

Web Crawling, Natural Language Processing & Information Extraction

Gianni Pantaleo

Dipartimento di Ingegneria dell'Informazione, DINFO

Università degli Studi di Firenze

Via S. Marta 3, 50139, Firenze, Italy

Tel: +39-055-2758517

DISIT Lab

<http://www.disit.org>

pierfrancesco.bellini@unifi.it

gianni.pantaleo@unifi.it

Security And Knowledge Management

Prof. Pierfrancesco Bellini

A.A. 2019/20



Outline

- 1. Sistemi di Web Crawling
 - 1.1 Introduzione
 - 1.2 Strategie di Crawling
 - 1.3 Robot Exclusion Protocol
 - 1.4 Concorrenza
- 2. Tecniche di Parsing ed Estrazione di Informazioni
 - 2.1 Introduzione
 - 2.2 NLP: Cenni Storici
 - 2.3 Fasi dell'Elaborazione in Linguaggio Naturale
 - 2.4 NLP Tools
- 3. Applicazioni



Introduzione: Dati e Informazione sul Web

- **The Zettabyte Era (1 ZB = 10^{12} GB):** Secondo studi e stime di **CISCO**, la quantità di dati prodotti ed archiviati nel cloud data center globale sarà pari nel **2020**, a **1.8 Zettabytes** (1.8 mila miliardi di Gigabytes), equivalente alla quantità di spazio archiviabile in circa **45 milioni di DVD all'ora**, con una crescita di 5 volte rispetto al 2015 (*).



- La quantità di traffico dati nel web globale (*global data center traffic*), arriverà a **15.3 Zettabyte nel 2020 (**)**.

* [Cisco Global Cloud Index: Forecast and Methodology, 2015–2020 White Paper](#) (Cisco Public Knowledge)

** <http://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/index.html>
(Source: Global Cloud Index Infographic [GCI 2016](#))

Introduzione: Era dell'Information Technology



LISA GROSSMAN SCIENCE 10.19.10 01:30 PM

TWITTER CAN PREDICT THE STOCK MARKET



The emotional roller coaster captured on Twitter can predict the ups and downs of the stock market, a new study

BUSINESS INSIDER UK FINANCE

The ECB says Twitter can predict the stock market

Oscar Williams-Grut
Jul. 22, 2015, 11:40 AM 1,790

FACEBOOK LINKEDIN TWITTER EMAIL PRINT

The European Central Bank (ECB) just put out an interesting study looking at whether Twitter and Google can be used to predict stock market moves — and its conclusion is, for Twitter, it can.

The ECB says: "Twitter bullishness has a statistically and economically significant predictive value in respect of share prices in the United States, the United Kingdom and Canada."



In other words, what people are Twitter CEO Dick Costolo, right, celebrates the Twitter IPO with Twitter ders, from left, Jack Dorsey, Biz Stone and Evan Williams on the



Outline

- 1. Sistemi di Web Crawling

- 1.1 Introduzione 
- 1.2 Strategie di Crawling
- 1.3 Robot Exclusion Protocol
- 1.4 Concorrenza

- 2. Tecniche di Parsing ed Estrazione di Informazioni

- 2.1 Introduzione
- 2.2 NLP: Cenni Storici
- 2.3 Fasi dell'Elaborazione in Linguaggio Naturale
- 2.4 NLP Tools

- 3. Applicazioni



1.1 - Introduzione: Generalità di un Web Crawler

Definizione di **Web Crawler / Spider / Scraper / Bot**

➤ Un **Crawler Web** è un software usato per collezionare ed archiviare il contenuto di pagine web, documenti presenti in una rete o in un database



➤ Definito anche **Web Spider, Web Robot, Web Scraper** o semplicemente **Bot**

1.1 - Introduzione: Aree Applicative

Un **Web Crawler** può essere usato per:

- Creare un indice generale (**general Web search**)
- Creare indici di topics o argomenti specifici (**vertical Web search**)
- Archiviare il contenuto di pagine web e documenti (**Web archival**)
- Analizzare il contenuto di pagine e siti web per produrre statistiche e analisi aggregate (**Web characterization**)
- Tenere copie o repliche di siti web (**Web mirroring**)
- Analisi di siti web (**Web site analysis**)



1.1 - Introduzione: Processo di Crawling



1.1 - Introduzione: Principali Search Bot

Bot Name	% of Sites Crawled	Bot Type
Googlebot	96%	Search Bot
Baidu Spider	89%	Search Bot
MSN Bot/BingBot	82%	Search Bot
Yandex Bot	73%	Search Bot
Soso Spider	61%	Search Bot
ExaBot	35%	Search Bot
Sogou Spider	31%	Search Bot
Google Plus Share	24%	Crawler
Facebook External Hit	24%	Crawler
Google Feedfetcher	22%	Feed Fetcher

Fonte: <https://www.incapsula.com/blog/know-your-top-10-bots.html>

1.1 - Introduzione: Principali Crawler Open Source



<https://webarchive.jira.com/wiki/display/Heritrix>



<http://nutch.apache.org/>



<https://www.cs.cmu.edu/~rcm/websphinx/>



PORTIA

<https://portia.readthedocs.io/en/latest/>

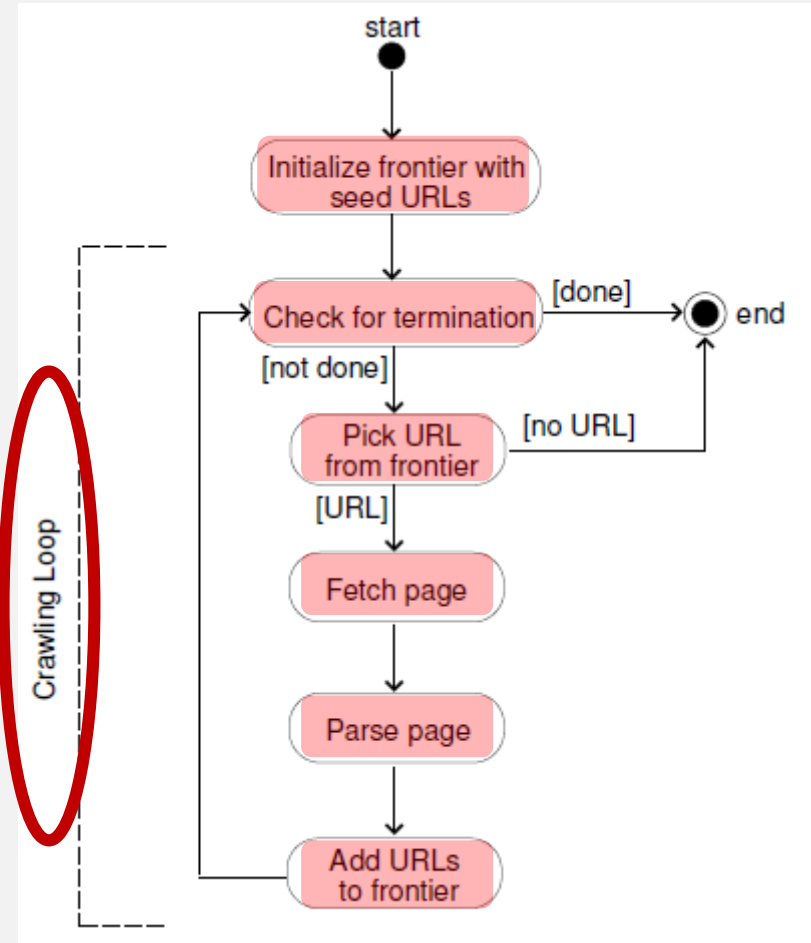
1.1 - Introduzione: Principi di Funzionamento



➤ Il web è visto come un grafo: i nodi sono le pagine web e gli archi sono i link (*hyperlinks*)

➤ Un crawler viene di solito utilizzato per archiviare le pagine web visitate per una loro successiva elaborazione (estrazione di informazioni, indicizzazione, ecc..)

➤ Analizza le pagine partendo da una lista di indirizzi web (**URL**) iniziali. Gli URL sono contenuti in una struttura dati che si chiama **frontiera**



1.1 - Introduzione: La Frontiera di un Crawler

- E' una struttura dati dinamica che contiene URL non ancora visitati
- Può riempirsi molto velocemente rispetto alle pagine web via via crawlate
- Con una media (stimata) di n links per pagina, la velocità di popolazione della frontiera è lineare, ma cresce di circa n volte rispetto al numero delle pagine già crawlate
- Occorre limitare la sua dimensione con un valore massimo



Occorre stabilire un meccanismo per decidere quale URL ignorare in caso di limite raggiunto

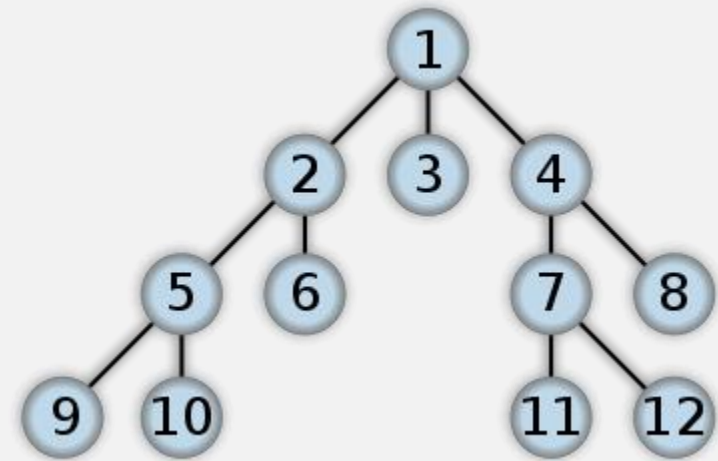
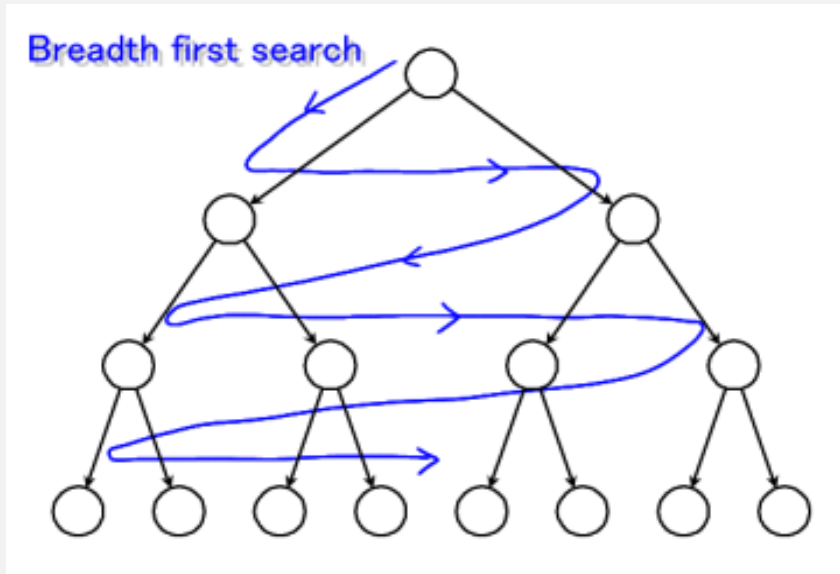
Outline

- 1. Sistemi di Web Crawling
 - 1.1 Introduzione
 - 1.2 Strategie di Crawling 
 - 1.3 Robot Exclusion Protocol
 - 1.4 Concorrenza
- 2. Tecniche di Parsing ed Estrazione di Informazioni
 - 2.1 Introduzione
 - 2.2 NLP: Cenni Storici
 - 2.3 Fasi dell'Elaborazione in Linguaggio Naturale
 - 2.4 NLP Tools
- 3. Applicazioni



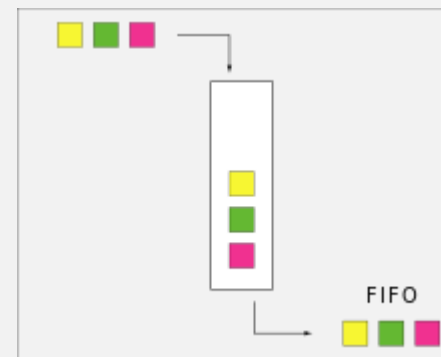
1.2 - Strategie di Crawling: Breadth First Search

➤ Breadth First Search



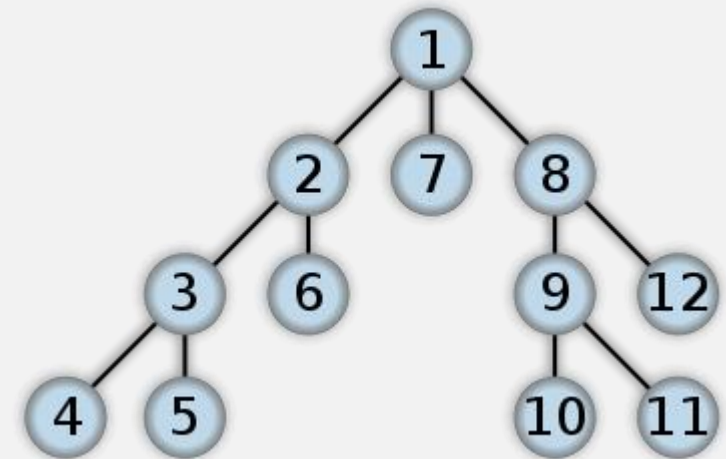
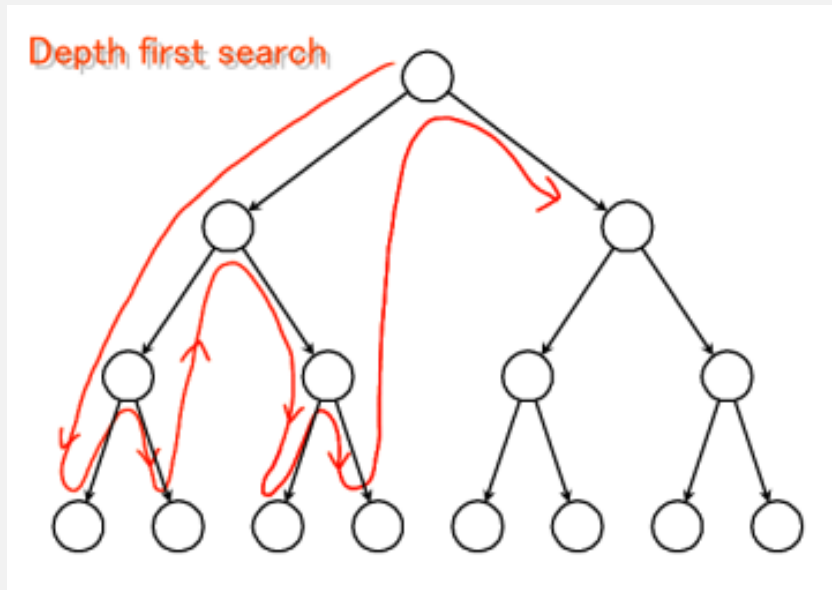
➤ Struttura dati lineare
con politica

FIFO : First In First Out (CODA)



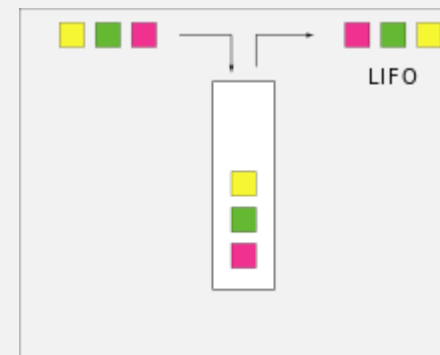
1.2 - Strategie di Crawling: Depth First Search

➤ Depth First Search



- Struttura dati lineare con politica

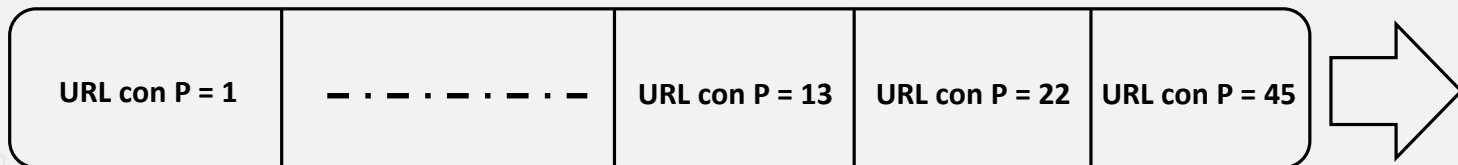
LIFO: Last In First Out (PILA o STACK)



1.2 - Strategie di Crawling

➤ Priority Search

- Struttura dati lineare: l'URL con maggiore priorità viene scelto
- Array dinamico ordinato in base allo score attribuito ad ogni URL
- Gli URL sono aggiunti nella frontiera in maniera tale da preservarne l'ordine in base allo score



1.2 - Strategie di Crawling

- Evitare il fetch della stessa pagina:
 - Tenere in memoria una lista delle pagine visitate

