

Διαδικασία Συλλογής των δεδομένων.

Η διαδικασία με την οποία συλλέξαμε τα δεδομένα έγινε με την χρήση python και της βιβλιοθήκης BeautifulSoup λόγο καλύτερης εξοικείωσης με το συγκεκριμένο εργαλείο.








- Αρχικά έγινε crawl των link όλων των προηγούμενων εκδόσεων των πολιτικών απορρήτου με το παρακάτω script.

```
1 import pandas as pd
2 import bs4 as bs
3 import urllib.request
4 import re
5 import sys
6 import json
7
8
9 def rm_main():
10
11     #CONCATING URL AND ARGUMENT
12     args = ["el", "da", "de", "en", "es", "fr", "it", "nl", "pt", "tr"]
13     links = []
14     for a in args:
15         url = "https://policies.google.com/privacy/archive?hl="
16         url2 = url+a
17
18         #BEAUTIFUL SOUP STUFF
19         source = urllib.request.urlopen(url2).read()
20         soup = bs.BeautifulSoup(source, 'lxml')
21
22         #GETTING CONTENT OF THE UL
23         link_list = soup.find('ul', {"id": "archives"})
24
25         #LINKS
26         temp = 0;
27         for li in link_list.findAll('li'):
28             if (temp >= 2):
29                 try:
30                     links.append(re.search('<a href="https://(.+?)">', str(li)).group(1))
31                 except AttributeError:
32                     links = 'fail'
33             if (temp < 2):
34                 temp +=1
35
36         #DATATABLE WITH INFO
37         df = pd.DataFrame({'Links':links})
38
39     return df
```

- Στη συνέχεια έγινε classification, φτιάχνοντας ένα attribute class με το operator generate attributes, ξεχωρίζοντας τα Links ανάλογα με τα τελευταία δύο γράμματα του url όπου είναι και η ανάλογη κωδικοποίηση των γλωσσών.

Expression

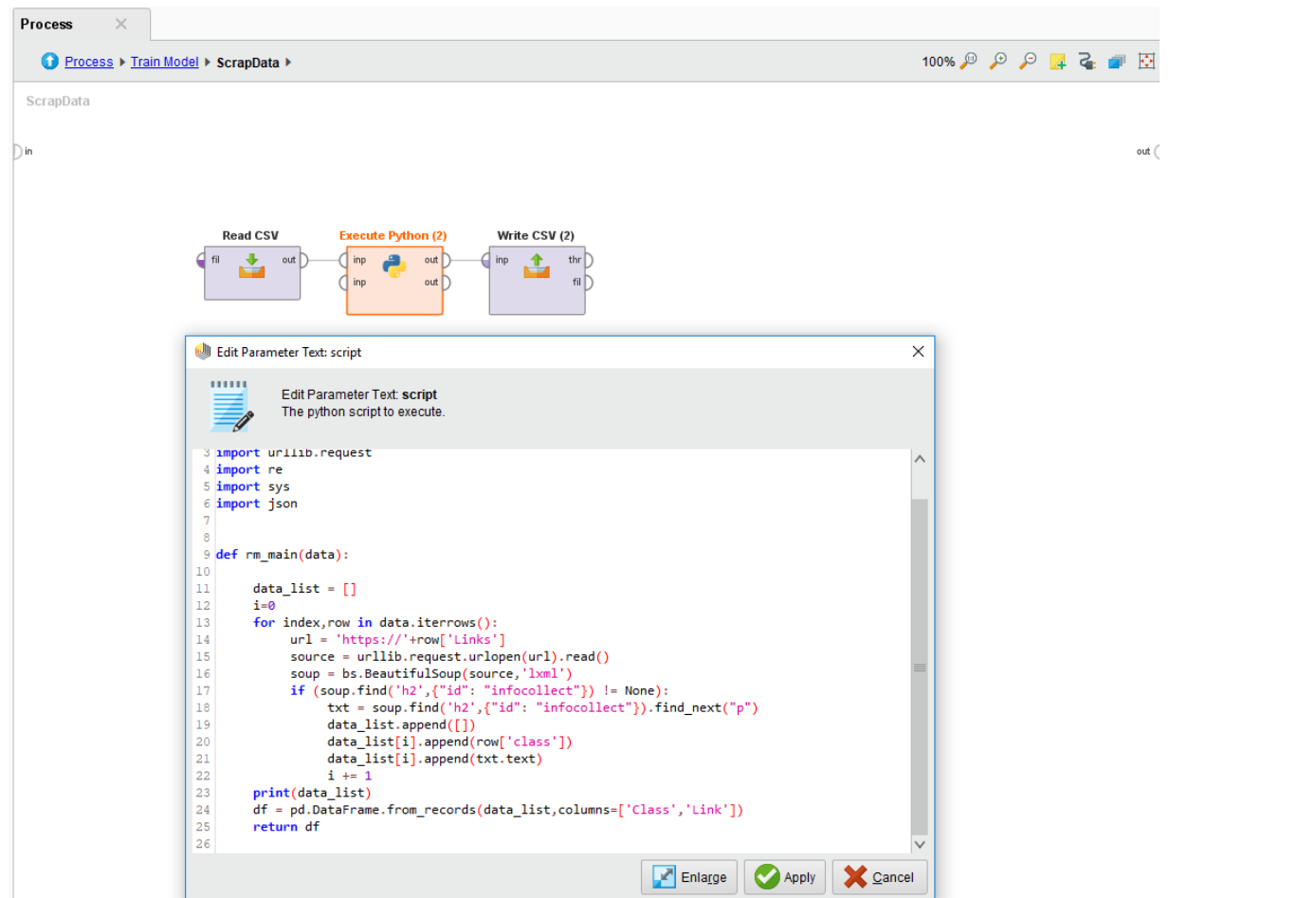
```
1 if(contains(Links,"el"),"Greek",if(contains(Links,"da"),
2 "Danish",if(contains(Links,"de"),
3 "German",if(contains(Links,"en"),
4 "English",if(contains(Links,"tr"),
5 "Turkish",if(contains(Links,"fr"),
6 "French",if(contains(Links,"it"),"Italian",if(contains(Links,"nl"),"Nederlands",if(contains(Links,"pt"),"Portuguese",if(contains
```

