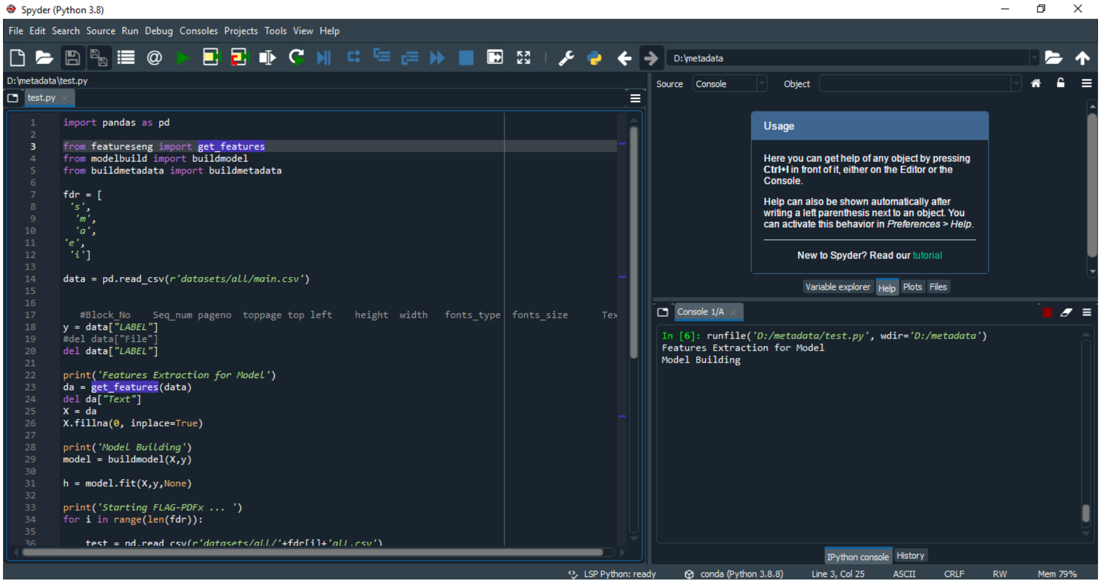
Help File to Execute the code and verify extracted Metadata

1. The folders “\metadata\datasets” has the code and extracted PDF content in form of CSV files
2. Please install python 3.7 version
3. In \metadata folder is test.py file. This is the main file which run the complete code



```
import pandas as pd

from featureseng import get_features
from modelbuild import buildmodel
from buildmetadata import buildmetadata

fdr = [
 's',
  'm',
  'a',
'e',
 'i']

data = pd.read_csv(r'datasets/all/main.csv')


   #Block_No    Seq_num pageno  toppage top left    height  width   fonts_type  fonts_size      Tex
y = data["LABEL"]
#del data["File"]
del data["LABEL"]

print('Features Extraction for Model')
da = get_features(data)
del da["Text"]
X = da
X.fillna(0, inplace=True)

print('Model Building')
model = buildmodel(X,y)

h = model.fit(X,y,None)

print('Starting FLAG-PDFx ... ')
for i in range(len(fdr)):

    test = pd.read_csv(r'datasets/all/'+fdr[i]+'all.csv')
```

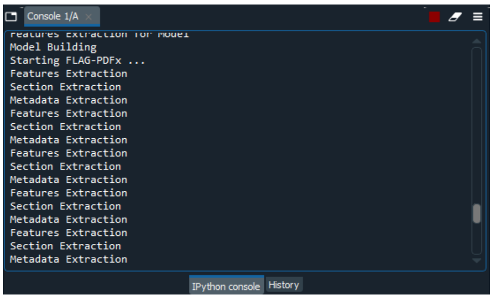
Here you can get help of any object by pressing **Ctrl+I** in front of it, either on the Editor or the Console.

Help can also be shown automatically after writing a left parenthesis next to an object. You can activate this behavior in *Preferences > Help*.

New to Spyder? Read our tutorial

```
In [6]: runfile('D:/metadata/test.py', wdir='D:/metadata')
Features Extraction for Model
Model Building
```

4. When test.py is executed then console file shows the process stages. The Model Building stage is based on system specifications.
5. When complete process is completed then the system generates the output file in form of metadata CSV files.

