ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Crawling Greek business websites

Master in Data Science

Panagiotis Kapsalis

October 2017, Athens

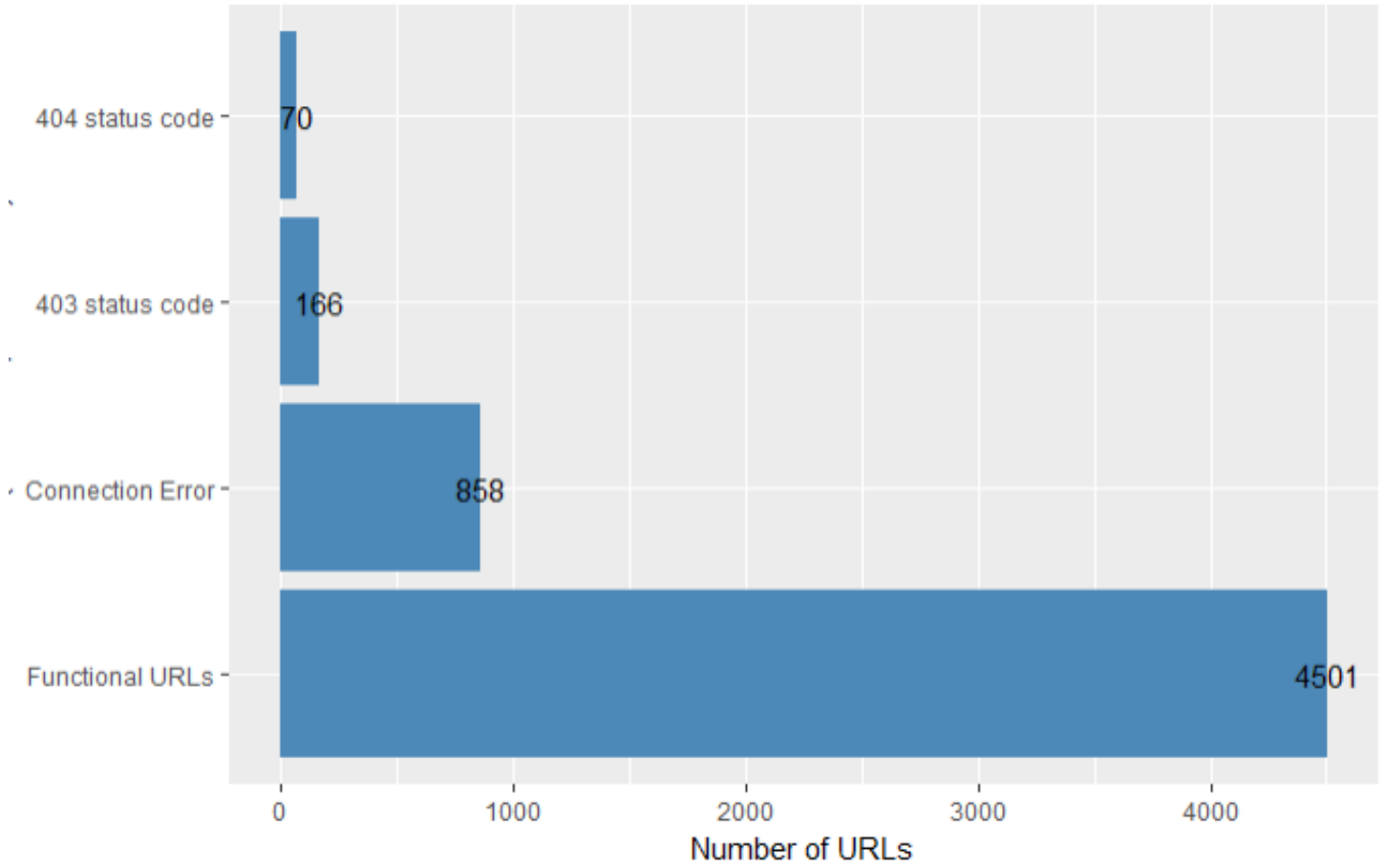
Aim of the thesis

- The goal of this thesis is to extract business data from their websites.
- Crawling and text engineering methods will be used, in order to investigate how feasible is to extract information about businesses from their websites.
- For this reason, we have implemented scrapers, which take as input Greek business websites and extract data from their content.
- We need to automatically identify common patterns on websites.