▶ CrawlData GenerateLabel ▶ 100%       

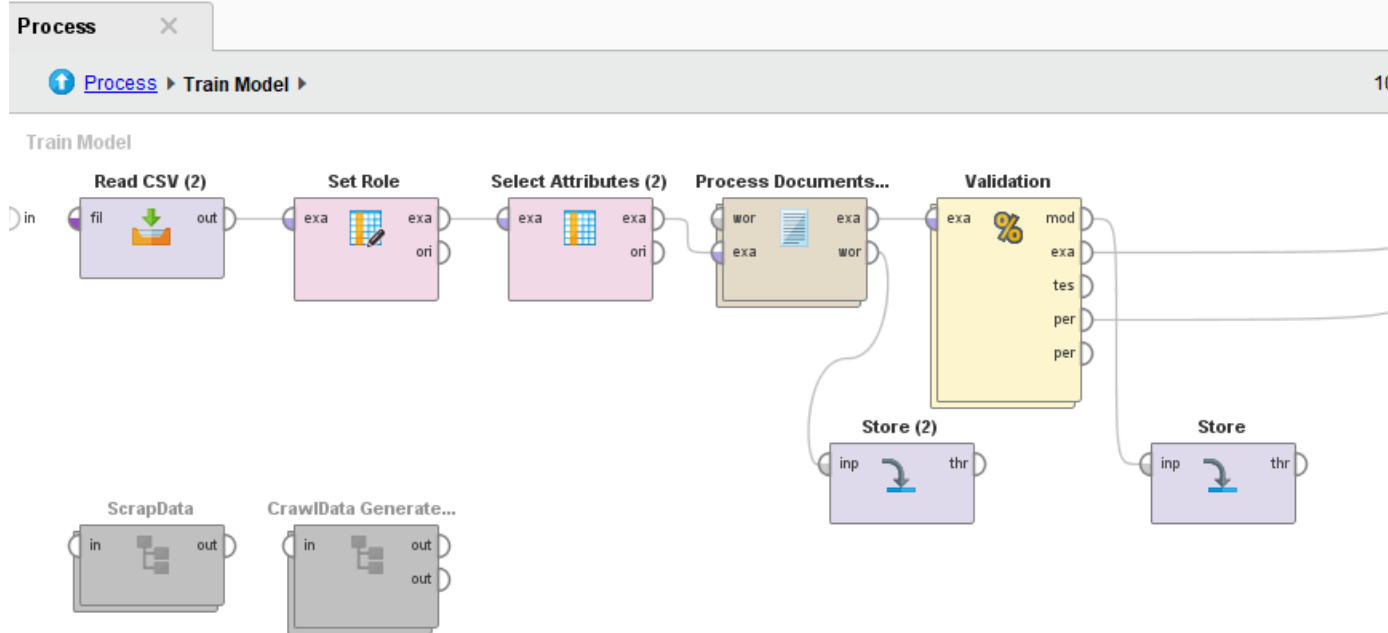
100%



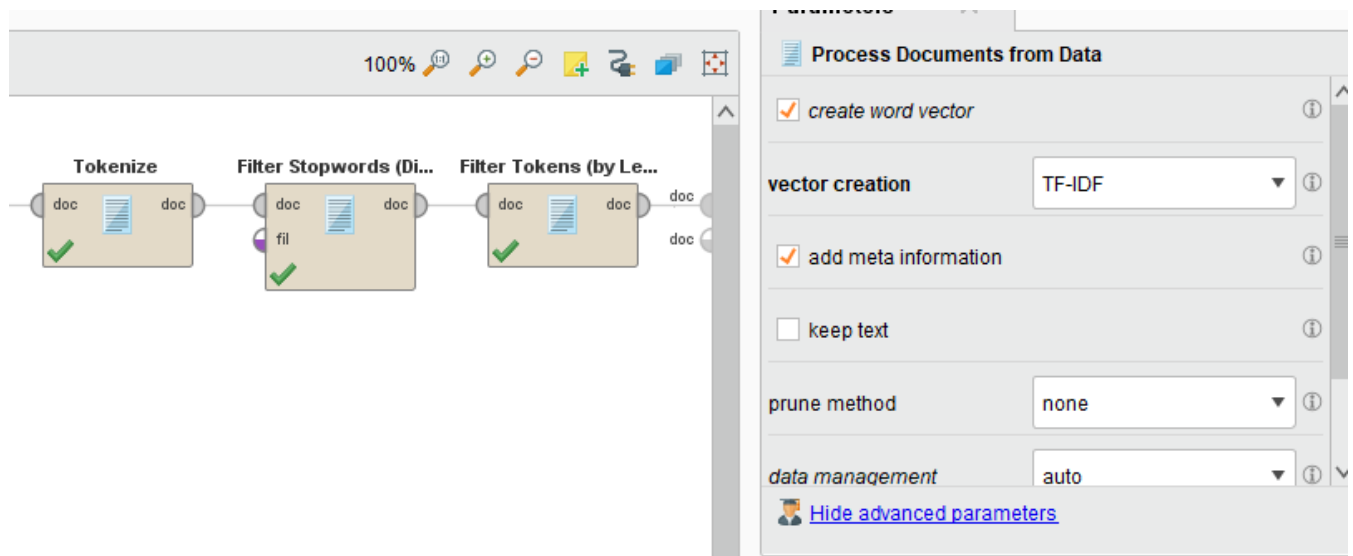
- Τέλος, αφού τα αποτελέσματα σώθηκαν σε csv αρχείο έγινε scrap της ζητούμενης παραγράφου σε κάθε ένα από τα link με το παρακάτω script.



Δημιουργία Μοντέλου.



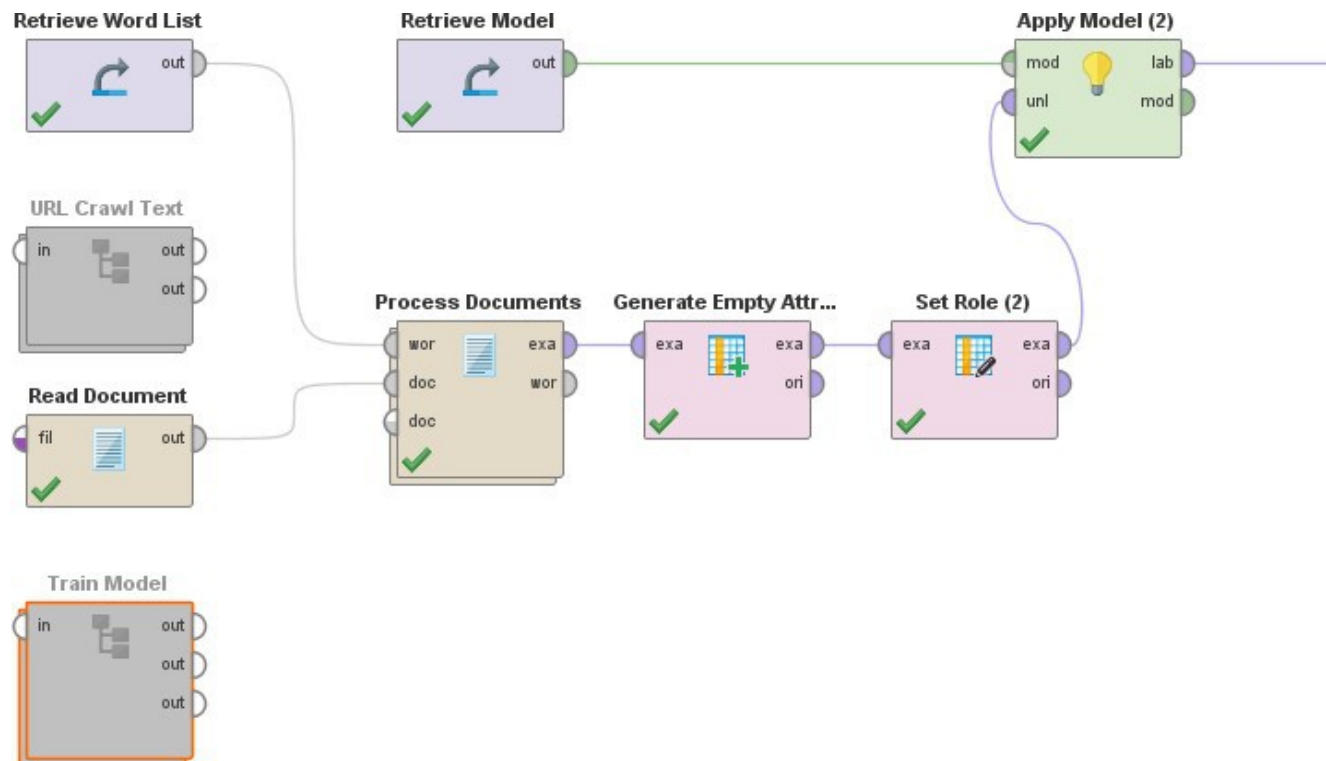
- Για την δημιουργία του μοντέλου αρχικά θέτουμε ως label το attribute class που δημιουργήσαμε πιο πριν και μετά γίνεται ένα select attributes για την αφαίρεση ενός ανεπιθύμητου attribute(η γραμμές του panda dataframe που φτιάξαμε πιο πριν στην python).
- Στη συνέχεια επεξεργαζόμαστε τα κείμενα κατάλληλα.