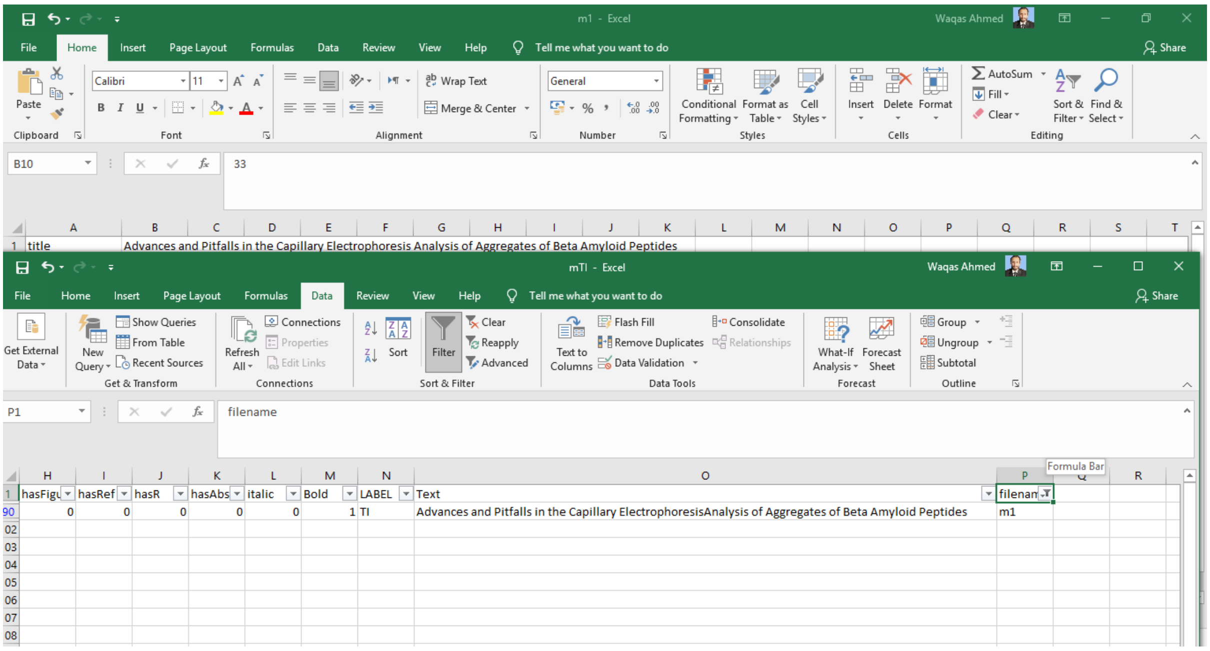


```
Features Extraction for Model
Model Building
Starting FLAG-PDFx ...
Features Extraction
Section Extraction
Metadata Extraction
Features Extraction
Section Extraction
Metadata Extraction
Features Extraction
Section Extraction
Metadata Extraction
Features Extraction
Section Extraction
Metadata Extraction
Features Extraction
Section Extraction
Metadata Extraction
```

6. The main.csv is used to build the model
7. The metadata\datasets\all has csv with file names like ialL (i, e, m, s, a represents different publishers). These files contain the extracted pdf content with geometric and fonts info.
8. The output of system is in metadata\datasets\ folder (with initial class label and section postfix).
   a. TI - Title
   b. AU- Authors
   c. AFF- Affiliations
   d. EM- Emails
   e. HL- Header/ footers
   f. H1- Heading level 1
   g. H2- Heading level 2
   h. H3- Heading level 3

i. TA- Table
j. FI- Figures
k. ACK- Acknowledgements
l. RE- References

9. Dataset Folders ds\Annotated\ has subfolders publisher wise.
10. Each csv file has the metadata of same file pdf
11. To verify output