- La dimensione della frontiera cresce velocemente
 - Può essere necessaria una politica di priorità sugli URLs

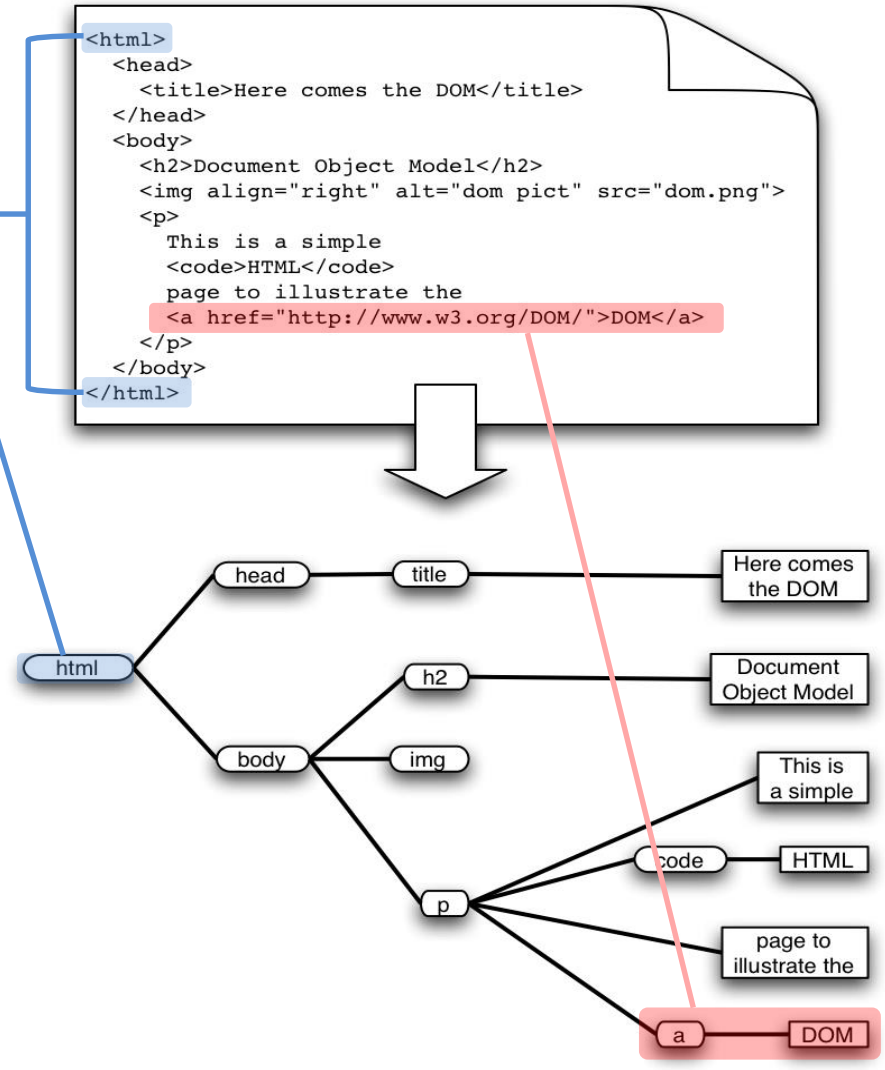
- Fetcher deve essere robusto
 - Evitare crash in caso di fallimenti nei downloads
 - Prevedere meccanismi di timeout

- Determinazione dei file non desiderati
 - Esaminare estensioni dei file (filtrando formati come multimedia, immagini, video ecc.)
 - Content-Type (MIME) headers



1.2 - Strategie di Crawling - Parsing per estrazione di Informazioni

- I documenti HTML hanno una struttura ad albero - DOM (Document Object Model) definite da tag
- Spesso i documenti HTML non rispettano gli standard di sintassi
- Occorre trattare le entità e i tag HTML
- Vi sono molti formati diversi di files:
 - ❖ Flash, SVG, RSS, AJAX...



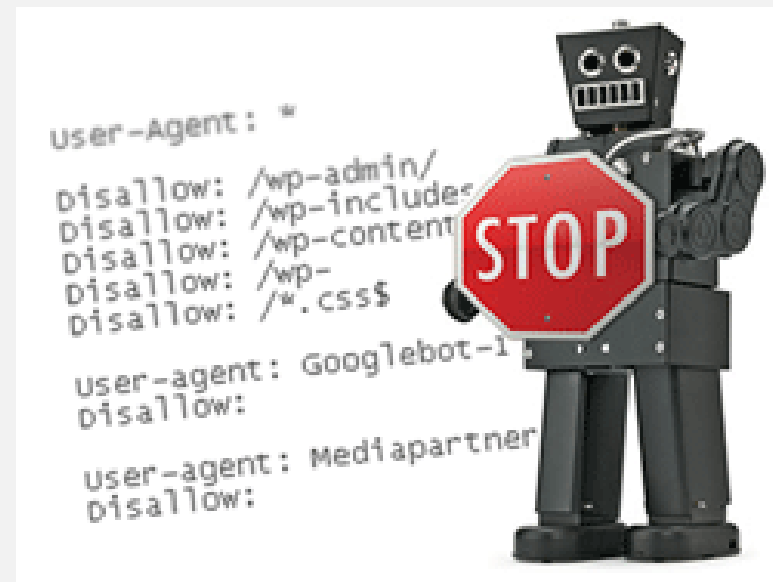
Outline

- 1. Sistemi di Web Crawling
 - 1.1 Introduzione
 - 1.2 Strategie di Crawling
 - 1.3 Robot Exclusion Protocol 
 - 1.4 Concorrenza
- 2. Tecniche di Parsing ed Estrazione di Informazioni
 - 2.1 Introduzione
 - 2.2 NLP: Cenni Storici
 - 2.3 Fasi dell'Elaborazione in Linguaggio Naturale
 - 2.4 NLP Tools
- 3. Applicazioni



1.3 Robot Exclusion Protocol

- Un server può specificare a quale parte dell'albero dei propri documenti può essere accessibile a un crawler (robot) esterno
- Questa informazione è nel file 'robots.txt' posto nell' HTTP root directory
- Un crawler dovrebbe sempre verificare la presenza di questo file prima di inviare richieste al server



1.3 Robot Exclusion Protocol

www.apple.com/robots.txt

```
# robots.txt for http://www.apple.com/
```

```
User-agent: *
```

```
Disallow:
```

Tutti i crawlers ...


...possono analizzare qualsiasi
pagina!



1.3 Robot Exclusion Protocol

www.microsoft.com/robots.txt

```
# Robots.txt file for http://www.microsoft.com
```

```
User-agent: *  
Disallow: /canada/Library/mnp/2/asp/   
Disallow: /communities/bin.aspx  
Disallow: /communities/eventdetails.aspx  
Disallow: /communities/blogs/PortalResults.aspx  
Disallow: /communities/rss.aspx  
Disallow: /downloads/Browse.aspx  
Disallow: /downloads/info.aspx  
Disallow: /france/formation/centres/planning.asp  
Disallow: /france/mnp_utility.aspx  
Disallow: /germany/library/images/mnp/  
Disallow: /germany/mnp_utility.aspx  
Disallow: /info/customerror.htm  
#etc...
```

Tutti i crawlers ...

... non hanno accesso a questi percorsi del grafo del sito web

1.3 Robot Exclusion Protocol

www.springer.com/robots.txt

Robots.txt for <http://www.springer.com> (fragment)

User-agent: Googlebot
Disallow: /chl/*
Disallow: /uk/*
Disallow: /italy/*
Disallow: /france/*

Google crawler può analizzare tutte le pagine eccetto queste

User-agent: slurp
Disallow:
Crawl-delay: 2

User-agent: MSNBot
Disallow:
Crawl-delay: 2

Yahoo e MSN/Windows Live possono analizzare tutte le pagine ma il processo deve essere "lento" (2 secondi)


User-agent: scooter
Disallow:

AltaVista non ha limiti

all others
User-agent: *
Disallow: /

A tutti gli altri crawlers non è permesso niente

Outline

- 1. Sistemi di Web Crawling
 - 1.1 Introduzione
 - 1.2 Strategie di Crawling
 - 1.3 Robot Exclusion Protocol
 - 1.4 Concorrenza 
- 2. Tecniche di Parsing ed Estrazione di Informazioni
 - 2.1 Introduzione
 - 2.2 NLP: Cenni Storici
 - 2.3 Fasi dell'Elaborazione in Linguaggio Naturale
 - 2.4 NLP Tools
- 3. Applicazioni



1.4 - Concorrenza

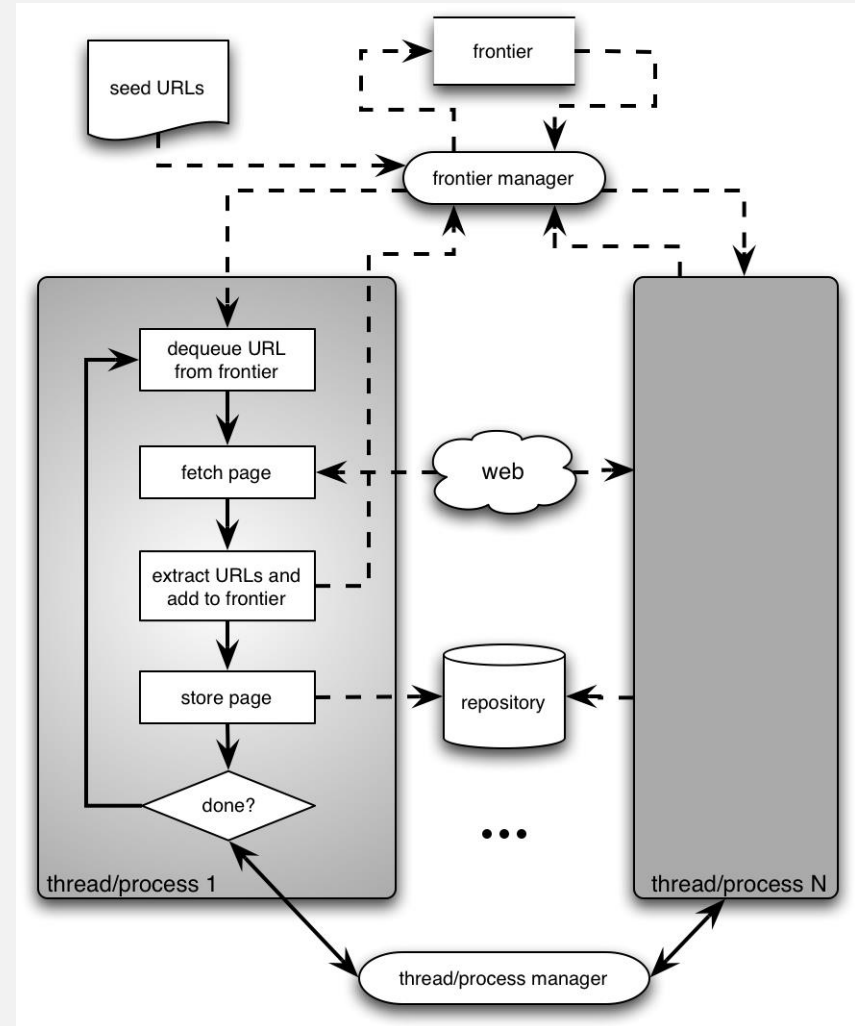
- Le operazioni effettuate dal crawler possono richiedere diverso tempo:
 - Risoluzione indirizzi IP
 - Connessioni al server e invio risposte
 - Ricezione pagina di risposta

- Soluzione: Ridurre i suddetti tempi eseguendo la scansione di più pagine in maniera concorrente




1.4 - Concorrenza

- Ogni thread lavora come un crawler sequenziale e condivide le strutture dati: frontiera e repository (concorrenza in lettura)
- Le strutture dati condivise devono essere sincronizzate (concorrenza in scrittura)



Outline

- 1. Sistemi di Web Crawling
 - 1.1 Introduzione
 - 1.2 Strategie di Crawling
 - 1.3 Robot Exclusion Protocol
 - 1.4 Concorrenza
- 2. Tecniche di Parsing ed Estrazione di Informazioni
 - 2.1 Introduzione 
 - 2.2 NLP: Cenni Storici
 - 2.3 Fasi dell'Elaborazione in Linguaggio Naturale
 - 2.4 NLP Tools

3. Applicazioni



2.1 - Tecniche di Parsing: Introduzione

➤ Natural Language Processing

➤ Scenario / Requisiti

- ❖ Dotare l'IA delle abilità linguistiche proprie dell'essere umano
- ❖ Comprensione e generazione di testo non strutturato (*linguaggio naturale*)
- ❖ Contesto multi-language: differenti regole e strutture a seconda della lingua

➤ Applicazioni

- ❖ Generalizzazione delle query nei motori di ricerca
 - *"Chi si occupa di sistemi distribuiti nell'Università di Firenze?"*
- ❖ Supporto automatizzato per Help-Desk
- ❖ Tutoring assistito (e-tutoring, e-teaching...)
- ❖ Summarization: creare compendi da una collezione eterogenea di documenti
- ❖ Machine translation: tradurre testi in lingue diverse



2.1 - Tecniche di Parsing: Introduzione

➤ Scenario / Requisiti

❖ I linguaggi sono purtroppo ambigui.

❖ Le ambiguità si possono avere a 4 livelli:

✓ Ambiguità lessicale: «*attacco*» (verbo, sostantivo)

✓ Ambiguità sintattica (strutturale): «*leri ho visto l'uomo col binocolo*»

✓ Ambiguità semantica: «*acuto*» (persona intelligente, tipo di suono)

✓ Ambiguità pragmatica: «*la porta è aperta...*»

L'intensione comunicativa può essere recepita diversamente dagli interlocutori:


- Invito ad entrare liberamente...
- Meglio chiudere, si è creata una corrente d'aria...



Ciò rende il processo di elaborazione automatica del linguaggio naturale un task molto complesso !



Outline

- 1. Sistemi di Web Crawling
 - 1.1 Introduzione
 - 1.2 Strategie di Crawling
 - 1.3 Robot Exclusion Protocol
 - 1.4 Concorrenza
- 2. Tecniche di Parsing ed Estrazione di Informazioni
 - 2.1 Introduzione
 - 2.2 NLP: Cenni Storici 
 - 2.3 Fasi dell'Elaborazione in Linguaggio Naturale
 - 2.4 NLP Tools

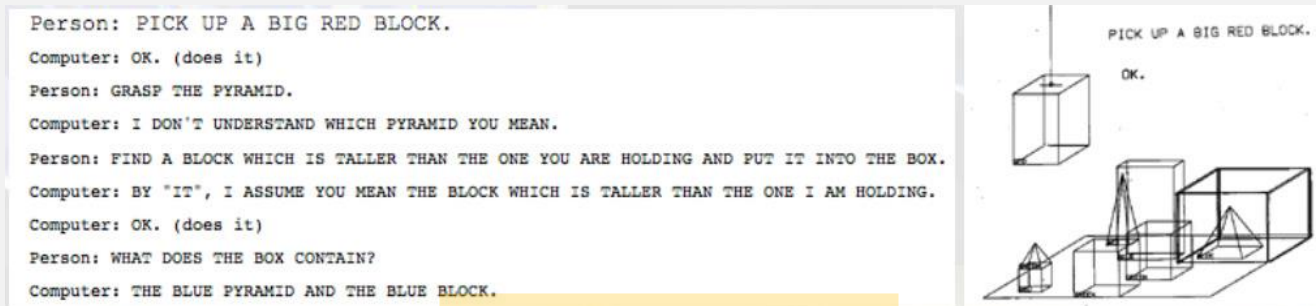
3. Applicazioni



2.2 - Tecniche di Parsing – NLP: Cenni Storici

➤ SHRDLU [Winograd - 1971]

- ❖ Interfaccia per automa preposto al semplice movimento di blocchi 3D
- ❖ Dominio limitato, query semplici



➤ Apple Siri [IOS Siri - 2010]

- ❖ Virtual personal assistant
- ❖ Knowledge navigator
- ❖ User recommendation system.