Greek business websites

- In order to implement the crawlers and to validate their results, we were given 5500 domain names.
- We have implemented a Python script which finds the full URL scheme for each domain name and identifies the URLs presented connection error, 404 status code and the URLs which raised 403 status code.
- 403 status code is raised because of server security features which block spider/ bot user agents
- In order to avoid this problem we use a known browser user agent.

URLs


404 status code
70
403 status code
166
Connection Error
858
Functional URLs
4501
0
1000
2000
3000
4000
Number of URLs

First Crawler

The **First Crawler** takes as input, a business website and extracts the following elements:

- Social network names (text)
- Social network links (text)
- If the website provides multi language option (0 or 1)
- If the website provides newsletter (0 or 1)
- If the website provides Search option (0 or 1)
- If the website provides Blog (0 or 1)
- If the website provides mobile application (0 or 1)
- If the website provides E-shop (0 or 1)

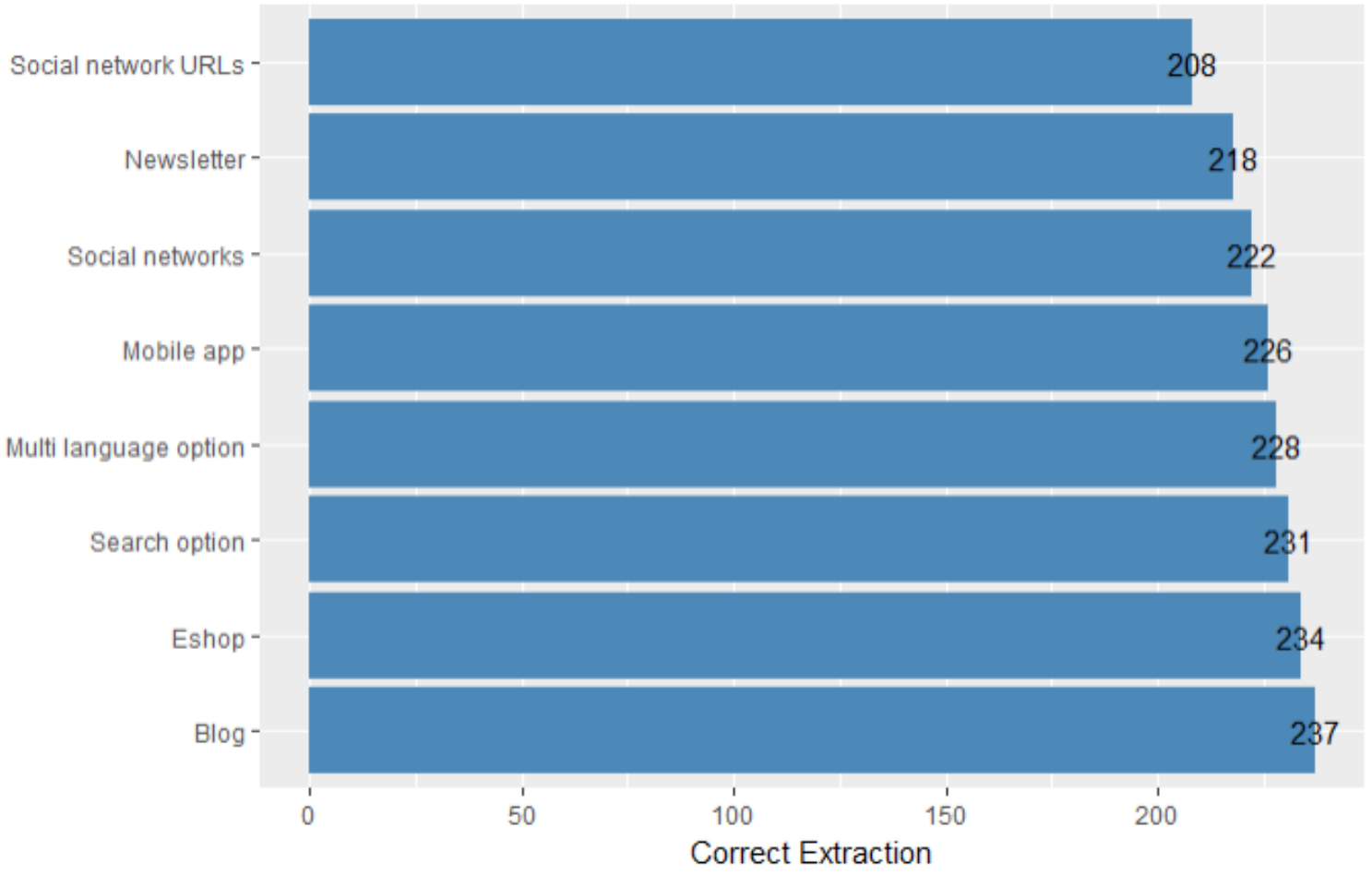
First Crawler demo

```
In [8]: first_crawler('http://couniniotis.gr/')

Out[8]: ['http://couniniotis.gr/',
         ['Facebook', 'Twitter', 'Linkedin', 'Youtube'],
         ['http://www.youtube.com/user/CouniniotisGroup',
          'http://www.linkedin.com/company/couniniotis-group',
          'https://www.youtube.com/user/CouniniotisGroup',
          'http://www.facebook.com/couniniotis.gr'],
         1,
         0,
         1,
         1,
         0,
         0]
```