Συγκεκριμένα γίνεται χωρισμός των λέξεων σε ξεχωριστά tokens, φιλτράρονται ανούσιες λέξεις σύμφωνα με λεξικό (στη συγκεκριμένη περίπτωση δεν θα χρειαζόταν εφόσον και απ' τις ανούσιες λέξεις μπορεί να αναγνωριστεί μια γλώσσα αλλά μπήκε για να μικρύνει το example set) , κόψιμο πολύ μεγάλων και πολύ μικρών λέξεων. Τέλος ,οι λέξεις αντιπροσωπεύονται με το σχήμα TF-IDF όπου παίρνουν τιμές ανάλογα με το πόσο συχνά εμφανίζονται στο κείμενο.

- Σώσιμο του word list για αναφορά του αργότερα και δημιουργία του μοντέλου με Cross Validation και την βοήθεια του αλγορίθμου Naïve Bayes όπου μετά από δοκιμές καταλήξαμε ότι είναι ο πιο κατάλληλος για text categorization.

Αναγνώριση Γλώσσας σε κείμενο και URL.



- Αφού διαβαστεί το κείμενο (ή το περιεχόμενο ενός url σε μορφή κειμένου) περνάει απ' την ίδια διαδικασία που πέρασε και το training set παίρνοντας την ίδια μορφή.
- Παράδειγμα αναγνώρισης ενός Δανικού κειμένου :

ExampleSet (Apply Model (2))

ExampleSet (1 example, 13 special attributes, 256 regular attributes) Filter (1)

Row No.	Class	prediction(Class...	confidence(Greek)	confidence(Danish)	confidence(German)	confidence(English)
1	?	Danish	0	1	0	0

D:\RapidMinerProjects\Project\text.txt - Notepad++

File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ? X

charts_table.php bubble_table.php countryGeo.php kbubble.php kmeans.php text.f

```

1 Alle mennesker
2 er født frie og lige i værdighed
3 og rettigheder. De er udstyret med fornuft og samvittighed,
4 og de bør handle mod hverandre i en broderskabets ånd.

```

length: 172 lin: Ln: 4 Col: 1 Sel: 0 | 0 Windows (CR LF) UTF-8 INS

Παράδειγμα αναγνώρισης γλώσσας από το url : <http://www.icsd.aegean.gr/icsd/>

Process

Process > URL Crawl Text > 100%

Crawl Text

Get Page out doc doc doc doc doc doc

Extract Content (2) Unescape HTML Doc...

Parameters

Get Page

url <http://www.icsd.aegean.gr/icsd/>

☐ random user agent

user agent

connection timeout 10000

read timeout 10000

[Hide advanced parameters](#)

ExampleSet (Apply Model (2))

ExampleSet (1 example, 13 special attributes, 256 regular attributes)

Row No.	Class	prediction(C...	confidence(Greek)	confidence(Danish)	confidence(...	co
1	?	Greek	1	0	0	0