➤ Google Assistant

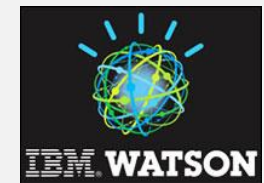


➤ Microsoft Cortana



➤ IBM Watson [IBM Watson - 2012]

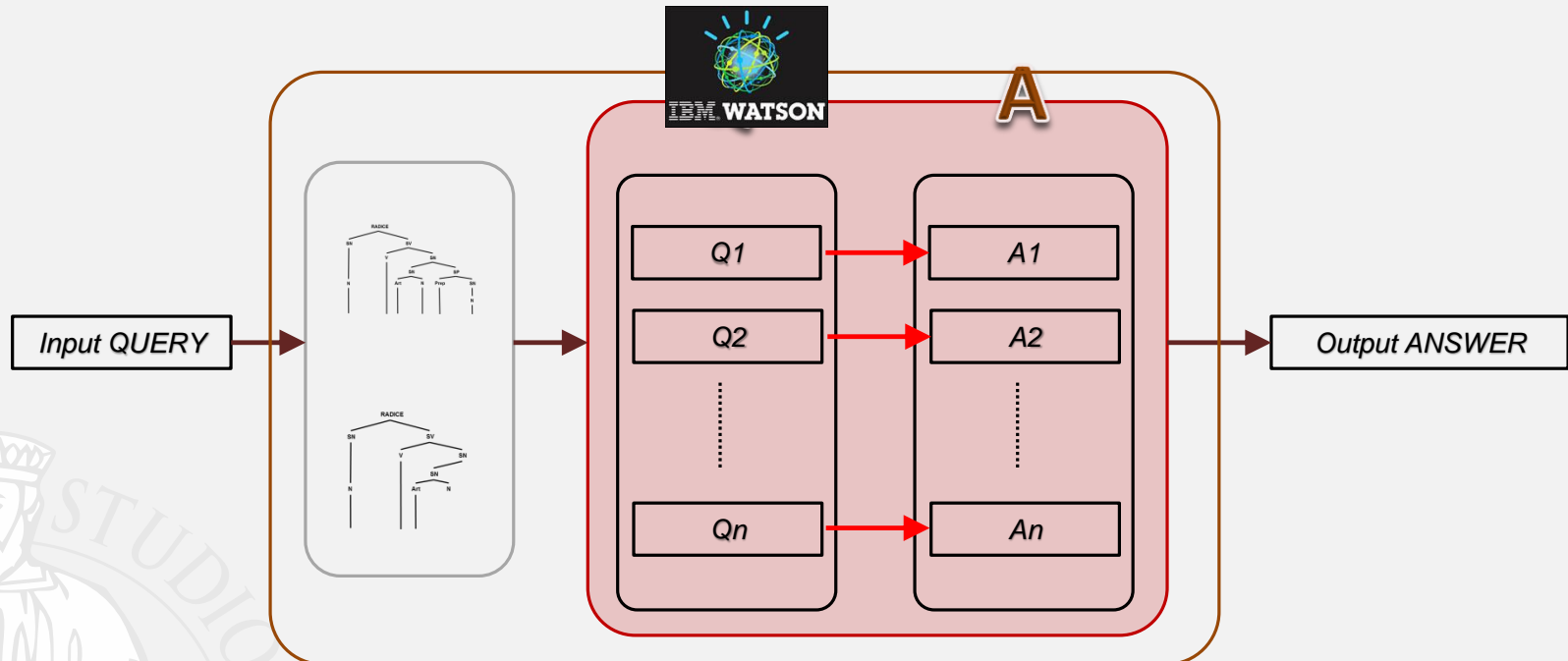
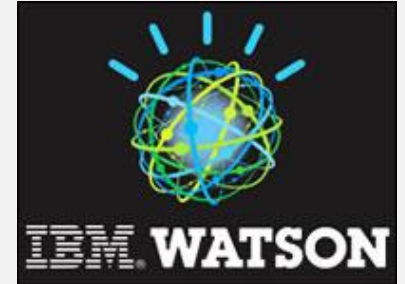
- ❖ Nato come sistema Question & Answer per il quiz televisivo americano *Jeopardy!*, successivamente è stato sviluppato come *cognitive system* completo con meccanismi di auto-apprendimento.




2.2 - Tecniche di Parsing – NLP: Cenni Storici

➤ IBM Watson

- ❖ Database built-in in cui sono già presenti le domande relazionate con le rispettive risposte corrette.
- ❖ Il sistema si limita ad analizzare la query in input, attraverso l'elaborazione del grafo sintattico, cercando un match con una delle domande presenti nel set.




Outline

- 1. Sistemi di Web Crawling
 - 1.1 Introduzione
 - 1.2 Strategie di Crawling
 - 1.3 Robot Exclusion Protocol
 - 1.4 Concorrenza
- 2. Tecniche di Parsing ed Estrazione di Informazioni
 - 2.1 Introduzione
 - 2.2 NLP: Cenni Storici
 - 2.3 Fasi dell'Elaborazione in Linguaggio Naturale 
 - 2.4 NLP Tools


3. Applicazioni




2.3 – Fasi di Elaborazione in Linguaggio Naturale



Morphological Analysis: le parole vengono analizzate (distinzione dei morfemi che le compongono) ed i simboli (punteggiature) vengono separati dalle parole .



Syntactic Analysis: Le sequenze di parole sono trasformate in strutture che mostrano come le parole sono in relazione l'una con l'altra.



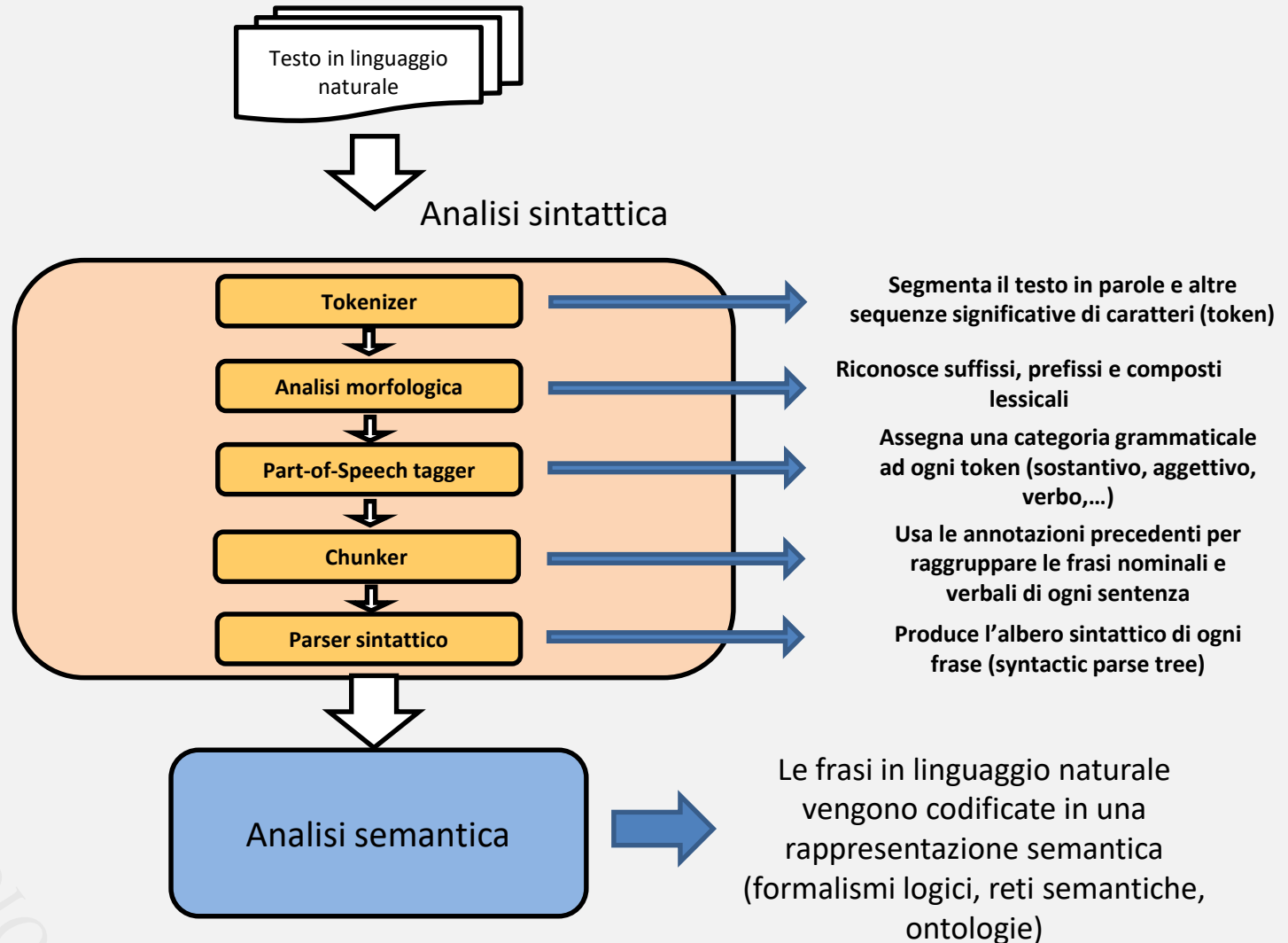
Semantic Analysis: Viene assegnato un significato alle strutture sintattiche trovate.

Discourse integration: il significato di una frase spesso dipende dalla frase che la precede e può influenzare quello della frase che la segue.

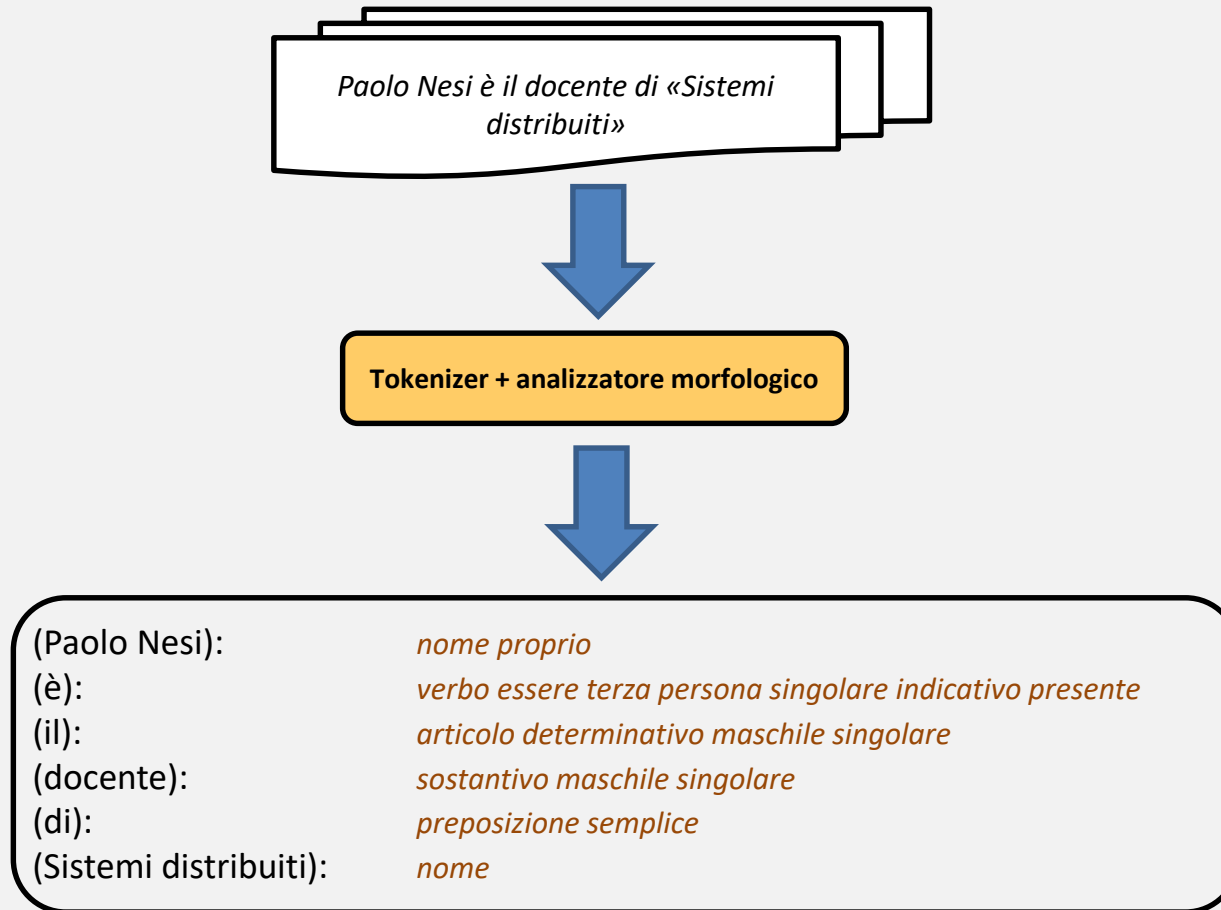
Pragmatic Analysis: la frase è reinterpretata per determinare il significato specifico della frase stessa.



2.3 – Fasi di Elaborazione in Linguaggio Naturale



2.3 – Fasi di Elaborazione in Linguaggio Naturale



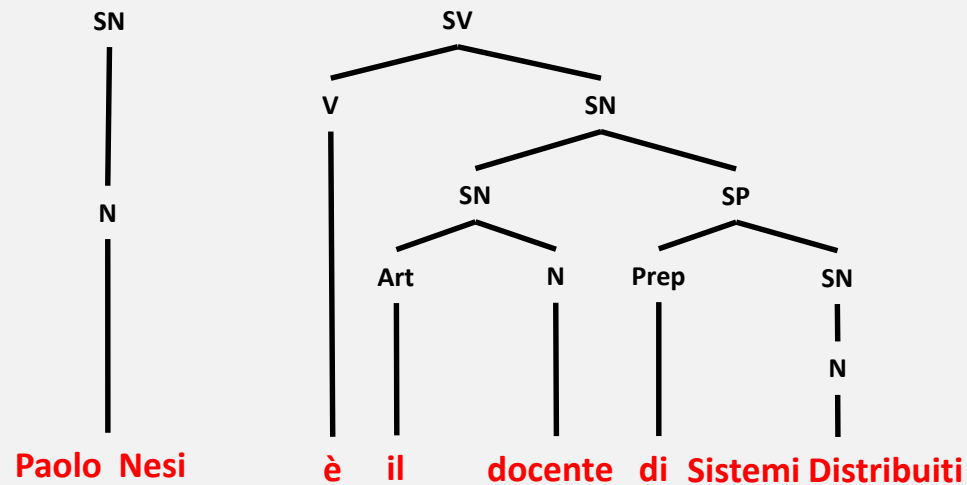
(Paolo Nesi):	<i>nome proprio</i>
(è):	<i>verbo essere terza persona singolare indicativo presente</i>
(il):	<i>articolo determinativo maschile singolare</i>
(docente):	<i>sostantivo maschile singolare</i>
(di):	<i>preposizione semplice</i>
(Sistemi distribuiti):	<i>nome</i>



**Part-of-Speech + chunker +
parser sintattico**



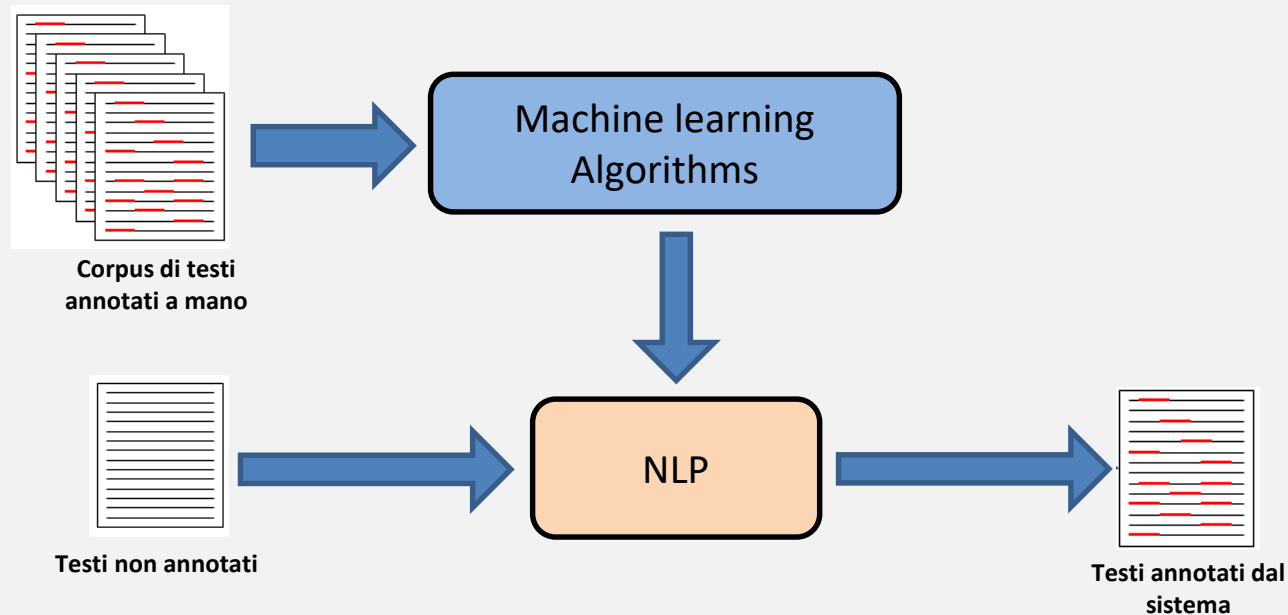
RADICE




SN: sentenza nominale
SV: sentenze verbale
SP: sentenza preposizionale
N: nome
V: verbo
Art: articolo
Prep: preposizione