First Crawler: Evaluation

- We use the First Crawler for the URLs we have in our disposal (4501 + 166)
- The execution time needed is ~**1 h 40 m**
- In order to evaluate First Crawler results, we take 250 random samples and we compare them with websites' information

First Crawler: Evaluation



Second Crawler

The **Second Crawler** extract from business websites the following elements:

- Email contacts
- Phone contacts
- Business name
- Quality certifications
- Countries in which the business operates
- Business scope of activities

In order to make more efficient the extraction of the above elements we distribute the work load in 5 functions.

Second Crawler: “emails”

·The function “email_crawler”, extracts businesses email contacts.

·This function use regular expressions in order to identify email addresses from the website

·Selenium Web driver is used in order to disable JavaScript from websites which use it to protect their emails from bots.

·Selenium Web driver is a very slow module, every time is activated a new browser window presents

Second Crawler: “phones”

- The function “finder” extract business phone numbers from their websites.
- In order to extract phone numbers we use regular expressions.
- We found all the possible ways where a phone number can be written, for example:
  - 2106845691
  - (210) 684 56910
  - +30 2106845691
  - 210.684.56910
  - 26510 57 580

Second Crawler: “Quality certifications”

- The function “find_certf” extract quality certifications, such as “**ISO**”, “**HACCP**”, “**AGROCERT**”, “**OHSAS**”, “**TUV**” from business website.
- In order to find certifications the function searches on internal links relating to businesses’ history, businesses’ quality, businesses’ identity.

Second Crawler: “Business name”

·The function “find_bname”, extracts Business name from its website.

· More specifically, it tries to locate capitalized words before keywords such as “**A.E**”, “**Ε.Π.Ε**”, “**A.T.E**”, “**O.E**”, “**A.B.E.E**”, “**E.E**”, “**S.A**”, “**Ltd**” etch.

· In order to achieve this we use regular expressions to identify the business names before these keywords.

· In case these keywords can’t be found, the function tries to locate capitalized words (business name) after keywords such as “company”, “εταιρεία”, “οργανισμός”, “επιχείρηση”, “’ομιλος” etch.

Second Crawler: “Country data”

·The function “country_data”, tries to locate country names in which a company operates and its scope of activities.

· More specifically the function checks for country names on websites’ internal links relating to company profile or companies exports or companies imports and activities.

·Business scope of activities is “local” if the function does not locate the name of any country name, is “national” if “Greece” is located and “international” if more than one country names located.

·The function returns country names and (0 for local scope of activities, 1 for national scope of activities and 2 for international scope of activities).