m1 - Excel

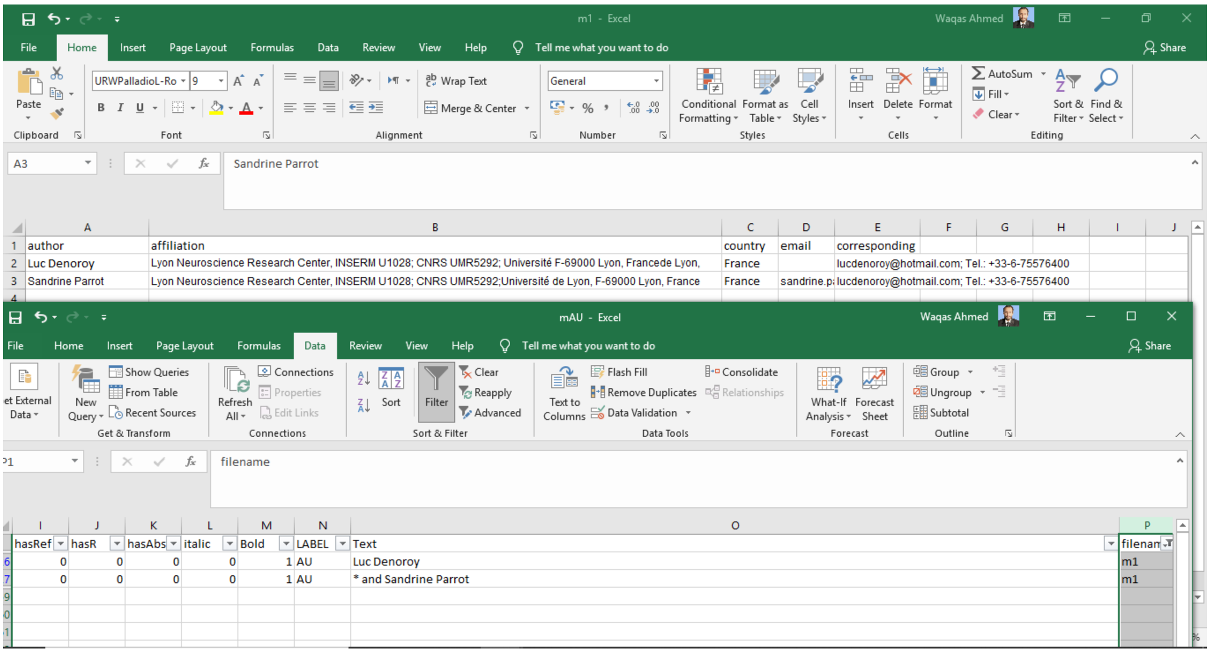
B10: 33

	A	B
1	title	Advances and Pitfalls in the Capillary Electrophoresis Analysis of Aggregates of Beta Amyloid Peptides

mTI - Excel

P1: filename

	H	I	J	K	L	M	N	O	P
1	hasFigu	hasRef	hasR	hasAbs	italic	Bold	LABEL	Text	filenam
90	0	0	0	0	0	1	TI	Advances and Pitfalls in the Capillary ElectrophoresisAnalysis of Aggregates of Beta Amyloid Peptides	m1
02									
03									
04									
05									
06									
07									
08									