2.3 – Fasi di Elaborazione in Linguaggio Naturale

I sistemi di NLP usano principalmente algoritmi di machine learning addestrati su grandi corpus di testi annotati a mano



Outline

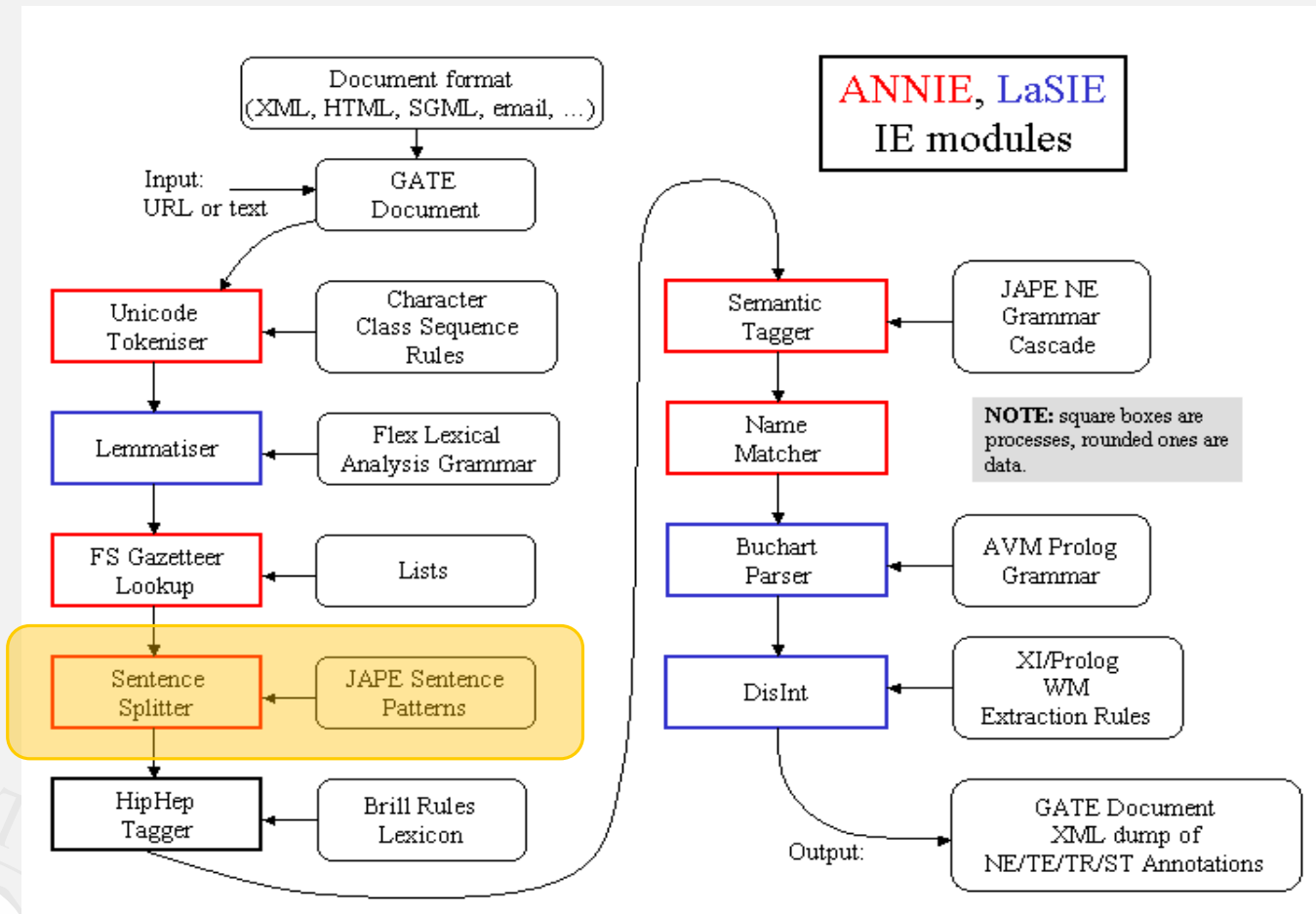
- 1. Sistemi di Web Crawling
 - 1.1 Introduzione
 - 1.2 Strategie di Crawling
 - 1.3 Robot Exclusion Protocol
 - 1.4 Concorrenza
- 2. Tecniche di Parsing ed Estrazione di Informazioni
 - 2.1 Introduzione
 - 2.2 NLP: Cenni Storici
 - 2.3 Fasi dell'Elaborazione in Linguaggio Naturale
 - 2.4 NLP Tools 

3. Applicazioni



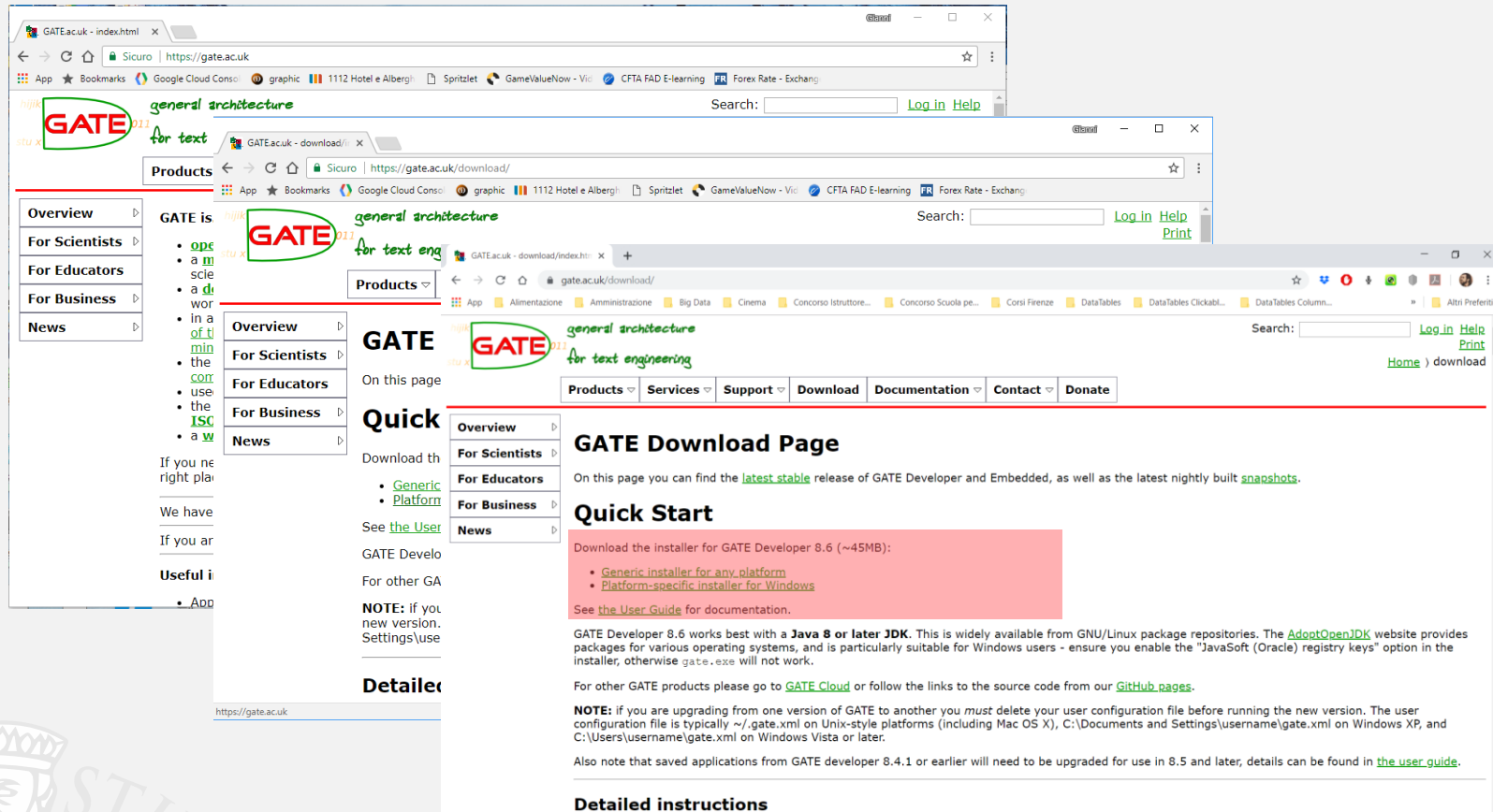
2.4 - NLP Tools: GATE

- **GATE – General Architecture for Text Engineering** (<https://gate.ac.uk/>)
 - ❖ Supporta documenti plain text, HTML, XML ...



2.4 - NLP Tools: GATE

- **GATE – General Architecture for Text Engineering** (<https://gate.ac.uk/>)



The screenshot shows the GATE website interface. The main content area is titled "GATE Download Page" and includes a "Quick Start" section. The "Quick Start" section contains the following text:

Download the installer for GATE Developer 8.6 (~45MB):

- [Generic installer for any platform](#)
- [Platform-specific installer for Windows](#)

See [the User Guide](#) for documentation.

GATE Developer 8.6 works best with a **Java 8 or later JDK**. This is widely available from GNU/Linux package repositories. The [AdoptOpenJDK](#) website provides packages for various operating systems, and is particularly suitable for Windows users - ensure you enable the "JavaSoft (Oracle) registry keys" option in the installer, otherwise `gate.exe` will not work.

For other GATE products please go to [GATE Cloud](#) or follow the links to the source code from our [GitHub pages](#).

NOTE: if you are upgrading from one version of GATE to another you *must* delete your user configuration file before running the new version. The user configuration file is typically `~/gate.xml` on Unix-style platforms (including Mac OS X), `C:\Documents and Settings\username\gate.xml` on Windows XP, and `C:\Users\username\gate.xml` on Windows Vista or later.

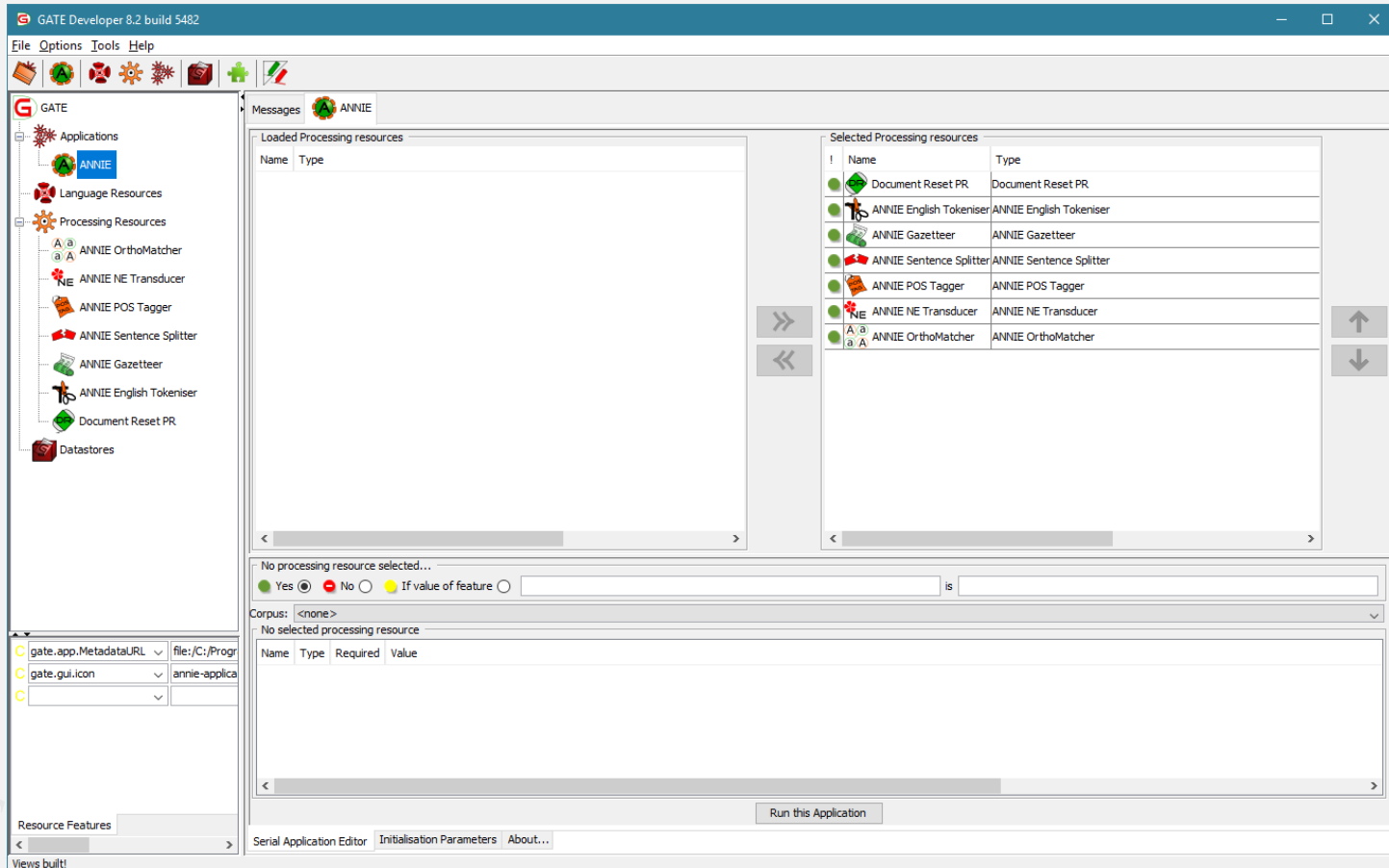
Also note that saved applications from GATE developer 8.4.1 or earlier will need to be upgraded for use in 8.5 and later, details can be found in [the user guide](#).

The "Detailed instructions" section is also visible at the bottom of the page.



2.4 - NLP Tools: GATE

➤ GATE: Interfaccia grafica e caricamento plugin principali



2.4 - NLP Tools: GATE

➤ ANNIE (A Nearly New IE System)

- ❖ Utilizza regole per il **Part of Speech (POS) Tagging**. Tali regole sono predefinite (tag html) o customizzate dall'utente

- ❖ Una regola ha un left hand side (LHS) e un right hand side (RHS)
 - **LHS**: espressione regolare da riscontrare nel testo in input
 - **RHS**: descrive le annotazioni che devono essere aggiunte all'*AnnotationSet*
 - Utilizzano regole predefinite (tag html) o customizzate dall'utente

- ❖ Sintassi:
 - ❖ `{LHS} > {Annotation type}; {attribute1}={value1};...;{attribute n}={value n}`

- ❖ Es.:
 - ❖ `"UPPERCASE_LETTER" "LOWERCASE_LETTER"* > Token; orth=upperInitial; kind=word.`

- ❖ Tipi di Token previsti: Word, Number, Symbol, Punctuation, Space Token

- ❖ Guida online: <https://gate.ac.uk/sale/thakker-jape-tutorial/GATE%20JAPE%20manual.pdf>

2.4 - NLP Tools: GATE

➤ JAPE (Java Annotations Pattern Engine)

- ❖ Permette di ricercare espressioni regolari nel testo in input
- ❖ Regole composte da LHS e RHS

❖ Es. di sintassi di una regola JAPE:

```
Phase: Address_Retrieval
Input: Token Lookup
Options: control = appelt
```

Headers della regola

```
Rule: FindStreetAddress
Priority: 20
```

Nome della regola e priorità

```
(
  ({Token.string == "Via"} | Token.string == "Piazza"} | {Token.string == "Largo"})
  ({{Lookup.majorType == NomeProprio} | {Lookup.majorType == Cognome}} |
  {{Lookup.majorType == Cognome}})
)
```

```
:address
```