Second Crawler demo

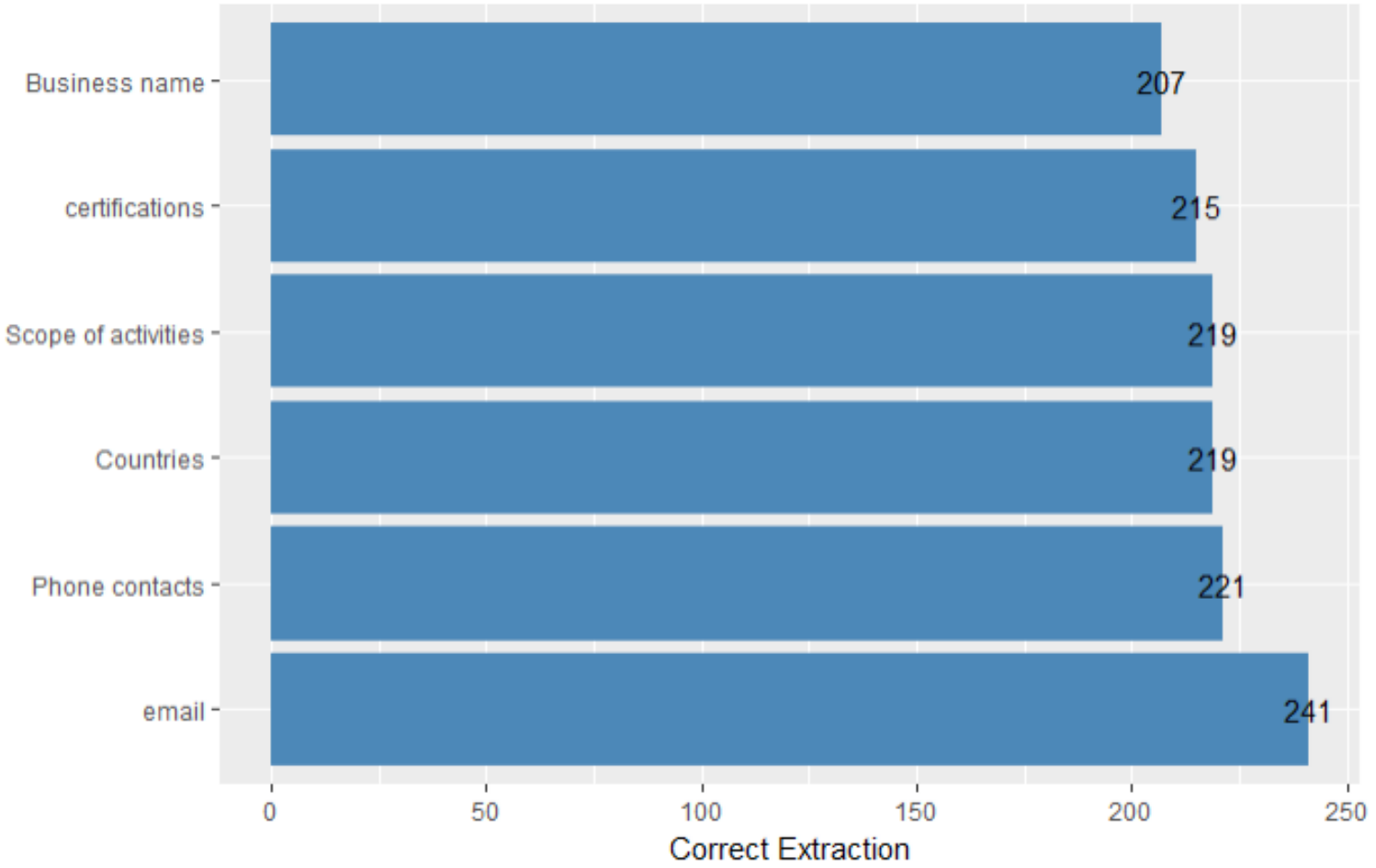
```
In [95]: run_me('http://hellasfrost.gr/')

         Emails:
         ['info@hellasfrost.gr']

         Quality certifications:
         ['AGRO 2-1', 'AGRO 2-2', 'ISO 22000:2005', 'ISO ', 'EUROCERT']

         Phones:
         2553051721,2553051720

         Business name:
         Ελληνική Οικοαγροτική Α.Ε

         Countries and scope of activities:
         (['Ελλάδα'], 1)
```

Second Crawler: Evaluation

The table bellow demonstrates the execution time needed by the functions to extract these elements from 4600 business websites.

