Γλωσσική Τεχνολογία

Εργασία 2020-2021

έΥΑ ντουρου (παλαιοσ αμ 235854, νεοσ αμ 1041567)

2021

Περιεχόμενα

[Μέρος Α 2](#_Toc81479601)

[Προσκομιστής Ιστοσελίδων 2](#_Toc81479602)

[Προεπεξεργασία Δεδομένων 2](#_Toc81479603)

[**Politico** 2](#_Toc81479604)

[**Euronews** 3](#_Toc81479605)

[Αναφορές 4](#_Toc81479606)

# Μέρος Α

## Προσκομιστής Ιστοσελίδων

Ο προσκομιστής ιστοσελίδων υλοποιήθηκε με τη χρήση του Scrapy (1), το οποίο λαμβάνει ως πηγές τα RSS feeds που φαίνονται στον πίνακα 1 και εξάγει το περιεχόμενό τους σε μορφοποίηση html, το οποίο και αποθηκεύει στο directory του project. Η υλοποίησή του φαίνεται στο αρχείο parker.py. (./scraper\_tintin/Scraper/NewsScrape/NewsScrape/spiders/parker.py)

Πίνακας 1 Πηγές του crawler

|  |  |
| --- | --- |
| **Όνομα πηγής** | **RSS feed Link** |
| Eureporter | https://www.eureporter.co/feed/ |
| Aljazeera | https://www.aljazeera.com/xml/rss/all.xml |
| CNN | http://rss.cnn.com/rss/edition\_world.rss |
| Euronews | http://feeds.feedburner.com/euronews/en/home/ |
| BBC | http://feeds.bbci.co.uk/news/world/rss.xml |
| Washington Post | http://feeds.washingtonpost.com/rss/world |
| Politico | https://www.politico.eu/feed/ |
| NY Times | https://www.nytimes.com/svc/collections/v1/publish/https://www.nytimes.com/section/world/rss.xml |

Από τις παραπάνω ιστοσελίδες, οι NY times δεν επέτρεπαν πρόσβαση σε scraper μέσω του robots.txt.

Οι σελίδες αυτές μετά το scrape, αποθηκεύονται σε αρχεία html, για την προεπεξεργασία.

## Προεπεξεργασία Δεδομένων

Για την προεπεξεργασία δεδομένων παίρνουμε ξεχωριστά την κάθε αποθηκευμένη σελίδα και με χρήση του εργαλείου Beautiful Soup, εξάγουμε από αυτή τους τίτλους. Συγκεκριμένα:

### **Politico**

Κάθε άρθρο περιλαμβάνεται υπό ένα <item> tag.

Οι τίτλοι και τα links των άρθρων βρίσκονται στα αντίστοιχα tags.

To RSS του politico περιλαμβάνει και ολόκληρα τα άρθρα, στο tag <content:encoded>, το οποίο το Beautiful Soup μετατρέπει απλά σε <encoded>. Το εξάγουμε μέσω της findAll() και στη συνέχεια το μετατρέπουμε σε καλύτερη μορφή, αφαιρώντας τα tags :

article = re.sub('<.\*?>', '', article)

και τα newlines:

article = ''.join(article.strip().split('\\n'))

Αξίζει να σημειωθεί πως ανάλογη επεξεργασία γίνεται και στα links, καθώς περιλαμβάνουν ειδικούς χαρακτήρες \r, \n οι οποίοι δυσκολεύουν την χρήση τους.

### **Euronews**

Αντίστοιχα με το Politico, οι τίτλοι και τα links εξάγονται από τα αντίστοιχα tags.

Το Euronews δεν περιλαμβάνει ολόκληρα άρθρα στο RSS feed του, επομένως χρησιμοποιούμε το κάθε link για να κάνουμε εκ νέου request και στη συνέχεια να εξάγουμε το άρθρο από τη σελίδα του μέσω του Beautiful Soup.

Συγκεκριμένα εξάγουμε όλα τα div του class 'c-article-content' και στη συνέχεια από αυτά όλα τα p tags, δηλαδή τις παραγράφους του άρθρου. Έπειτα της συνδέουμε όλες για να αποθηκεύσουμε ολόκληρο το άρθρο.

### **Washington Post**

To RSS feed γίνεται parse ακριβώς όπως στο Euronews.

Στη συνέχεια, πρέπει να πάρουμε το άρθρο από το link του, με εκ νέου scrape του.

### **BBC**

### **Aljazeera**

### **CNN**

### **Eureporter**

# Αναφορές

1. **Scrapy.** Scrapy. [Ηλεκτρονικό] https://scrapy.org/.

2. **Crummy.** Beautiful Soup 4. [Ηλεκτρονικό] https://www.crummy.com/software/BeautifulSoup/.

3. **NLTK.** Natural Language Toolkit. [Ηλεκτρονικό] https://www.nltk.org/.