Put OCRed Text into One Dataframe

In [1]:

```
import pandas as pd
import numpy as np
import os
from glob import glob
import re
```

In [3]:

```
# Use the text files to create dataframe
ar = []
x = ""
cum = []
result = [y for x in os.walk(r'catalog\new_text') for y in glob(os.path.join(x[0], '*.txt'))]
for res in result:
    with open(res, 'r',encoding="utf8") as source:
        lines = [line.replace("\n"," ") for line in source.readlines()]
        for l in lines:
            # if line is not end of line, add the text
            if not (l.startswith('\x0c')):
                x+=l
            # when reaching end of line, append a new entry to list which contains all the text before end of l
            else:
                ar = np.append(ar,x)
                # Starting after "\x0c", add text
                x =l[1:]
        cum.append(len(ar))
df = pd.DataFrame(ar)
df.to_csv("catalog\main_file_all_text.csv")
```

Out[3]:

```
(50368, 1)
```