Second Crawler	Phone Contacts	~ 3653 sec	~ 1 h
	Countries in which a company operates	~ 5769 sec	~ 1h 35 m
	Scope of activities		
	Business name	~ 11929 sec	~ 3h 20 m
	Email	~ 23907 sec	~ 6 h 30 m
	Certifications	~ 27801 sec	~ 7 h

Second Crawler: Evaluation



Third Crawler

The third crawler uses third party services in order to obtain the following elements:

- Website development quality
- Website last modified date
- Total visits/year
- Unique visits/year

The third crawler is consisted of three functions, each of them finds one of the above elements.

Third Crawler: “Website development quality”

·The function “website_dev” uses Google Insights API in order to get information about websites’ development stats.

·More particularly, the Google Insights API returns percentages which represent websites’ development quality.

· According to Google documentation 0-65% means Poor development quality, 66-79% Average development quality and 80-100% Good quality.

·This function returns 1 for Poor, 2 for Average, 3 for Good quality.

Third Crawler: “Website last modified date”

·The information regarding “last modified date” of a page is usually reserved for web site owners (via direct access to the web site files / database with their associated time stamps).

·However Internet Archive is a way to get an approximation of a page’s last modified date.

·The function “wayback_machine” uses information from Internet Archive consuming its API.

Third Crawler: “visits/year”

·In order to get websites’ total visits per year and websites’ unique visits per year, the function “advanced_stats” consumes information from StatsShow.com.

· StatsShow.com is a website analysis tool which provides estimated data of websites, using mathematical and statistical methods this tool can estimate websites’ total visits per year and unique visits per year.

·The function “advanced_stats” returns total visits/year and unique visits/year.

Third Crawler demo

```
In [5]: run_me('http://www.kallas-pap.com/')
```

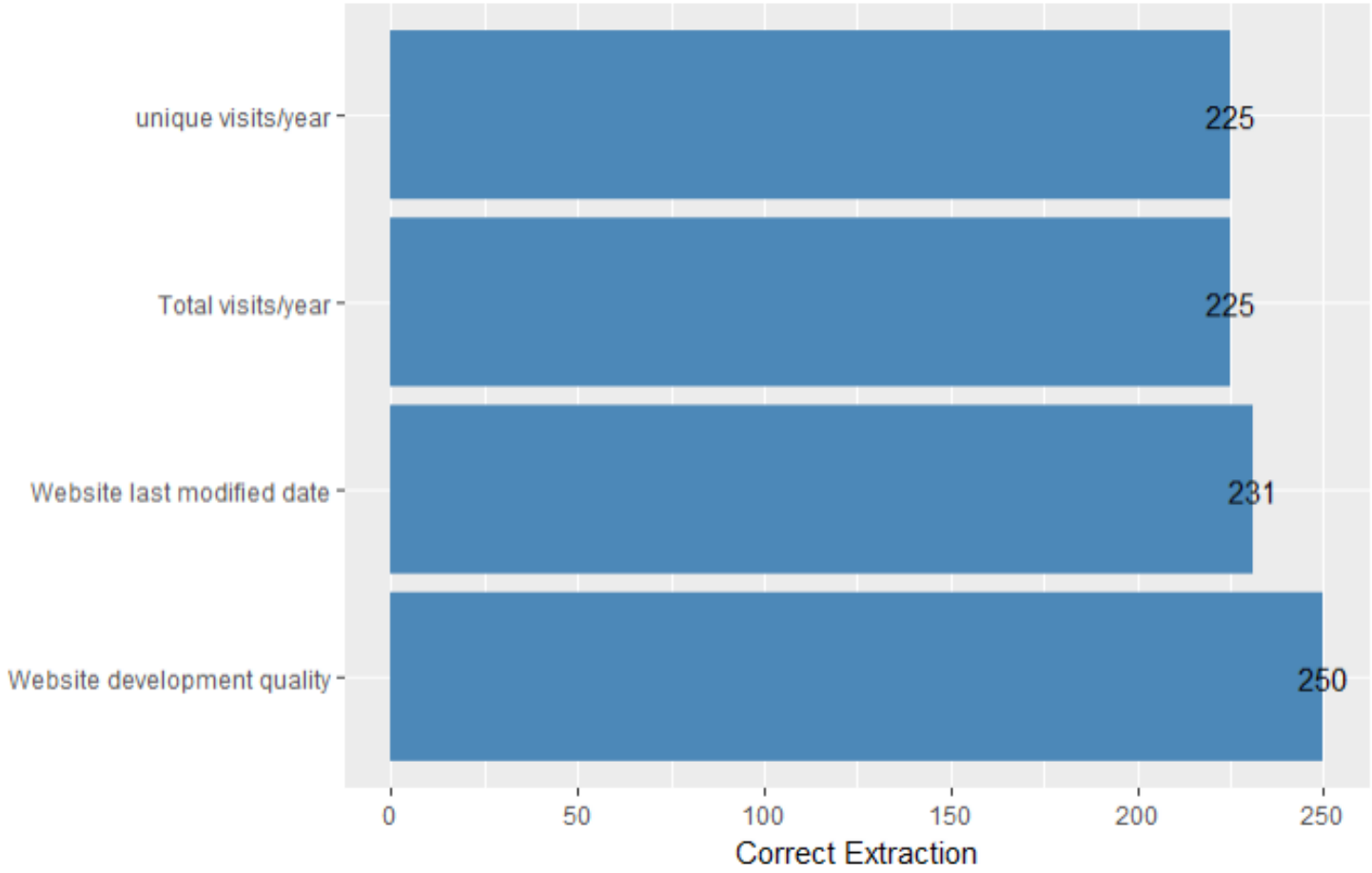
```
Last  modified date: 16/09/2017
Website development quality: 1
Total visits per year: 86,505
Unique visits per year: 39,055
```