Label

```
-->
```

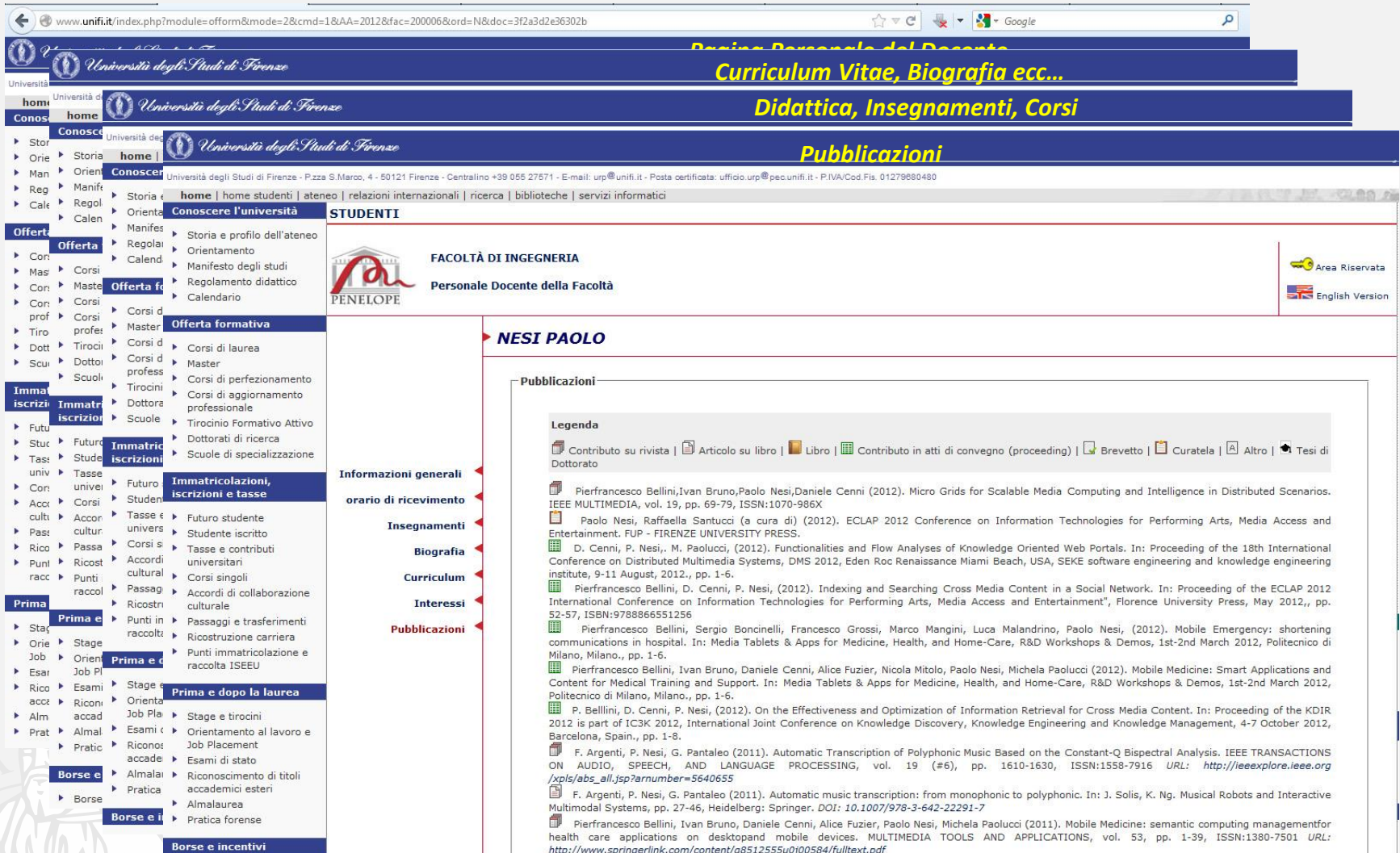
RHS

LHS

```
: address.indirizzo = {rule = "FindStreetAddress"}
```



2.4 - NLP Tools: GATE



The screenshot shows the website of the University of Florence (www.unifi.it). The page is titled "NESI PAOLO" and displays a list of publications. The navigation menu on the left includes sections like "Storia", "Orientamento", "Manif", "Regol", "Calen", "Offerta", "Cor", "Mas", "Cor", "Cor", "prof", "Tiro", "Dott", "Scu", "Immatricolazioni", "Futu", "Stuc", "Tas", "univ", "Cor", "Acc", "cult", "Pas", "Rico", "Punt", "racc", "Prima", "Stag", "Orie", "Job", "Esar", "Rico", "acc", "Alm", "Prat", "Borse", "Borse e i", and "Borse e incentivi".

The main content area is titled "Pubblicazioni" and includes a "Legenda" section with icons for different publication types: Contributo su rivista, Articolo su libro, Libro, Contributo in atti di convegno (proceeding), Brevetto, Curatela, Altro, and Tesi di Dottorato.

The list of publications includes:

- Pierfrancesco Bellini, Ivan Bruno, Paolo Nesi, Daniele Cenni (2012). Micro Grids for Scalable Media Computing and Intelligence in Distributed Scenarios. IEEE MULTIMEDIA, vol. 19, pp. 69-79, ISSN:1070-986X
- Paolo Nesi, Raffaella Santucci (a cura di) (2012). ECLAP 2012 Conference on Information Technologies for Performing Arts, Media Access and Entertainment. FUP - FIRENZE UNIVERSITY PRESS.
- D. Cenni, P. Nesi, M. Paolucci, (2012). Functionalities and Flow Analyses of Knowledge Oriented Web Portals. In: Proceeding of the 18th International Conference on Distributed Multimedia Systems, DMS 2012, Eden Roc Renaissance Miami Beach, USA, SEKE software engineering and knowledge engineering institute, 9-11 August, 2012., pp. 1-6.
- Pierfrancesco Bellini, D. Cenni, P. Nesi, (2012). Indexing and Searching Cross Media Content in a Social Network. In: Proceeding of the ECLAP 2012 International Conference on Information Technologies for Performing Arts, Media Access and Entertainment", Florence University Press, May 2012., pp. 52-57, ISBN:9788866551256
- Pierfrancesco Bellini, Sergio Boncinelli, Francesco Grossi, Marco Mangini, Luca Malandrino, Paolo Nesi, (2012). Mobile Emergency: shortening communications in hospital. In: Media Tablets & Apps for Medicine, Health, and Home-Care, R&D Workshops & Demos, 1st-2nd March 2012, Politecnico di Milano, Milano., pp. 1-6.
- Pierfrancesco Bellini, Ivan Bruno, Daniele Cenni, Alice Fuzier, Nicola Mitolo, Paolo Nesi, Michela Paolucci (2012). Mobile Medicine: Smart Applications and Content for Medical Training and Support. In: Media Tablets & Apps for Medicine, Health, and Home-Care, R&D Workshops & Demos, 1st-2nd March 2012, Politecnico di Milano, Milano., pp. 1-6.
- P. Bellini, D. Cenni, P. Nesi, (2012). On the Effectiveness and Optimization of Information Retrieval for Cross Media Content. In: Proceeding of the KDIR 2012 is part of IC3K 2012, International Joint Conference on Knowledge Engineering and Knowledge Management, 4-7 October 2012, Barcelona, Spain., pp. 1-8.
- F. Argenti, P. Nesi, G. Pantaleo (2011). Automatic Transcription of Polyphonic Music Based on the Constant-Q Bispectral Analysis. IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, vol. 19 (#6), pp. 1610-1630, ISSN:1558-7916 URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5640655
- F. Argenti, P. Nesi, G. Pantaleo (2011). Automatic music transcription: from monophonic to polyphonic. In: J. Solis, K. Ng. Musical Robots and Interactive Multimodal Systems, pp. 27-46, Heidelberg: Springer. DOI: 10.1007/978-3-642-22291-7
- Pierfrancesco Bellini, Ivan Bruno, Daniele Cenni, Alice Fuzier, Paolo Nesi, Michela Paolucci (2011). Mobile Medicine: semantic computing management for health care applications on desktop and mobile devices. MULTIMEDIA TOOLS AND APPLICATIONS, vol. 53, pp. 1-39, ISSN:1380-7501 URL: <http://www.springerlink.com/content/q8512555u0j00584/fulltext.pdf>

Outline

- 1. Sistemi di Web Crawling
 - 1.1 Introduzione
 - 1.2 Strategie di Crawling
 - 1.3 Robot Exclusion Protocol
 - 1.4 Concorrenza
- 2. Tecniche di Parsing ed Estrazione di Informazioni
 - 2.1 Introduzione
 - 2.2 NLP: Cenni Storici
 - 2.3 Fasi dell'Elaborazione in Linguaggio Naturale
 - 2.4 NLP Tools
- 3. Applicazioni ←



3 - Applicazioni: Motori di Ricerca Semantici



<https://www.google.it/intl/it/insidesearch/features/search/knowledge.html>



<https://searchengineland.com/library/bing/bing-satori>

3 - Applicazioni: Motori di Ricerca Semantici

Introduzione della tecnologia Google Knowledge Graph nelle ricerche su Google:

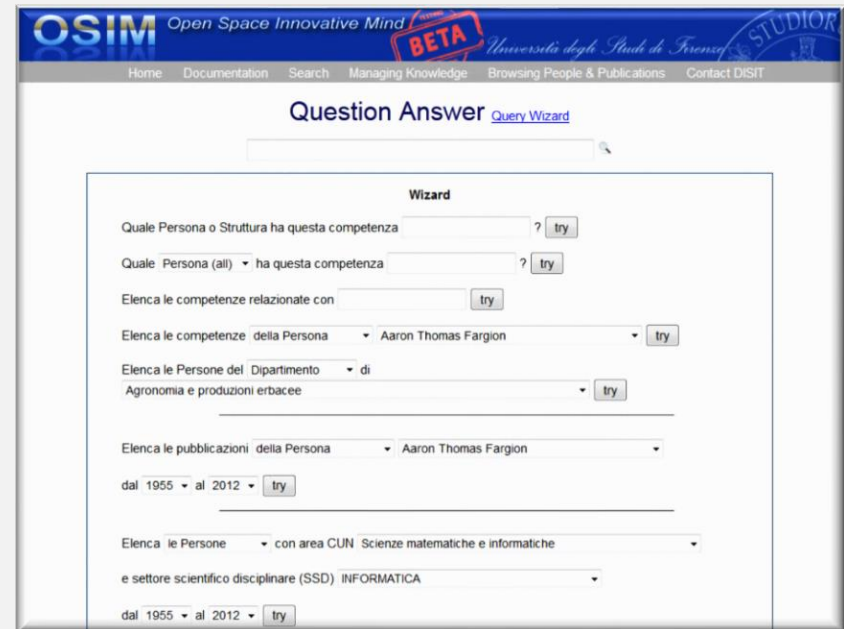
The image shows a Google search result for "bill gates". On the right side, a Knowledge Graph panel is highlighted with a red border. This panel provides a structured overview of Bill Gates, including his occupation as an entrepreneur, investor, and author, his net worth of 105.5 billion USD in 2019, his birth date of October 28, 1955, in Seattle, Washington, and his spouse Melinda Gates. It also lists his children and related books like "The Road Ahead" and "Le travail à la vitesse d...". At the bottom, it suggests related searches for other tech figures like Jeff Bezos, Steve Jobs, Mark Zuckerberg, Warren Buffett, and Melinda Gates.

3 - Applicazioni: OSIM (Open Space Innovative Mind)

- **Open Space Innovative Mind** (<http://openmind.disit.org>) [Bellandi et al., 2012] è un progetto per la realizzazione di un portale nel quale industrie, istituti di ricerca, ricercatori, studenti possono effettuare ricerche per individuare, all'interno dell'Ateneo Fiorentino:
 - le competenze possedute dai gruppi di ricerca, dai laboratori e dal personale universitario
 - le competenze offerte da corsi specifici nell'ambito dei vari corsi di laurea.
 - Le relazioni esistenti tra competenze diverse
 - Le pubblicazioni scientifiche
 - Le relazioni di conoscenza tra docenti
 - ...

Nell'architettura di OSIM sono implementate tutte le tecnologie presentate:

- ❖ **Data Mining**
- ❖ **Web Crawling**
- ❖ **NLP**
- ❖ **Semantic Web**
- ❖ **Interrogazione della conoscenza e Reasoning (OSIM Query Wizard)**



The screenshot shows the OSIM Query Wizard interface. At the top, there is a navigation bar with links for Home, Documentation, Search, Managing Knowledge, Browsing People & Publications, and Contact DISIT. The main heading is "Question Answer" with a "Query Wizard" link. Below this is a search input field. The wizard itself consists of several steps, each with a "try" button:

- Wizard**
- Quale Persona o Struttura ha questa competenza [input] ? try
- Quale Persona (all) ha questa competenza [input] ? try
- Elenca le competenze relazionate con [input] try
- Elenca le competenze della Persona [dropdown: Aaron Thomas Fargion] try
- Elenca le Persone del Dipartimento [dropdown: Agronomia e produzioni erbacee] di try
- Elenca le pubblicazioni della Persona [dropdown: Aaron Thomas Fargion] dal 1955 al 2012 try
- Elenca le Persone con area CUN [dropdown: Scienze matematiche e informatiche] e settore scientifico disciplinare (SSD) [dropdown: INFORMATICA] dal 1955 al 2012 try

3 - Applicazioni: OSIM (Open Space Innovative Mind)

I motori di ricerca attuali, ad esempio Google, sono keyword-based

Vengono restituiti i documenti che contengono esattamente le parole specificate dall'utente nella ricerca, senza tener conto della semantica di quanto espresso dall'utente stesso



Se per esempio un utente chiede una specifica competenza non è desiderabile avere come risposta tutte le pagine che contengono quella competenza, ma ci si aspetta di conoscere quali Professori o ricercatori hanno quella competenza, in quali corsi universitari può essere acquisita e quali gruppi di ricerca hanno maggior esperienze in quel settore.

La risposte trovate dovrebbero essere ordinate secondo una certa rilevanza



In riferimento, ad esempio, alla ricerca di chi possieda una specifica competenza, un docente/ricercatore può avere più rilevanza se possiede una competenza più specifica rispetto a quella cercata, di un docente/ricercatore che possiede una competenza più generale rispetto a quella cercata

3 - Applicazioni: OSIM (Open Space Innovative Mind)



ha cercato nell'area di INGEGNERIA

Cerca

Ricerca base

Forse cercavi: sistemi distribuiti

Dottorato in Informatica, Sistemi Distribuiti
... Il corso punterà a completare la formazione in progettazione ed analisi. ... Sicurezza
www.unifi.it/dist/cmpro-v-p-6.html - 23

Dottorato in Informatica, Sistemi Distribuiti
... Sono di interesse i sistemi di gestione software, i sistemi distribuiti, il software
www.unifi.it/dist/cmpro-v-p-16.html - 2

Dipartimento di Energetica - Laboratorio IBIS
... F., Morin F., Miglioramento delle prestazioni tramite sistemi di Manutenzione ... della catena logistica tramite simulazione
www.ibis.unifi.it/cmpro-l-s-4.html - 32k - 2009-05-19

Informatica e Applicazioni

Referente: Rosario Pugliese

Obiettivi:

Scopo del Dottorato è la formazione verso gli aspetti applicativi. Questo (che debbono fare i conti con frequenze soprattutto ad allargare la base culturale lavorative non collegate alla ricerca. Il corso punterà a completare la formazione

- Algoritmi per sistemi distribuiti
- Elaborazione delle immagini
- Metodi formali per la specificazione
- Progettazione di algoritmi di controllo
- Progettazione ed analisi di sistemi
- Sicurezza di sistemi distribuiti
- Strumenti formali per l'analisi
- Trattamento numerico e modelli

Visto l'esiguo numero di dottorandi interessati e che garantisce gli obiettivi

DIPARTIMENTO DI ENERGETICA - LABORATORIO IBIS

[home ateneo](#) | [home polo](#) | [home dipartimento](#) | [home laboratorio](#)

Lingua - Language



Menù

- ▶ Home
- ▶ Aree di ricerca
- ▶ Collaborazioni
- ▶ Progetti
- ▶ Prodotti
- ▶ Strumenti
- ▶ Persone
- ▶ Contatti
- ▶ Dove siamo

Utilità

- ▶ Mappa
- ▶ Statistiche
- ▶ Redazione

Aree di ricerca



Condition Monitoring & Condition Based Maintenance

Progettazione di sistemi per l'acquisizione dei dati di campo e per il monitoraggio remoto delle condizioni di impianti industriali, macchinari, flotte, ecc.), sviluppo sistemi e modelli per la manutenzione predittiva

Reliability analysis & Expert Systems

Modellazione e analisi di affidabilità e di disponibilità di sistemi complessi, sviluppo di modelli diagnostici (previsione vita utile residua del bene).

Service Management and Engineering

Analisi del valore, sviluppo service concept, ingegnerizzazione servizi, due diligence tecnologica, progettazione logistica di supporto.

Di seguito si illustrano le competenze acquisite in riferimento alle tematiche di cui al presente documento membri del comitato scientifico del laboratorio.

