Finding PDF about taxi licence for Bucuresti : PDF at <http://www.pmb.ro/adrese_utile/transport_urban/autorizatii_taxi/autorizatii_TAXI.php>

Thinking about application + web site about this -with a simple search and possible OCR ( take image of the car plate and recognizing number) . Also for international ( GPS for phones to know the city). Also IOT for monitoring the illegal taxi movement( could be also illegal to do this)

The Proof of Concept: A WebAPI + WebSite( that can be accessed also by mobile) to enter a plate number and find if it is legal. Maybe later an Android App.

***Step 1*** : acquiring data from <http://www.pmb.ro/adrese_utile/transport_urban/autorizatii_taxi/autorizatii_TAXI.php>

. Download situatia\_autorizatiilor\_taxi\_20171208.PDF

***Step 2*** : cleaning data - read with Word and transform into CSV.

Trying <table>. ConvertToText - not good, it preserves the return character and data can not be read safely after that.

The solution: read row by row and replace CR/LF with empty for each cell.

After this, problem with Bell character – replace this also the 7 character

After this, problem with repeating headers for each table in the csv

This is the VBA code

( maybe doing same in R : <https://datascienceplus.com/extracting-tables-from-pdfs-in-r-using-the-tabulizer-package/> )

***Step 3*** : Making objects to support

First creating objects to support this . Car, City, LicenseState, Licensee, TaxiAutorization.

Creating test - the City should be unique - even if Bucarest is multiple times, the City object should be the same.

***Step 4:***  Consolidating data from CSV to objects

Creating objects to mimic CSV data . Return to step 2 and put separator | instead of ,

Problem with parsing data

- a date could not exist, so it will be nullable

– could be either dd.mm.yyyy , either d.mm.yyyy either 27.02.202

Separate the lines with errors from the lines without errors . And return the result as a tuple

Making test in order to see the errors

***Step 5***: Application Web + WebAPI to display data

Making WEBAPI for seeing all and some taxis

See that valid taxis are not parsed correctly ( Validat vs Valid) . Modify test

Discover there are some licenses with state…

***Step 6***: Deploy sources to GitHub