Third Crawler

The table bellow demonstrates the execution time needed by these functions, in order to get website development quality, websites' last modified date, total and unique visits/year.

Third Crawler	Website development quality	~ 2390 sec	~ 40 m
	Website last modified date	~ 3153 sec	~ 50 m
	Total visits/year	~10370 sec	~ 2 h 45 m
	Unique visits/year		

Third Crawler Evaluation



Fourth Crawler

The **Fourth Crawler** extract from business websites the following data:

- Business street address
- Business location geographical coordinates
- Business zip code

Fourth Crawler

·Firstly, the crawler tries to estimate, if a website has Google Maps, which indicates business location.

· If the crawler finds Google Maps on website, we take business geographical coordinates from its source code and we apply reverse geocoding.

·In case we cannot locate google maps on website, we have stored on a file the Greek domain's zip codes.

·The crawler tries to match every possible zip code on each website's text. When the zip code is found, we apply reverse geocoding in order to find the area in which this zip code belongs.

Fourth Crawler demo

```
In [11]: run_me('http://www.veler.gr/')
```

```
Sapfous 29-35, Acharnes 136 71, Greece,38.0629975797081,23.73970891578256,136 71
```

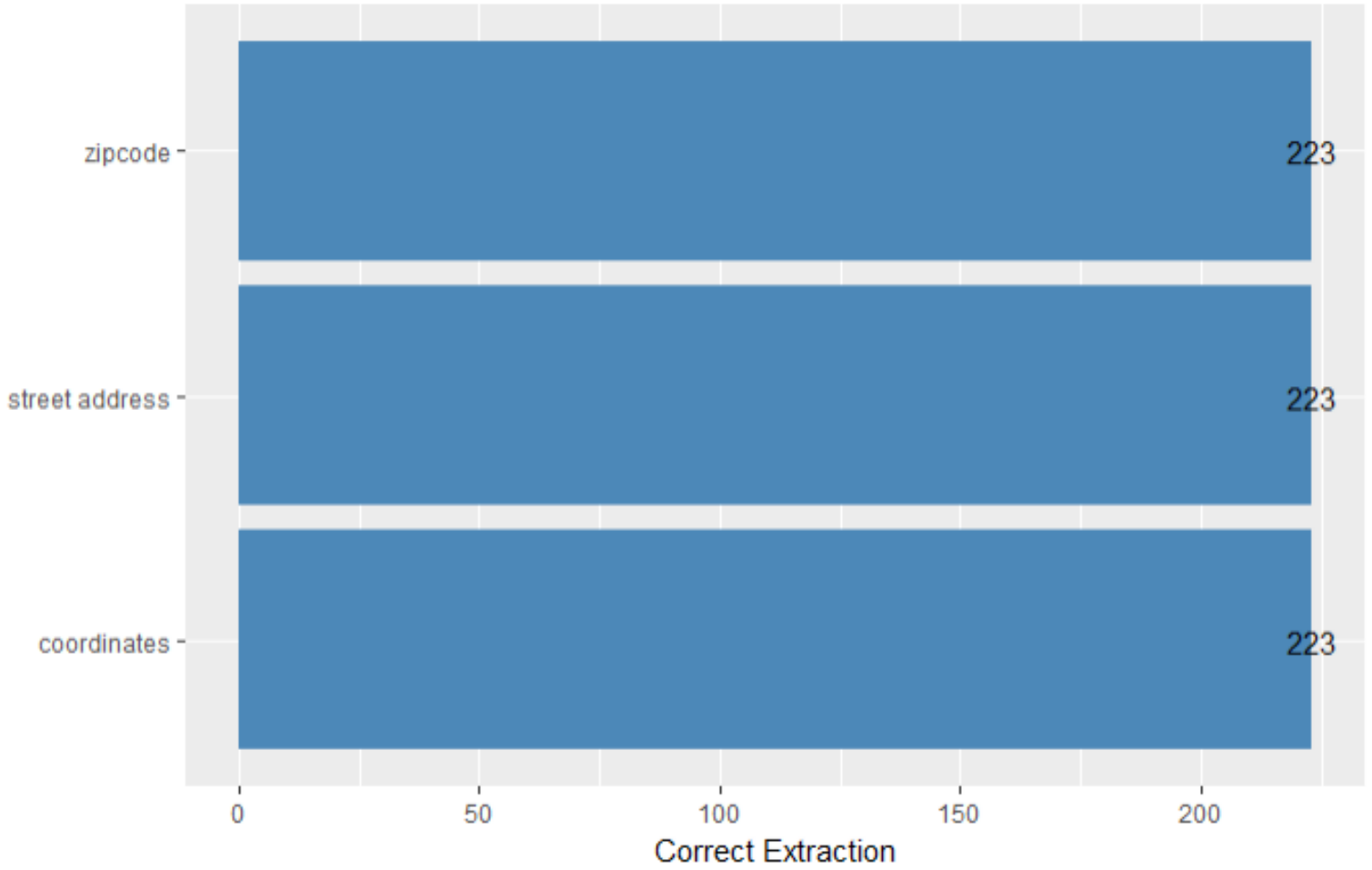
Fourth Crawler: Evaluation

·The time, **fourth crawler** needs to extract street address, zip code and geographical coordinates from the websites we have in our disposal is ~ **8** hours.

·And that's because of the use of the Selenium Web driver, which is a very slow module.

·In order to evaluate the results, we picked 250 random samples from these fourth crawler's results and we compared them with websites' raw information. For each field that the crawler extracts, we found the following:

Fourth Crawler



Fifth Crawler

The fifth crawler tries to locate terms on websites' text. More specifically tries to locate if the following terms are referred on websites:

- If "Corporate social responsibility" is referred (0 or 1)
- If "exports" is referred (0 or 1)
- If "imports" is referred (0 or 1)
- If "Customer support" is referred (0 or 1)
- If "private facilities" is referred (0 or 1)
- If "awards" is referred (0 or 1)
- If "representation" is referred (0 or 1)

