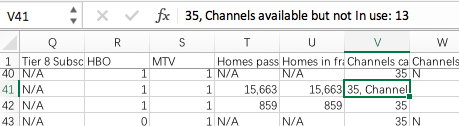
* “TWP.” in a town name is very likely to be recognized as “TOP.” in the TXT file
* If a town name contains “PITT”, it is likely to be recognized as “Pin”. A town name is usually all in capital letters, so this should be quite obvious.

1. If you see an entire block of original text in Excel for a town, usually it is also caused by the wrong em dash/hyphen after the town name.

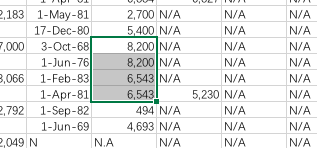


1. After fixing all the problems in Excel (or re-running the program), sort by column “copied” → “Sort smallest to largest” → “Expand the selection”
2. Some final sanity check suggestions:
   1. Skim through the “Population” column, and check if there are unusual numbers, such as two-digit numbers (a town is unlikely to have such a low population). This might be a case of punctuation issue, where the comma (,) in the number is misread as a period (.), and only the part before the period in the number is displayed.
   2. Irrelevant punctuations or information, such as the comma (,) in cell 143D in the previous picture.

Or something like this, which is caused by the comma (,) after number 35. It should be a period (.) instead:

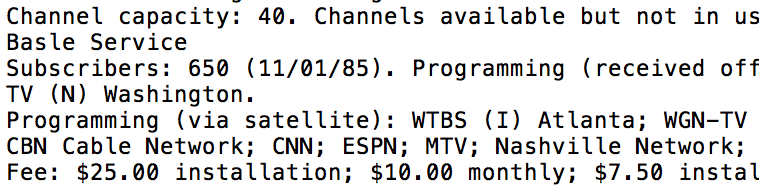


* 1. Skim through the columns for “Basic Subscribers”, “Homes passed”, “Homes in franchised area”, “Channels capacity”, and “Channels available but not in use”. If you see two consecutive towns that have the same number, it is suggested to refer back to original PDF and check if that is actually the case. For example,



This is happening because the program looks for key words to identify the information. If we go back to TXT file, the word “Basic Service” for these towns are misspelled, which leads to the problem. At this point, it is still challenging for the program predict these mistakes, and we could only identify them manually.

(Note: It turns out that “Basic” is recognized as “Basle” by OCR quite often. We could search for “Basle” in the TXT file and replace them.)



Thanks for reading. Both the program and this manual are written hastily, and there might be mistakes. Please feel free to contact me ([yucai@vassar.edu](mailto:yucai@vassar.edu)) should you run into any problems.