Data Cleaning

> https://colab.research.google.com/drive/1Y5X2A70huga6D-_VgBTI3y-TSiA0M5bS

In [0]:

```
!wget https://www.dropbox.com/s/toswrzisqljgkb3/Agile.zip?dl=0 #Dataset saved in dropbox
```

Loding the file

In [0]:

```
from zipfile import ZipFile

with ZipFile('/content/Agile.zip?dl=0', 'r') as zipObj:
   zipObj.extractall()
```

In [0]:

```
import pandas as pd
import codecs

doc = codecs.open('/content/agile_1.txt','rU','UTF-16')
wos_df = pd.read_csv(doc, sep='\t')
```

In [0]:

```
for i in range(1,4):
      doc = codecs.open('/content/agile_'+str(i)+'.txt','rU','UTF-16')
      df = pd.read_csv(doc, sep='\t')
      wos_df=pd.concat([df,wos_df], sort=False)
```

Choosing the features

In [0]:

```
print(wos_df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1516 entries, J to J
Data columns (total 67 columns):
PT    1516 non-null object
AU    0 non-null float64
BA    0 non-null float64
BE    0 non-null float64
GP    1516 non-null object
AF    0 non-null float64
BF    0 non-null float64
CA    1516 non-null object
TI    1516 non-null object
SO    0 non-null float64
SE    0 non-null float64
BS    1516 non-null object
LA    1516 non-null object
DT    32 non-null object
CT    32 non-null object
CY    32 non-null object
CL    16 non-null object
SP    13 non-null object
HO    1350 non-null object
DE    1248 non-null object
ID    1496 non-null object
AB    1497 non-null object
C1    1502 non-null object
```

```
CI     1502 non-null object
RP     1327 non-null object
EM     440 non-null object
RI     545 non-null object
OI     251 non-null object
FU     245 non-null object
FX     1478 non-null object
CR     1516 non-null int64
NR     1516 non-null int64
TC     1516 non-null int64
Z9     1516 non-null int64
U1     1516 non-null int64
U2     1516 non-null object
PU     1516 non-null object
PI     1516 non-null object
PA     1516 non-null object
SN     1143 non-null object
EI     0 non-null float64
BN     1516 non-null object
J9     1515 non-null object
JI     877 non-null object
PD     1504 non-null float64
PY     1490 non-null object
VL     1428 non-null object
IS     2 non-null object
PN     8 non-null float64
SU     172 non-null object
SI     0 non-null float64
MA     1466 non-null float64
BP     1466 non-null object
EP     51 non-null object
AR     1374 non-null object
DI     0 non-null float64
D2     18 non-null object
EA     1516 non-null int64
PG     1516 non-null object
WC     1516 non-null object
SC     1516 non-null object
GA     1516 non-null object
UT     8 non-null float64
PM     233 non-null object
OA     18 non-null object
HC     18 non-null object
HP     1516 non-null object
DA     0 non-null float64
dtypes: float64(15), int64(6), object(46)
memory usage: 805.4+ KB
None
```

In [0]:

```
wos_df=wos_df[['GP','CA','TI','HO','DE','ID','FX','PU','PD','PG','AR']]
wos_df.columns=['Author','Title','Journal','Sub-field','Keywords','Abstract','References
','City of pub','Year','Field','DOI']
```

Dropping articles missing values in the most important features, dropping potential dublicates and resetting the index.

In [0]:

```
wos_df.dropna(subset=['Author','Abstract','References','DOI'],inplace=True)
```

In [0]:

```
wos_df=wos_df.drop_duplicates()
```

In [0]:

```
wos_df.reset_index(drop=True,inplace=True)
```

In [0]:

```
wos_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 901 entries, 0 to 1015
Data columns (total 11 columns):
Author          901 non-null object
Title           901 non-null object
Journal         901 non-null object
Sub-field       834 non-null object
Keywords        778 non-null object
Abstract        901 non-null object
References      901 non-null object
City of pub     901 non-null object
Year            895 non-null float64
Field           901 non-null object
DOI             901 non-null object
dtypes: float64(1), object(10)
memory usage: 84.5+ KB
```

Exporting the file so we can use it in the two notebooks

In [0]:

```
from google.colab import files

wos_df.to_csv('wos_df.csv',index = True , header=True,sep='\t')
files.download('wos_df.csv')
```