Fifth Crawler demo

```
In [24]: run_me('http://www.biofresh-sa.com/')
```

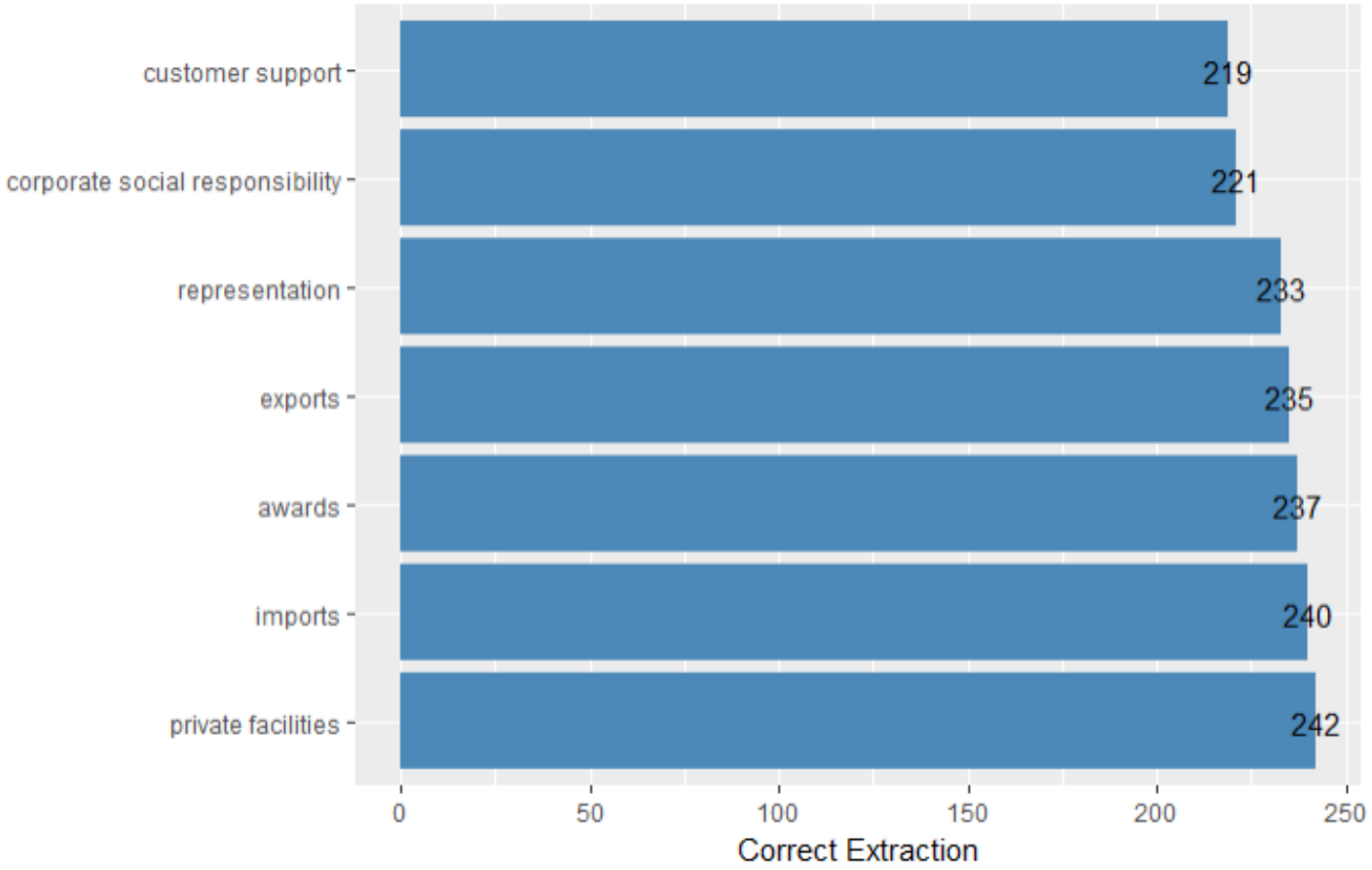
```
Αναφορά σε Ε.Κ.Ε: 0
Αναφορά σε εξαγωγές: 1
Αναφορά σε εισαγωγές: 0
Αναφορά σε αντιπροσωπείες: 1
Αναφορά σε Υποστήριξη Πελατών: 0
Αναφορά σε Ιδιόκτητες εγκαταστάσεις: 0
```

Fifth Crawler

- In order to share the work load, the Third Crawler is consisted of 3 functions.
- The first function tries to locate on website's text the terms "corporate social responsibility", "exports", "imports".
- The second function tries to locate "representation", "customer support", "private facilities" and the third function locates if "awards" is referred on websites' text.

Fifth Crawler	If "awards" (0 or 1)	~ 8543 sec	~ 2 h 20 m
	If "representation"	~ 15630 sec	~ 4 h 20 m
	If "private facilities"		
	If "customer support"		
	If "exports"	~ 19230 sec	~ 5 h
	If "imports"		
	If "Corporate social responsibility"		

Fifth Crawler: Evaluation



The code of this project is available in the following repository:
https://github.com/kapsali29/Crawler-for-Greek-business

Thank You

Questions?