Primo risultato

Secondo risultato

ensibilità
elle industrie
oni si punta
posizioni

condii loro

3 - Applicazioni: OSIM (Open Space Innovative Mind)



OSIM Open Space Innovative Mind **BETA** Università degli Studi di Firenze

Home Documentation Search Managing Knowledge Browsing People & Publications Contact DISIT

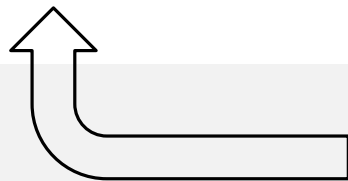
Question Answer [Query Wizard](#)

sistemi distribuiti

Results Displayed / Found: 1 - 3 / 3 in 21235 millisec

sistemi distribuiti (course)	Freqs: 1	score: 5.14	Paolo Nesi (full professor)	Freqs: 0	score: 4.91
sistemi distribuiti (skill)	Freqs: 0	score: 4.32			

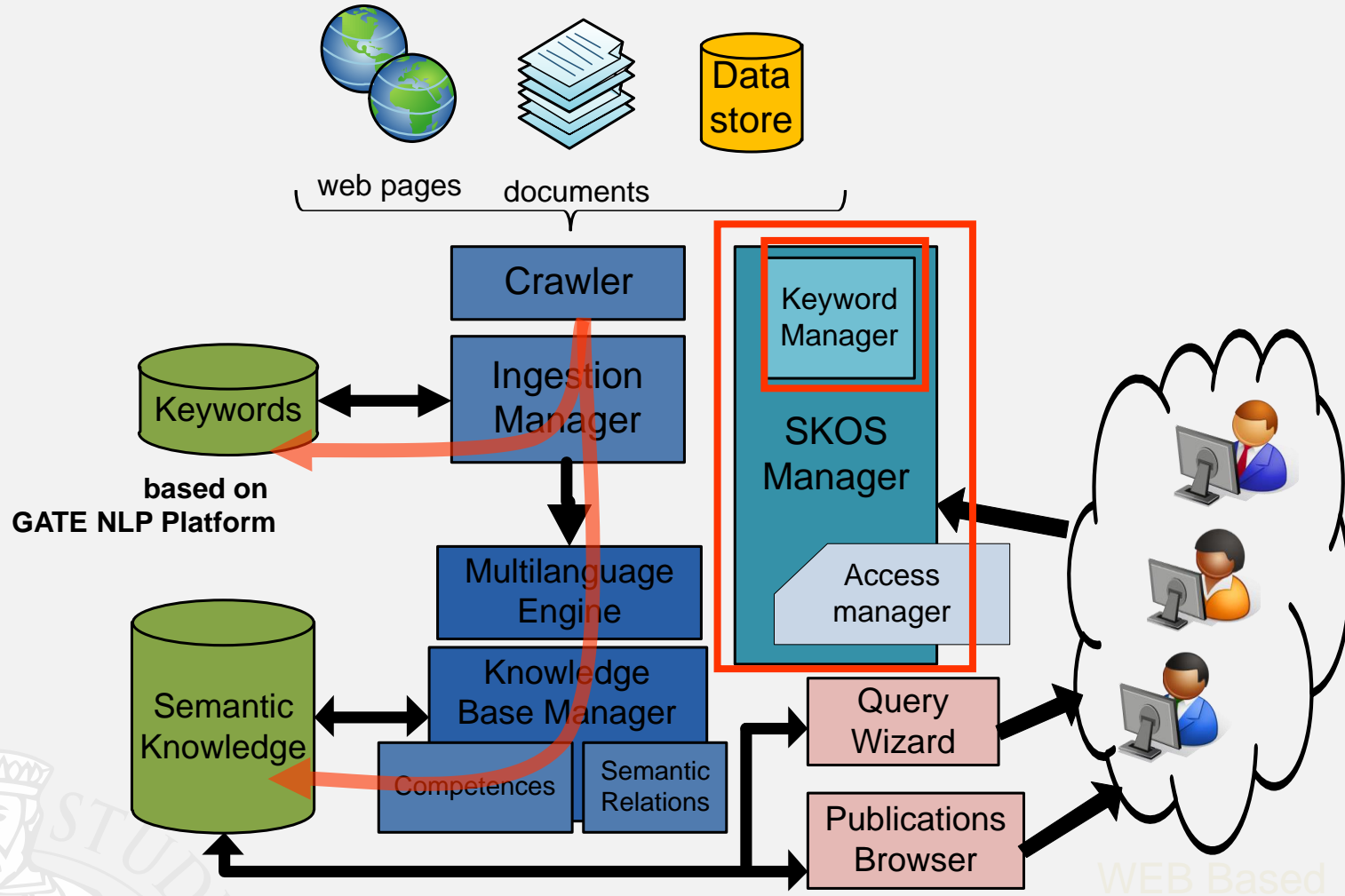
Previous 1 Next



I risultati non sono i documenti che contengono la parola "sistemi distribuiti"

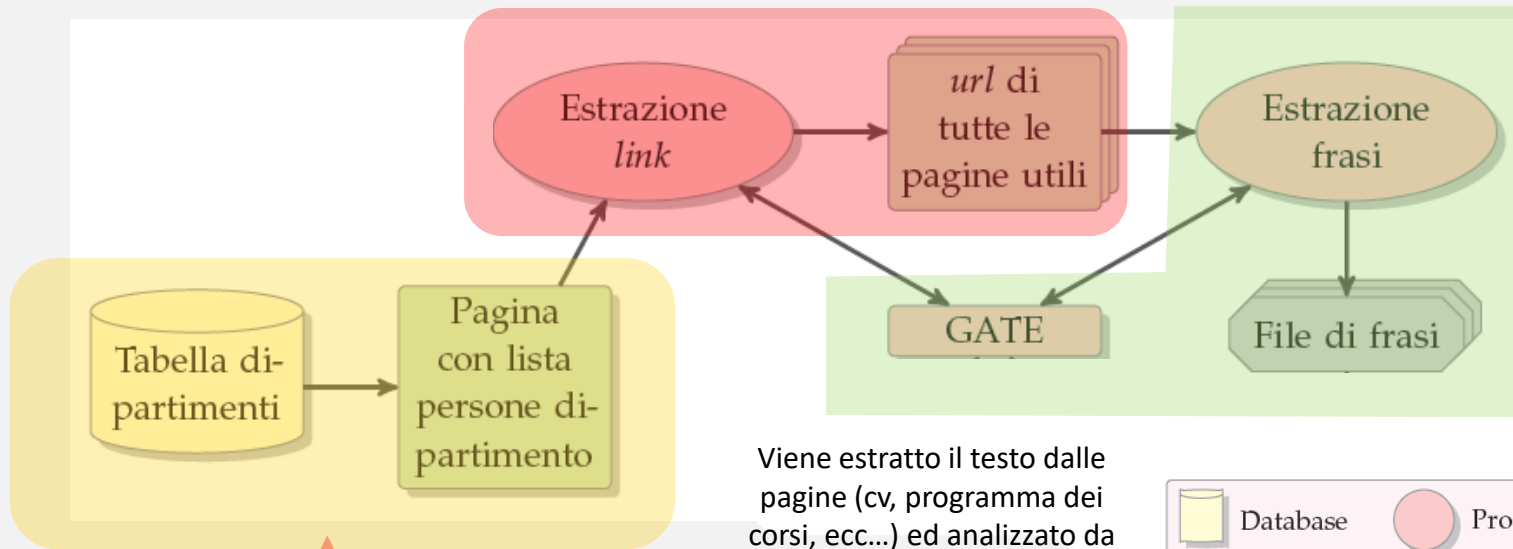
I risultati sono "il corso di Sistemi distribuiti" e la competenza "Sistemi distribuiti"

3 - Applicazioni: OSIM (Open Space Innovative Mind)



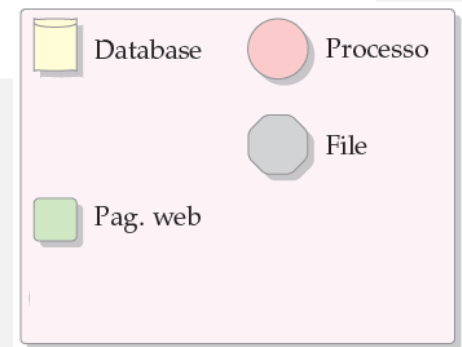
3 - Applicazioni: OSIM (Open Space Innovative Mind)

Vettore di URLs. Gli URLs sono scelti dalle apposite sezioni delle pagine, basandosi sui tags HTML delle pagine stesse




La seed list del crawler sono gli URLs delle pagine dipartimentali del Cerca Chi di UniFI

Viene estratto il testo dalle pagine (cv, programma dei corsi, ecc...) ed analizzato da GATE per il **Natural Language Processing**



Welcome root [Logout](#) [OSIM Managing Knowledge HOME](#)

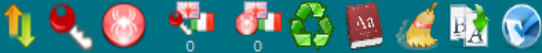
dipartimento di matematica per le decisioni - crawler is running

english 

ONTOLOGY MANAGER

KEYS SELECTION

RELATIONS MANAGER





all 20 / 113 (#2258)

id	value	translated values	occurencies	gazetteer	black list	no action	lang	Proposed
9390	algebra	algebra	9	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	en	0
1201	complementary	complementare	9	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	en	0
9139	changes	variazioni	9	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	en	0
9143	horizon	orizzonte	9	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	en	0
7611	decomposition	decomposizione	9	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	en	0
9148	infinite	infiniti	9	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	en	0
7365	laboratory	laboratorio	9	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	en	0
8646	angles	angoli	9	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	en	0
8397	embrechts	embrechts	9	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	en	0
9427	bond	obbligazione	9	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	en	0
9431	fields	campi	9	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	en	0
9689	edition	edizione	9	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	en	0
8673	min	min	9	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	en	0
482	decomposition	scomposizione	9	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	en	0
8426	year-old	anni	9	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	en	0

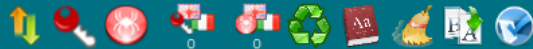
Trova: Maiuscole/minuscole



ONTOLOGY MANAGER

KEYS SELECTION

RELATIONS MANAGER



INSTANCES

filtered by black list

5



Concepts Repository

- 2
- A
 - able (18)
 - academic (25)
 - access (20)
 - access methods (5)
 - acm (21)
 - acm multimedia (9)
 - acquired (54)
 - acquired skills (48)
 - acquisition (6)
 - actions (8)
 - addresses (5)
 - addressing (5)
 - agreement (7)
 - allocation (26)
 - analyze (27)
 - and phase margin pulse crossing (5)
 - applications (105)
 - applied (6)

SKOS TREE

with frequencies



Concept Schema

- architectural (2)
- area of software engineering (1)
- artificial intelligence (2)
- automated control (0)
- computer science (0)
 - algorithm (95)
 - application (10)
 - code (8)
 - binary (4)
 - information (220)
 - notation (11)
 - xml (0)
 - database (0)
 - distributed systems (4)
 - life cycle (0)
 - programming (0)
- condition (0)
- e-commerce (0)
- e-learning (2)
- event (0)

LOG

1. skos tree node is re-loaded
2. skos tree node is re-loaded
3. [INFO]: LOOKUP FOR acquisition (6)
4. Related Subject:
5. http://www.unifi.it/off_form/insegnamenticc.php?cmd=2&cds=B070&cur=GEN&esa=B010480-FIRENZE&fac=200006<s=INGEGNERIA&AA=2009&codice=4480&bol=&coqnome=&nome=&f=s
6. http://www.unifi.it/off_form/insegnamenticc.php?cmd=2&cds=B070&cur=GEN&esa=B010480-FIRENZE&fac=200006<s=INGEGNERIA&AA=2010&codice=4480&bol=&coqnome=&nome=&f=s
7. Related Person:
8. Carlo Colombo (6)

CoSKOSAM

OSIM – Frammento Ontologia di dominio

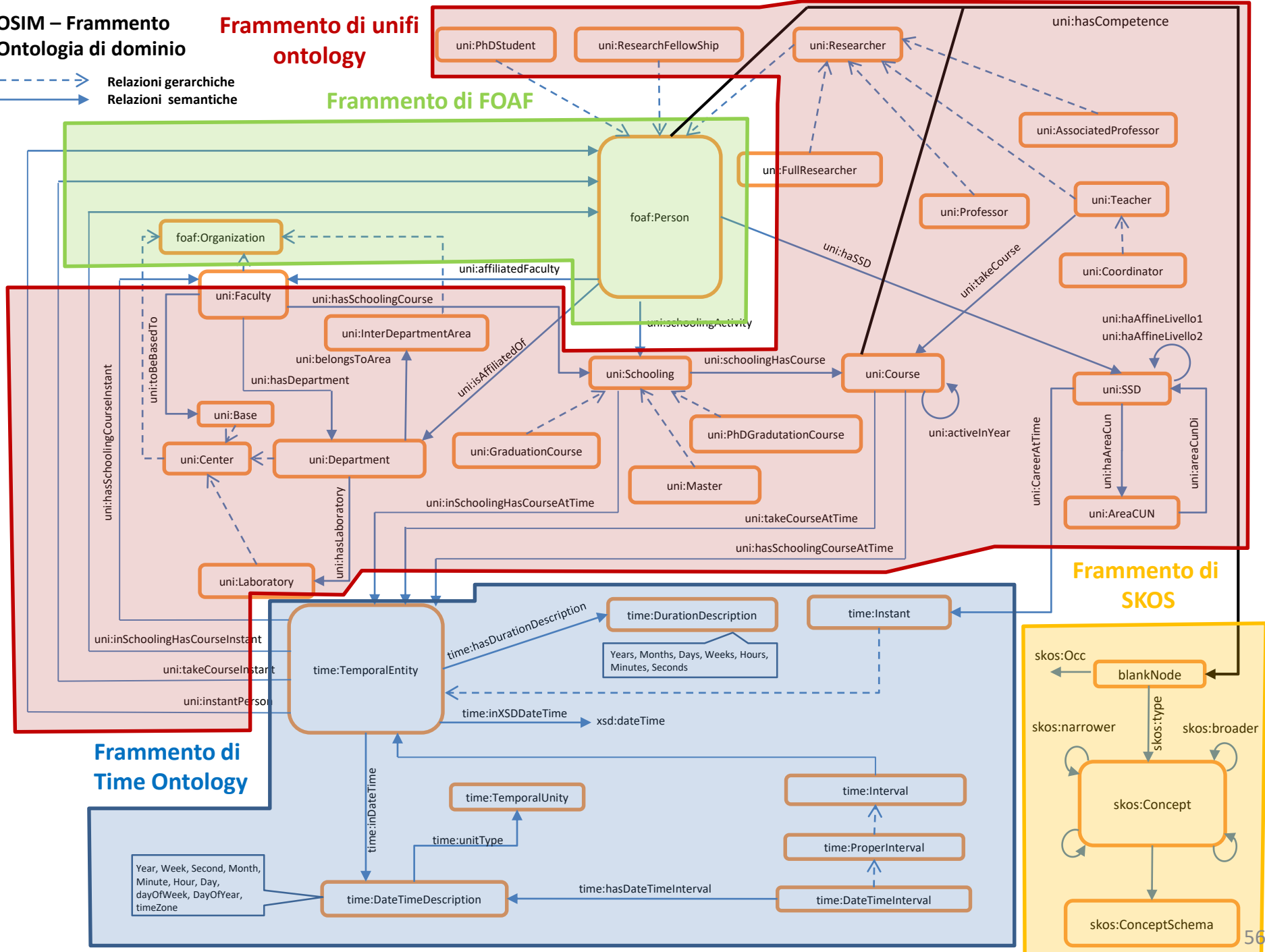
Frammento di unifi ontology

Frammento di FOAF

Frammento di SKOS

Frammento di Time Ontology

- - - - -> Relazioni gerarchiche
 - - - - -> Relazioni semantiche



3 - Applicazioni: OSIM (Open Space Innovative Mind)

OSIM Open Space Innovative Mind **BETA** Università degli Studi di Firenze

Home Documentation Search Managing Knowledge Browsing People & Publications Contact DISIT

italiano

Question Answer **Query Wizard**

Ricerca Full Text con Logica Fuzzy

Ricerca Assistita

Wizard

Quale Persona o Struttura ha questa competenza ?

Quale ha questa competenza ?

Elenca le competenze relazionate con

Elenca le competenze

Elenca le Persone del di

Elenca le Pubblicazioni

Dal Al

Elenca con Area CUN e settore scientifico disciplinare (SSD)

dal al

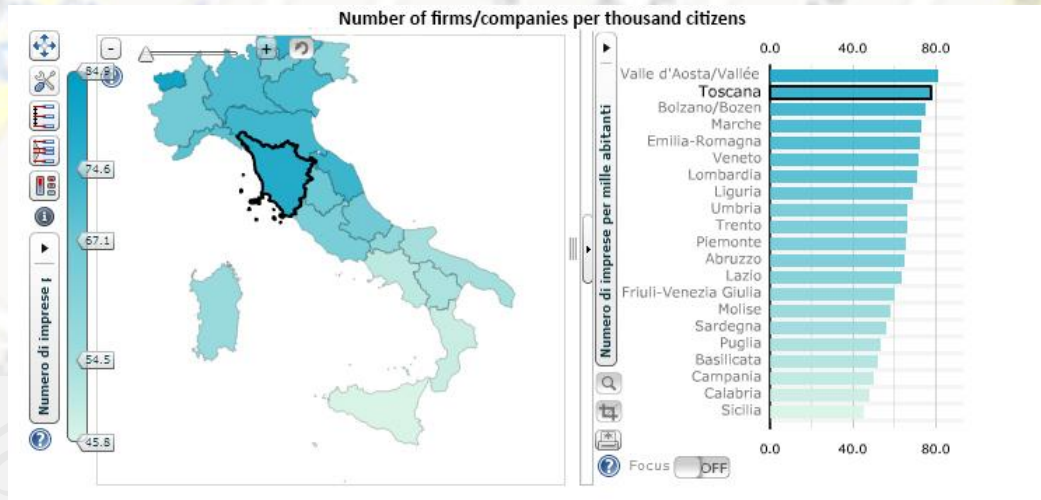
Quale persona o corso presenta aspetti legati a ?