m1 - Excel

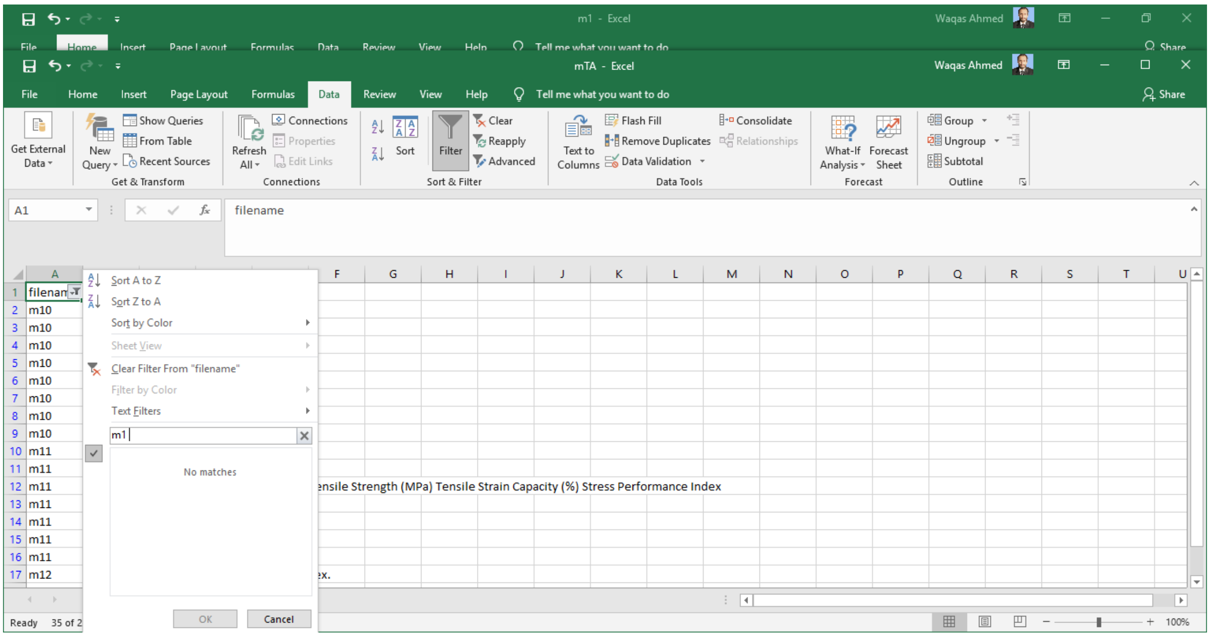
A3: Sandrine Parrot

	A	B	C	D	E
1	author	affiliation	country	email	corresponding
2	Luc Denoroy	Lyon Neuroscience Research Center, INSERM U1028; CNRS UMR5292; Université F-69000 Lyon, Francede Lyon,	France		lucdenoroy@hotmail.com; Tel.: +33-6-75576400
3	Sandrine Parrot	Lyon Neuroscience Research Center, INSERM U1028; CNRS UMR5292;Université de Lyon, F-69000 Lyon, France	France	sandrine.p	lucdenoroy@hotmail.com; Tel.: +33-6-75576400

mAU - Excel

P1: filename

	I	J	K	L	M	N	O	P
1	hasRef	hasR	hasAbs	italic	Bold	LABEL	Text	filenam
6	0	0	0	0	1	AU	Luc Denoroy	m1
7	0	0	0	0	1	AU	* and Sandrine Parrot	m1



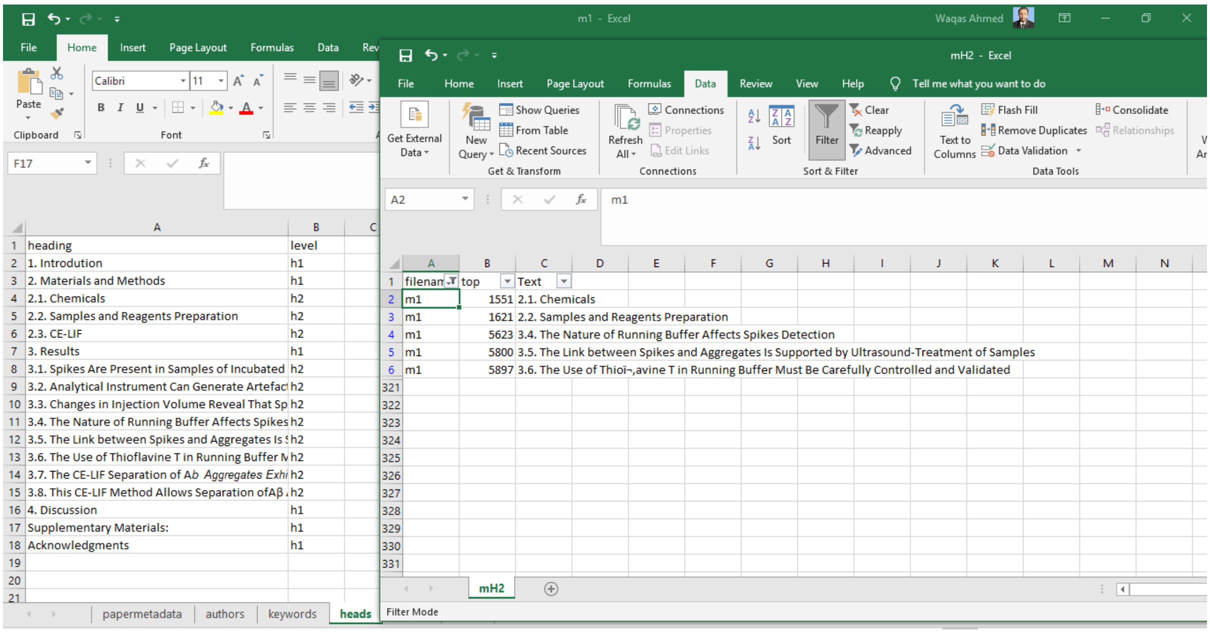
A1 | filename

Search: m1

No matches

	A	
1	filename	
2	m10	
3	m10	
4	m10	
5	m10	
6	m10	
7	m10	
8	m10	
9	m10	
10	m11	
11	m11	
12	m11	ensile Strength (MPa) Tensile Strain Capacity (%) Stress Performance Index
13	m11	
14	m11	
15	m11	
16	m11	
17	m12	ex.

