**Background:**

I have attached a text file, which I have converted from a Word document. It was in an inconsistent text + tables format in a Word document. This data contains the cutoff ranks and percentiles for engineering colleges in one state of India. Since I'm building software to utilise this data for writing some algorithms around it, I want this data in Excel format, so that it can be imported into a database.

**Task:**

This text file has a particular pattern which you need to follow for this extraction, and put all the data in the below columns of Excel as per the logic mentioned by me:

- Sr no

- Institute Code

- Institute Name

- District

- Branch Code

- Branch Name

- Institute Status

- Seat Description

- Stage

- Seat Category

- Rank

- Percentile  
  
**Logic:**You need to parse the file top to bottom. All the pattern recognition logic depends on this.

**First Line:  
*<institute\_code> - <institute\_name>, <district\_name>***

*Parameters:*

***Institute\_code:*** *5 digit code*

***Institute\_name:*** *String name, it will be space separated, read it until a comma comes up*

***District:*** *the string after comma and space is district (This is optional field, sometimes it won’t be there) In that case, you need to capture institute name till the end of the line*

*E.g., 01002 - Government College of Engineering, Amravati*

*In this* ***Institute\_code*** *is 01002****,*** *the* ***institute\_name is*** *Government College of Engineering and the* ***District*** *is Amravati*

**Second Line:  
*<branch\_code> - <branch\_name>***

***Branch\_code:*** *Generally 10 digit code, like 0100219110*

***Branch name:*** *comes after hyphen*

*E.g., 0100219110 - Civil Engineering*

*In this, 0100219110 is the branch code and Civil Engineering is the branch name*

**Third Line:**

**Status:** Government Autonomous

Simple, add this in the **Institute Status** column directly.

**Fourth Line:**

Status Level or some text like *Home University Seats Allotted to Home University Candidates*

*Add directly in the Seat Description column*

**Fifth Line:** **Stage**

*This is keyword or indication that seat category names are coming below. All the words below will be categories.*

The above 5 lines are fixed. After this, there will a set of categories and their rank/percentiles. Now look at the example below:

| GOPENS GSCS GOBCS GSEBCS LSTS LVJS LNT2S LSEBCS PWDOPENS ORPHAN EWS  I 28591 (90.4057549) 45021 (84.5242070) 35751 (87.9526086) 97487 (63.0701676) 145122 (34.5790896) 66354 (76.5404634) 72774 (73.8893284) 142686 (36.0381755) 168151 (14.0490445)  76123 (72.6095528)  VII          38929 (86.7801974) |
| --- |

1. Now you can see, there are categories from GOPENS to EWS. After that, there comes a value of stage - “I”. It could be I, II, VI, VII or a “Roman Digit-<text>”. For example, I-Non PWD.
2. This stage name must be added in the stage columns for all the categories listed above.
3. Now, all the numerical values coming below are Ranks and Percentiles. The percentiles are in the brackets. E.g., 143909 (39.4720482) where 143909 is the rank and 39.4720482 is the percentile.
4. **Mapping of category -> rank+percentile:** Since we already have a list categories captured for this stage, all the rank-percentiles will be mapped to those categories sequentially, i.e., category in first position will have to be mapped with the cutoff rank and percentile of the first position only.

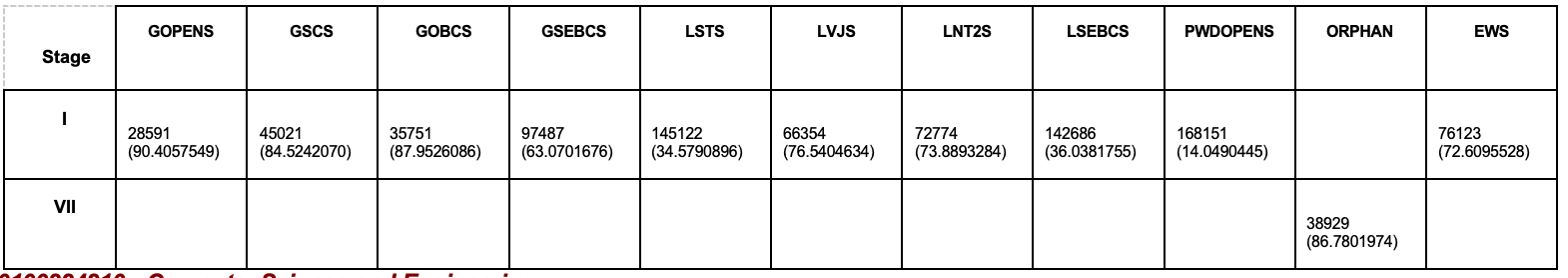
**The logic behind multiple stage captures:**

Now, if you carefully look at rank-percentile pairs, you will observe a gap between 9th and 11th pair. This is blank cell in the table. This blank line indicates that 10th number seat category didn’t get alloted in this stage, but in later of the stages in same branch.

Just after 11th pair which is “76123 (72.6095528)”, there is next stage coming : “VII”. So the 10th number category name should be mapped to this percentile (because 10th number was an empty line)

So, the last line **“38929 (86.7801974)”** after VII will come in the 10th number category, which is **ORPHAN** in this case.

Also note that, there are 9 empty lines before this rank-perc pair because the 9 pairs above it were part of previous stage.

So basically, an empty line is an indication skipped/blank cell for that stage. Look at table below from the original document:  
  


***BUT THIS “EMPTY LINE” CONDITION ONLY HAPPENS WHEN WE ARE INSIDE SOME INSTITUTE BRANCH. SOMETIMES, EMPTY LINE MIGHT ALSO COME WHEN BRANCH OR COLLEGE NAME IS ENDING. WE NEED TO ADD A CHECK WHERE THIS DOESN’T AFFECT ON THAT.***