Quali competenze legate all'attività di riguardano ?

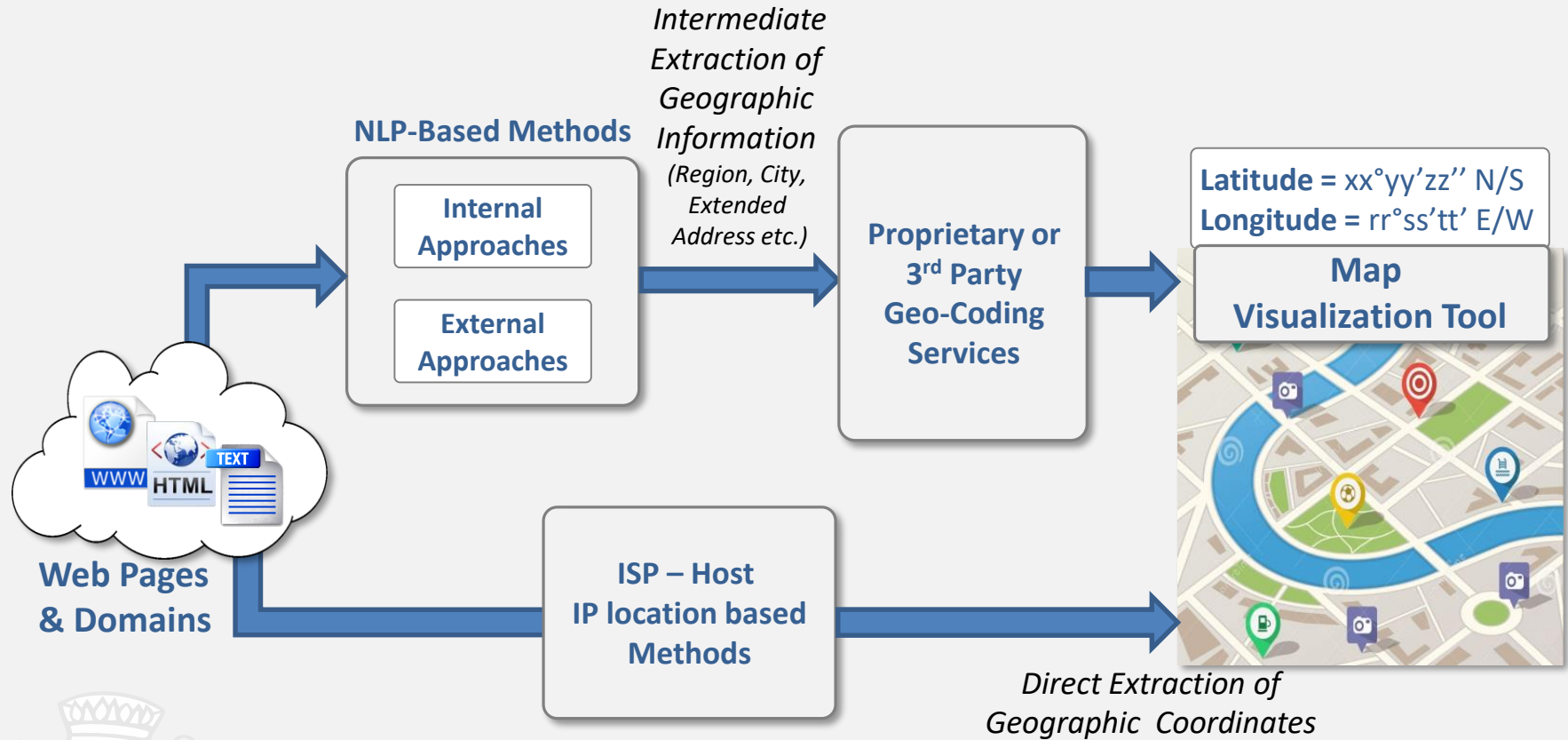
3 - Applicazioni: GeoLocator

- La Geolocalizzazione puntuale di un numero di servizi sempre crescent (di cui è possibile reperire risorse attraverso il web o gli Open Data delle Pubbliche Amministrazioni) sta diventando un requisito sempre più importante.
- Molti servizi, sepcialmente in ambito Smart City, sono basati su dati non strutturati e Open Data, spesso in formati diversi e non standardizzati tra loro, e soprattutto non geolocalizzati.

In 2015, Only 30000 commercial operators appears to be active in the Tuscany region, according Public Administration Open Data and among GoodRelations Ontology users.



3 - Applicazioni: GeoLocator



3 - Applicazioni: GeoLocator

GeoLocator [Nesi et al., 2016]

Address Extractor

I tag HTML (in particolare, *footer* e *header*) sono usati generalmente per contenere e visualizzare informazioni amministrative e fisiche dell'azienda / organizzazione / servizio proprietario del dominio web.

Estrazione degli indirizzi tramite tecniche di NLP, in particolare Pattern Matching:

- *General (high level detail) address pattern:*

[REGION] + [PROVINCE] + [POSTAL_CODE] + [CITY].

- *Specific (low level detail) address pattern:*

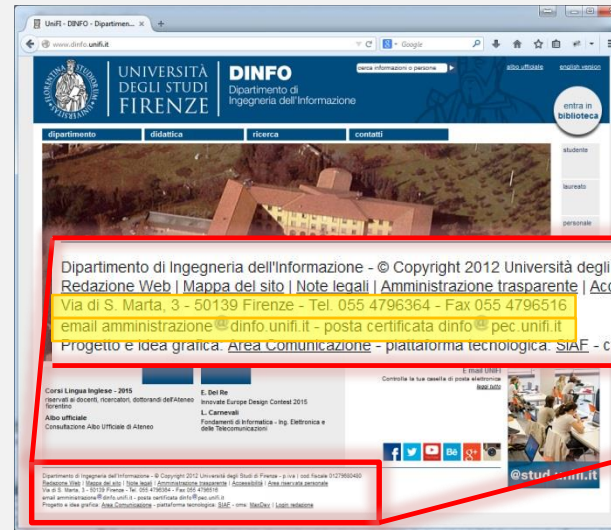
[STREET_IDENTIFIER] + [STREET_NAME] + [STREET_NUMBER].

- *Pattern per estrarre attributi speciali (e.g.: "Scala A, Interno 4"):*

[INNER_BLOCK_ID1] + [ID1_VALUE] + [INNER_BLOCK_ID2] + [ID2_VALUE].

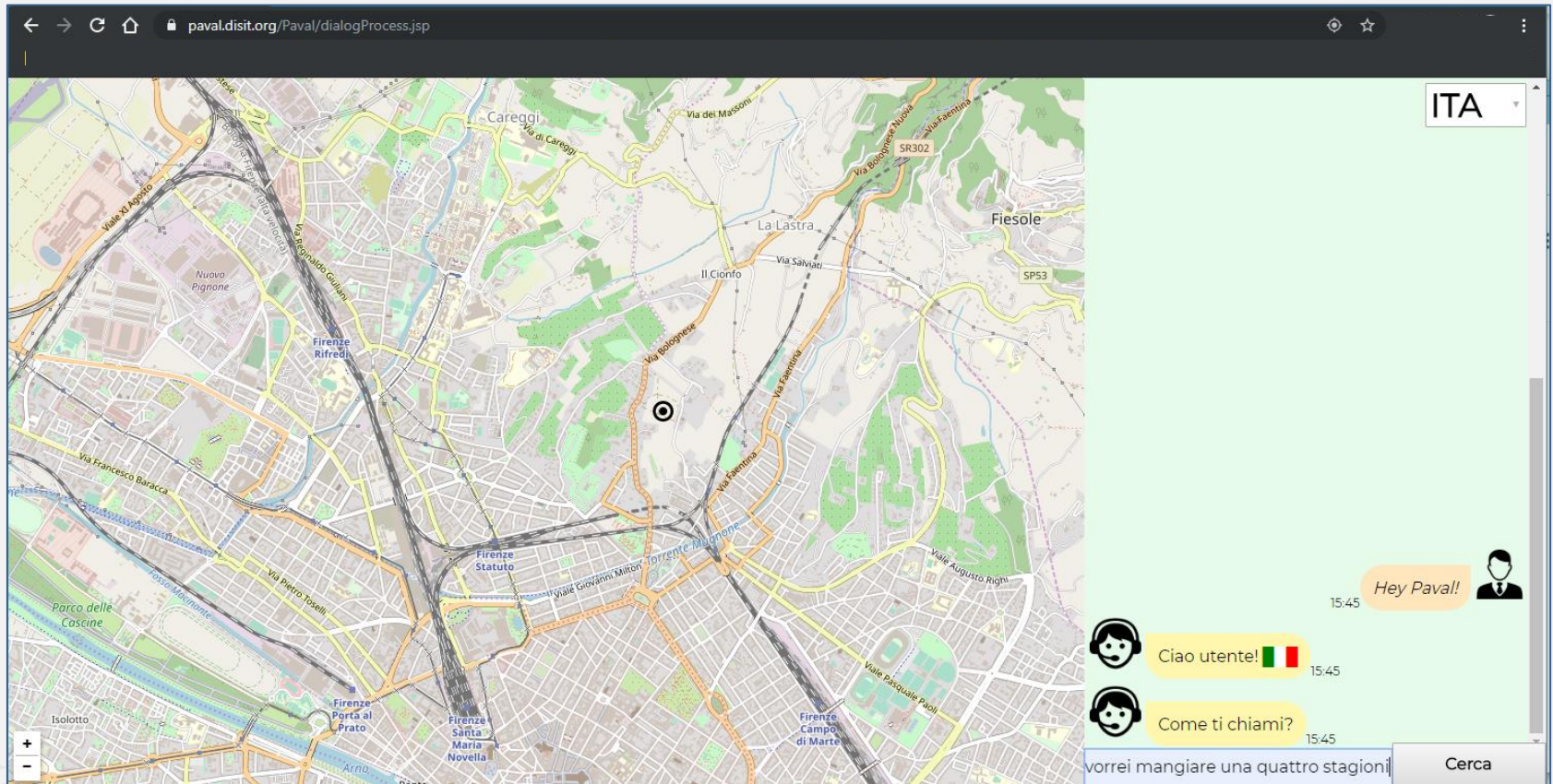
- *Pattern per estrarre coordinate geografiche (se presenti):*

[LATITUDE_IDENTIFIER] + [LATITUDE_VALUE] + [LONGITUDE_IDENTIFIER] + [LONGITUDE_VALUE].



3 - Applicazioni: Paval - A Location Aware Virtual Personal Assistant -

Paval [Massai et al., 2019] <https://paval.disit.org/>



3 - Applicazioni: Paval - A Location Aware Virtual Personal Assistant -

Paval [Massai et al., 2019] <https://paval.disit.org/>

The screenshot displays the Paval web application interface. The top navigation bar shows the URL `paval.disit.org/Paval/dialogProcess.jsp#bottom1569505532847`. The main area is split into two sections:

- Map:** A map of Florence, Italy, with several yellow speech bubble icons containing a pizza slice, indicating nearby pizzerias. The map includes labels for various streets and landmarks like "Via Lastra", "Fiesole", and "Via Salviati".
- Chat Window:** A vertical chat interface on the right side of the screen. It features a language selector set to "ITA". The chat history shows the following messages:
 - User: "Hey Paval!" (15:45)
 - Assistant: "Ciao utente! 🇮🇹" (15:45)
 - User: "Come ti chiami?" (15:45)
 - User: "vorrei mangiare una quattro stagioni" (15:47)
 - Assistant: "Ho trovato questi servizi che potrebbero interessarti: **Pizzeria**. I risultati sono mostrati sulla mappa nei dintorni della tua posizione." (15:47)

Below the map, there is a "Risultati" section with a plus sign icon. It lists the following results:

- Casa Della Pizza**
 - Pizzeria
- Le Follie**

At the bottom of the chat window, there is a text input field with the placeholder "Cosa vuoi fare? Dove?" and a "Cerca" button.

3 - Applicazioni: Paval - A Location Aware Virtual Personal Assistant -

Paval [Massai et al., 2019] <https://paval.disit.org/>

The screenshot displays the Paval application interface. On the left, a map of Florence, Italy, shows the current location and nearby streets. Two red location pins are visible on the map. Below the map, a search results section titled '+ Risultati' lists two nearby dental services:

- Pv Management Enterprise S.R.L.** → Dentista
- Oral Surgery S.A.S. Di Pasquale Paglianiti Societa' Tra**

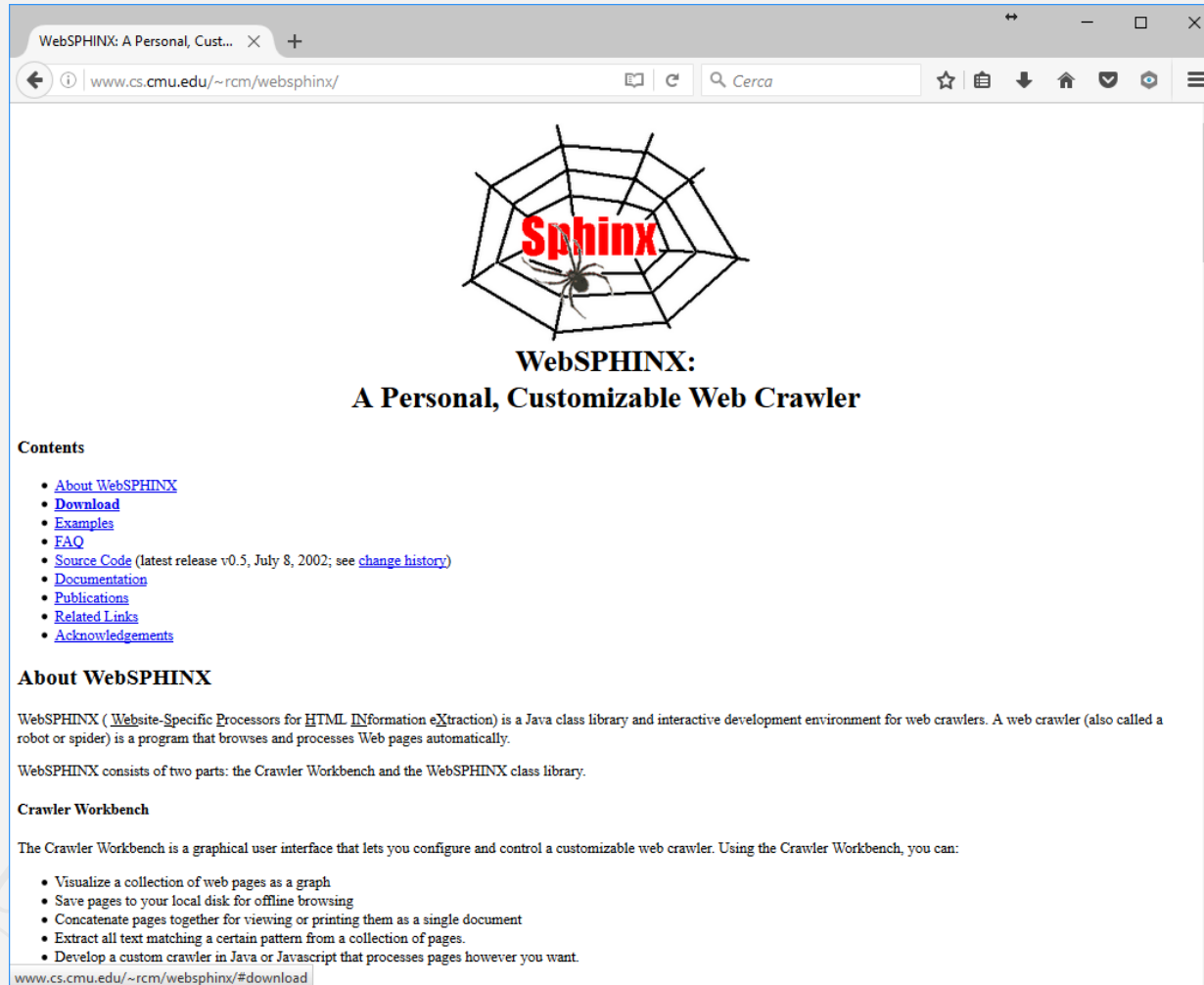
On the right side of the interface, a chat window is open, showing a conversation with the virtual assistant. The chat messages are:

- User: Hey Paval!
- Assistant: Ciao utente! 🇮🇹
- User: Come ti chiami?
- Assistant: ho le gengive arrossate
- User: Ho trovato questi servizi che potrebbero interessarti: **Dentista**. I risultati sono mostrati sulla mappa nei dintorni della tua posizione.

At the bottom of the chat window, there is a text input field with the placeholder text 'Cosa vuoi fare? Dove?' and a 'Cerca' button.