Ready 35 of 2



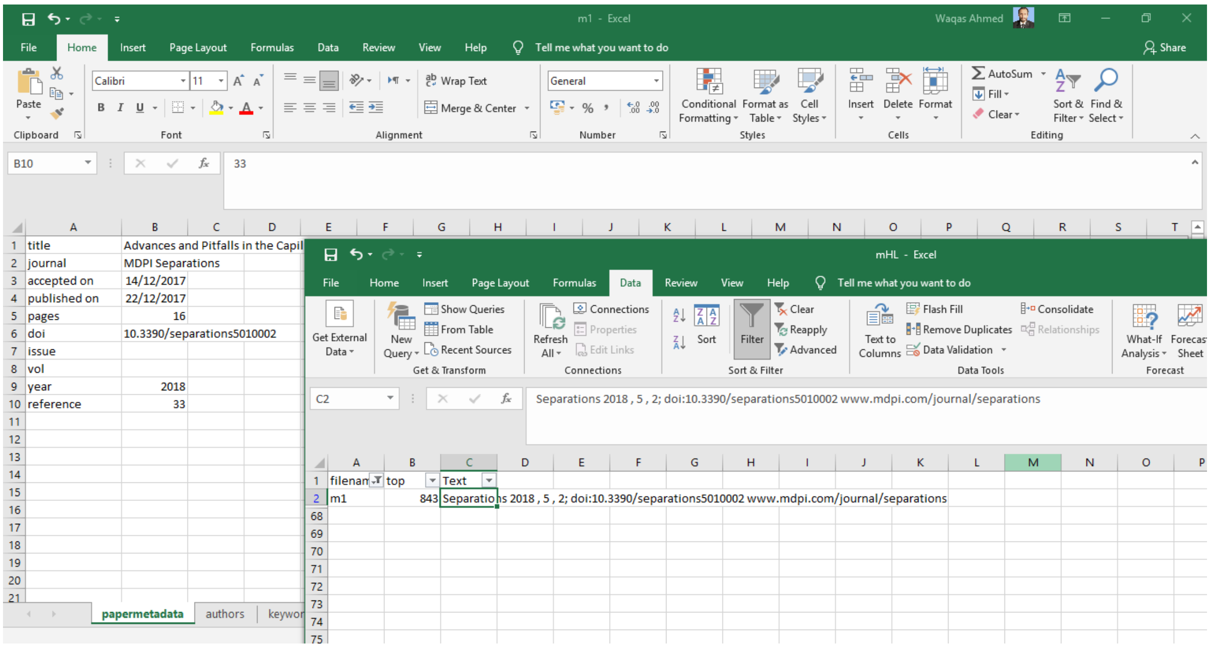
F17

	A	B
1	heading	level
2	1. Introdution	h1
3	2. Materials and Methods	h1
4	2.1. Chemicals	h2
5	2.2. Samples and Reagents Preparation	h2
6	2.3. CE-LIF	h2
7	3. Results	h1
8	3.1. Spikes Are Present in Samples of Incubated	h2
9	3.2. Analytical Instrument Can Generate Artefact	h2
10	3.3. Changes in Injection Volume Reveal That Sp	h2
11	3.4. The Nature of Running Buffer Affects Spikes	h2
12	3.5. The Link between Spikes and Aggregates Is S	h2
13	3.6. The Use of Thioflavine T in Running Buffer M	h2
14	3.7. The CE-LIF Separation of A*b* *Aggregates Exhi*	h2
15	3.8. This CE-LIF Method Allows Separation ofAβ A	h2
16	4. Discussion	h1
17	Supplementary Materials:	h1
18	Acknowledgments	h1
19		
20		
21		

A2 | m1

	A	B	C
1	filename	top	Text
2	m1	1551	2.1. Chemicals
3	m1	1621	2.2. Samples and Reagents Preparation
4	m1	5623	3.4. The Nature of Running Buffer Affects Spikes Detection
5	m1	5800	3.5. The Link between Spikes and Aggregates Is Supported by Ultrasound-Treatment of Samples
6	m1	5897	3.6. The Use of Thioï¬‚avine T in Running Buffer Must Be Carefully Controlled and Validated
321			
322			
323			
324			
325			
326			
327			
328			
329			
330			
331			

Filter Mode



m1 - Excel

B10 33

	A	B	C	D
1	title	Advances and Pitfalls in the Capil		
2	journal	MDPI Separations		
3	accepted on	14/12/2017		
4	published on	22/12/2017		
5	pages	16		
6	doi	10.3390/separations5010002		
7	issue			
8	vol			
9	year	2018		
10	reference	33		

mHL - Excel

C2 Separations 2018 , 5 , 2; doi:10.3390/separations5010002 www.mdpi.com/journal/separations

	A	B	C
1	filenam	top	Text
2	m1	843	Separations 2018 , 5 , 2; doi:10.3390/separations5010002 www.mdpi.com/journal/separations