Γλωσσική Τεχνολογία

Εργασία 2020-2021

έΥΑ ντουρου (παλαιοσ αμ 235854, νεοσ αμ 1041567)

2021

Περιεχόμενα

[Μέρος Α 2](#_Toc81307285)

[Προσκομιστής Ιστοσελίδων 2](#_Toc81307286)

[Αναφορές 2](#_Toc81307287)

# 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. (./Tintin/Scraper/NewsScrape/NewsScrape/spiders/parker.py)

Πίνακας 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 Aljazeera, CNN, NY times και

## Προεπεξεργασία Δεδομένων

# 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/.