3 - Applicazioni: WebSPHINX

<http://www.cs.cmu.edu/~rcm/websphinx/>



WebSPHINX: A Personal, Cust... × +

www.cs.cmu.edu/~rcm/websphinx/ Cerca



WebSPHINX: A Personal, Customizable Web Crawler

Contents

- [About WebSPHINX](#)
- [Download](#)
- [Examples](#)
- [FAQ](#)
- [Source Code](#) (latest release v0.5, July 8, 2002; see [change history](#))
- [Documentation](#)
- [Publications](#)
- [Related Links](#)
- [Acknowledgements](#)

About WebSPHINX

WebSPHINX ([W](#)eb-site-Specific [P](#)rocessors for [H](#)TML [I](#)nformation [E](#)xtraction) is a Java class library and interactive development environment for web crawlers. A web crawler (also called a robot or spider) is a program that browses and processes Web pages automatically.

WebSPHINX consists of two parts: the Crawler Workbench and the WebSPHINX class library.

Crawler Workbench

The Crawler Workbench is a graphical user interface that lets you configure and control a customizable web crawler. Using the Crawler Workbench, you can:

- Visualize a collection of web pages as a graph
- Save pages to your local disk for offline browsing
- Concatenate pages together for viewing or printing them as a single document
- Extract all text matching a certain pattern from a collection of pages.
- Develop a custom crawler in Java or Javascript that processes pages however you want.

www.cs.cmu.edu/~rcm/websphinx/#download

3 - Applicazioni: WebSPHINX

Crawler Workbench: websphinx.Crawler

File

Crawl: the subtree

Starting URLs: <https://finanza.repubblica.it/Borsaitalia/Azioni/>

Action: none

- none
- save
- concatenate
- extract
- highlight
- script

Start Pause Stop Clear

Graph Outline Statistics Options... Tear Off

La borsa italiana dalla A alla Z

Titolo	Ultimo	Var.%	Ora	Apertura	Minimo	Massimo
A.S. Roma	0,47	+1,42%	16.20	0,46	0,46	0,47
A2A	1,52	+0,46%	16.22	1,52	1,50	1,53
Abitare In	339,00	+0,30%	16.01	---	330,60	340,00
Acea	14,45	+1,19%	16.21	14,30	14,25	14,59
Acotel Group	4,50	-9,57%	16.20	5,05	4,50	5,05
Acsm-Agam	2,32	+4,98%	16.20	2,23	2,22	2,33
Adidas	185,20	INV.	16.21	---	---	---
Advanced Micro Devices	9,96	INV.	16.21	---	---	---
Aedes	0,48	+4,23%	16.21	0,47	0,47	0,49
Aeffe	2,19	+2,15%	16.21	2,12	2,12	2,21
Aegon	5,09	INV.	16.21	---	---	---

MARKET OVERVIEW

MERCATI MATERIE PRIME TITOLI DI STATO

Descrizione	Ultimo	Var.%
DAX	13.037	+0,47%

<https://finanza.repubblica.it/Borsaitalia/Azioni/>

3 - Applicazioni: WebSPHINX

5 Identificare la/le risorsa/e informative da Estrarre nel codice HTML e individuare un pattern HTML significativo con cui addestrare il crawler

6 Selezionare la gestione delle opzioni "Advanced"

7a Inserire il pattern HTML scelto

8 Specificare il nome sel file in cui salvare il risultato dell'elaborazione

7b Per stampare nel file di uscita delle etichette per le descrivere risorse informative estratte:
`<td><a>{?company}</td>`
`<td>{?price}</td>`

4 Visualizzare il codice HTML della pagina:
F12 (Explorer, Firefox, Google Chrome)
 Ctrl + U (Safari, Opera)

3 - Applicazioni: WebSPHINX

Crawler Workbench: websphinx.Crawler

File

Crawl: Links Pages Classifiers Limits << Simple

Crawl: the subtree Using: all links

Starting URLs: https://finanza.repubblica.it/Borsaitalia/Azioni/

9 Impostare la profondità di navigazione (numero di livelli del grafo web)

10 Impostare la strategia di navigazione

Depth: 5 hops Breadth first

Start Pause Stop Clear

Graph Outline Statistics Options... Tear Off

La borsa italiana dalla A a Z

Economia con Bloomberg

LA BORSA ITALIANA DALLA A ALLA Z

Titolo	Ultimo	Var%	Ora	Apertura	Minimo	Massimo
A.S. Roma	0.47	+1,42%	16.20	0.46	0.46	0.47
A2A	1.52	+0,46%	16.22	1.52	1.50	1.53
Abitare In	339.00	+0,30%	16.01	---	330,60	340.00
Acea	14.45	+1,19%	16.21	14.30	14.25	14.59

Apple Holds on to Tax Billions as

Elements Console Sources Network Performance Memory Application Security

```

<tbody>
  <tr></tr>
  <tr>
    <td class="long-fixed-col text-bold">
      <a id="ct100_ContentLeft_RoundedTableListing_gridListino_ct102_linkDescription" rel="IT0001008876" href="/Company/?symbol=ASR:IM" target="parent">A.S. Roma/</a> == $0
    </td>
    <td>
      <span id="ct100_ContentLeft_RoundedTableListing_gridListino_ct102_lblLastPrice">0,47</span>
    </td>
  </tr>
</tbody>

```

Styles Computed Event Listeners

```

Filter :hov .cls +
element.style {
  a:have global_min.css?...:171020112348:1 r {
    color: #ff7600;
  }
  a { global_min.css?...:171020112348:1 ecolor: #000; transition: all .2s ease-out; }
}

```

3 - Applicazioni: WebSPHINX

Risultato in Output:

File

Crawl >> Advanced

Crawl: the subtree

Starting URLs: https://finanza.repubblica.it/Borsaitalia/Azioni/

Action: extract

regions matching the HTML tag expression:

```
<td><a>(?{company})</a></td><td><span>(?{price})</span></td>
```

as HTML to file: D:\crawler_output_01.html

Start Pause Stop Clear

Graph Outline Statistics Options... Tear Off



Gianni

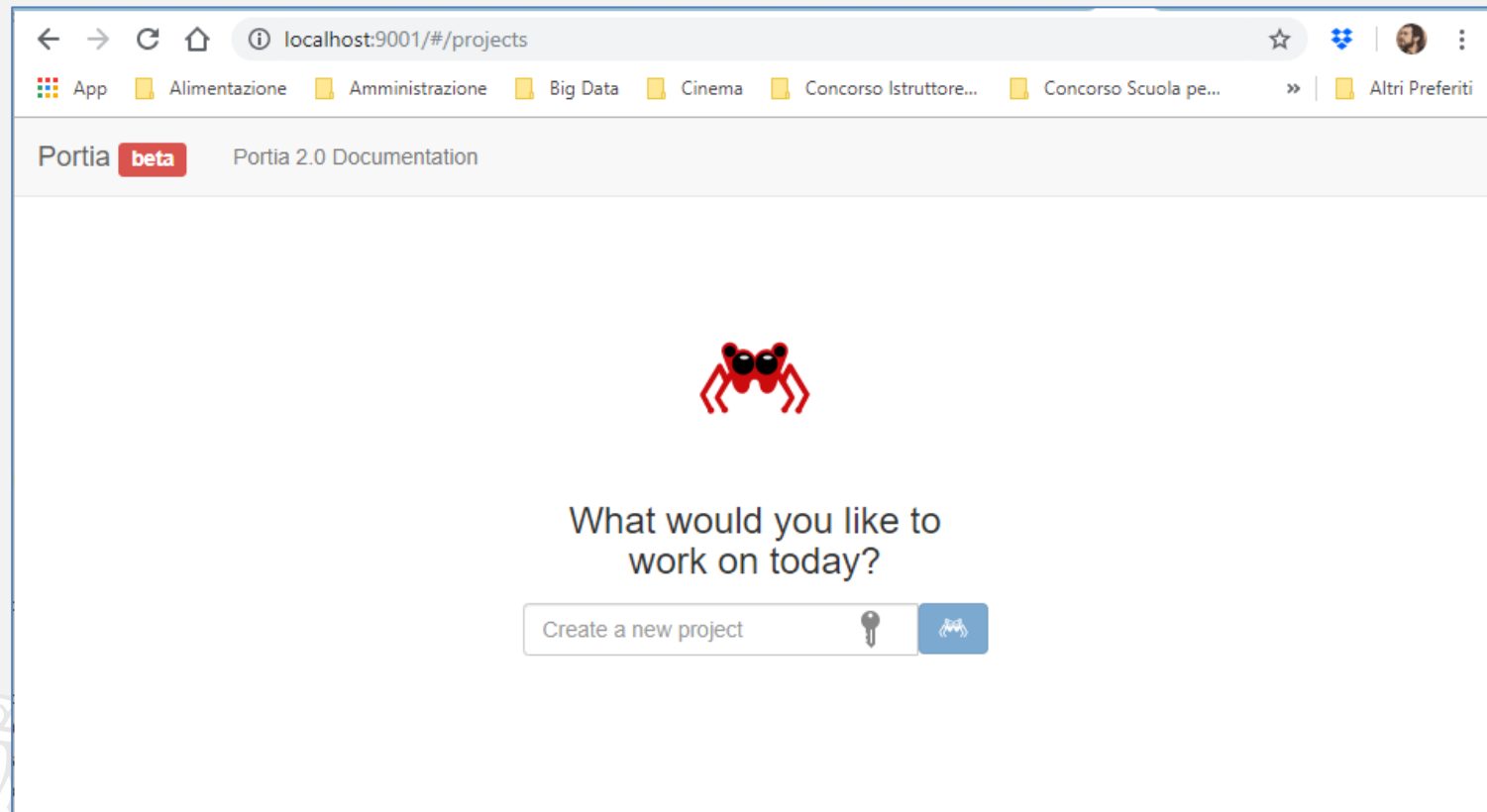
Extracted Records

file:///D:/crawler_output_01.html

	company	price
1.	A.S. Roma	0,46
2.	A2A	1,52
3.	Abitare In	339,00
4.	Acea	14,39
5.	Acotel Group	4,57
6.	Acsm-Agam	2,32
7.	Adidas	185,20
8.	Advanced Micro Devices	9,96
9.	Aedes	0,48
10.	Aeffe	2,20
11.	Aegon	5,09
12.	Aeroporto Guglielmo Marconi di Bologna	15,36
13.	Agatos	0,28
14.	Ageas	40,43
15.	Ahold Del	16,98
16.	Air Liquide	107,20
17.	Airbus	85,50
18.	Alerion	2,93
19.	Alfio Bardolla	7,40
20.	Allianz	196,80
21.	Alphabet Classe A	884,50
22.	Altaba	59,25
23.	Ambienthesis	0,39
24.	Ambromobiliare	3,65
25.	Amgen	150,00
26.	Amplifon	12,42
27.	Anheuser-Busch	97,55
28.	Anima Holding	5,59

3 - Applicazioni: Portia Web Scraper

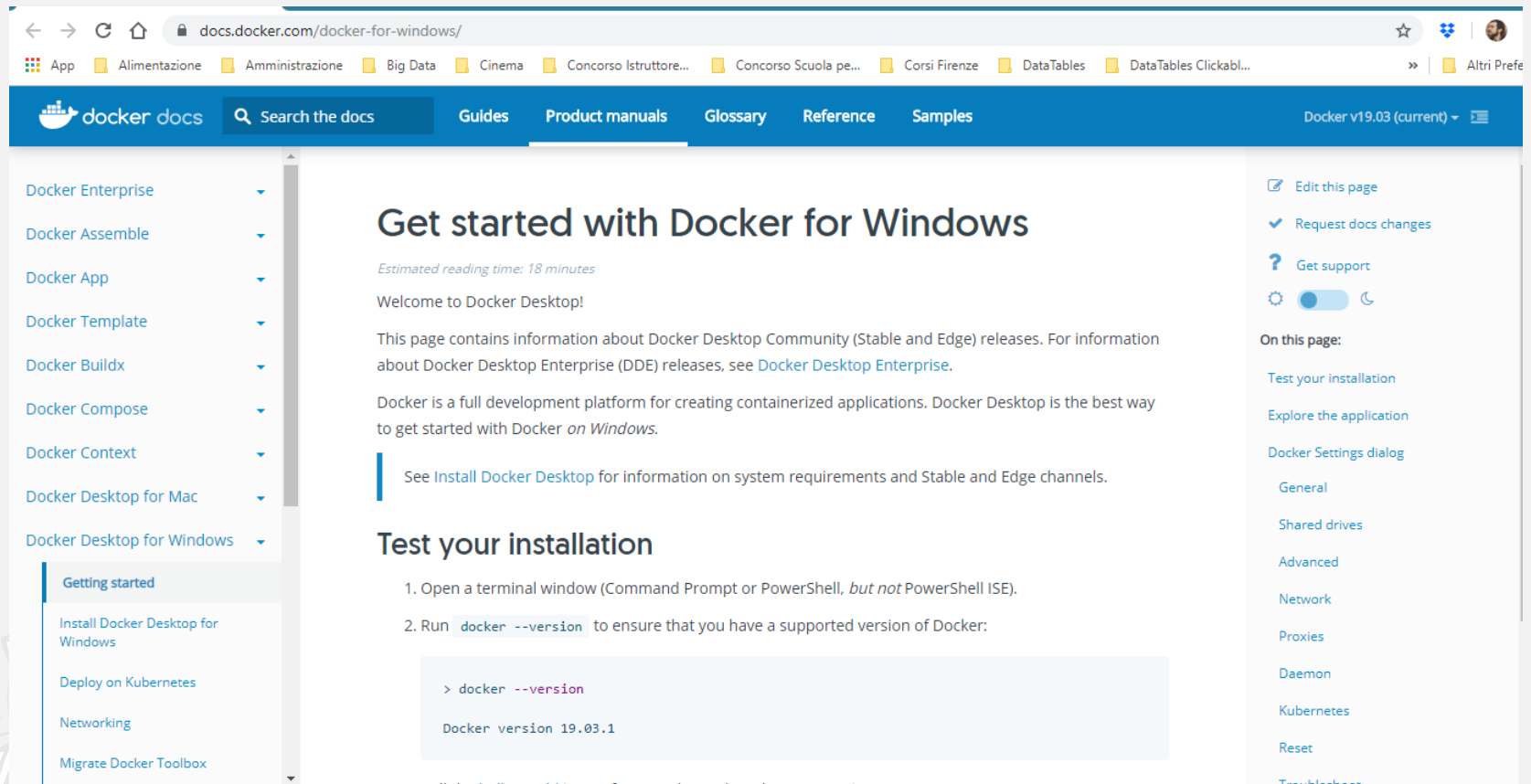
Strumento di crawling web orientato alla configurazione assistita tramite interfaccia grafica



3 - Applicazioni: Portia Web Scraper

Per installare e avviare Portia Web Scraper è necessario installare Docker:

<https://docs.docker.com/docker-for-windows/>



The screenshot shows the Docker documentation website for Windows. The main heading is "Get started with Docker for Windows". Below the heading, it says "Welcome to Docker Desktop!" and provides information about Docker Desktop Community and Enterprise releases. A section titled "Test your installation" lists two steps: 1. Open a terminal window (Command Prompt or PowerShell, but not PowerShell ISE). 2. Run `docker --version` to ensure that you have a supported version of Docker. A code block shows the command and its output: `> docker --version` resulting in `Docker version 19.03.1`. The left sidebar shows a navigation menu with "Getting started" selected. The right sidebar contains options like "Edit this page", "Request docs changes", "Get support", and "On this page" with a list of topics including "Test your installation", "Explore the application", "Docker Settings dialog", "General", "Shared drives", "Advanced", "Network", "Proxies", "Daemon", "Kubernetes", "Reset", and "Troubleshoot".

- Portia Web Scraper – Scrapinghub su Docker

The screenshot shows the Docker Quick Start page. On the left, there is a 'Quick Start' sidebar with a list of steps: 1. Download (highlighted), 2. Clone, 3. Build, 4. Run, and 5. Ship. The main content area features the Docker logo at the top, followed by the text: 'Let's get you started: you will need to download and install Docker to use the Docker command line interface (CLI)'. Below this text is an illustration of a laptop displaying a terminal window with code and a 'DOCKER IS RUNNING' notification. Further down, there is a 'Docker Desktop' section with a computer icon, the text 'The preferred choice for millions of developers that are building containerized applications', and a link 'Looking for Docker Engine Community?'. At the bottom, there are two buttons: a prominent blue button labeled 'Download Docker Desktop for Windows' and a smaller, lighter button labeled 'Docker Desktop for Mac'.

- Portia Web Scraper – Scrapinghub su Docker

Per avviare il programma su Docker:

- `docker run -i -t --rm -v <PROJECT_FOLDER>:/app/data/projects:rw -p 9001:9001 scrapinghub/portia`
- `docker run -i -t --rm -v d:/Docker:/app/data/projects:rw -p 9001:9001 scrapinghub/portia`

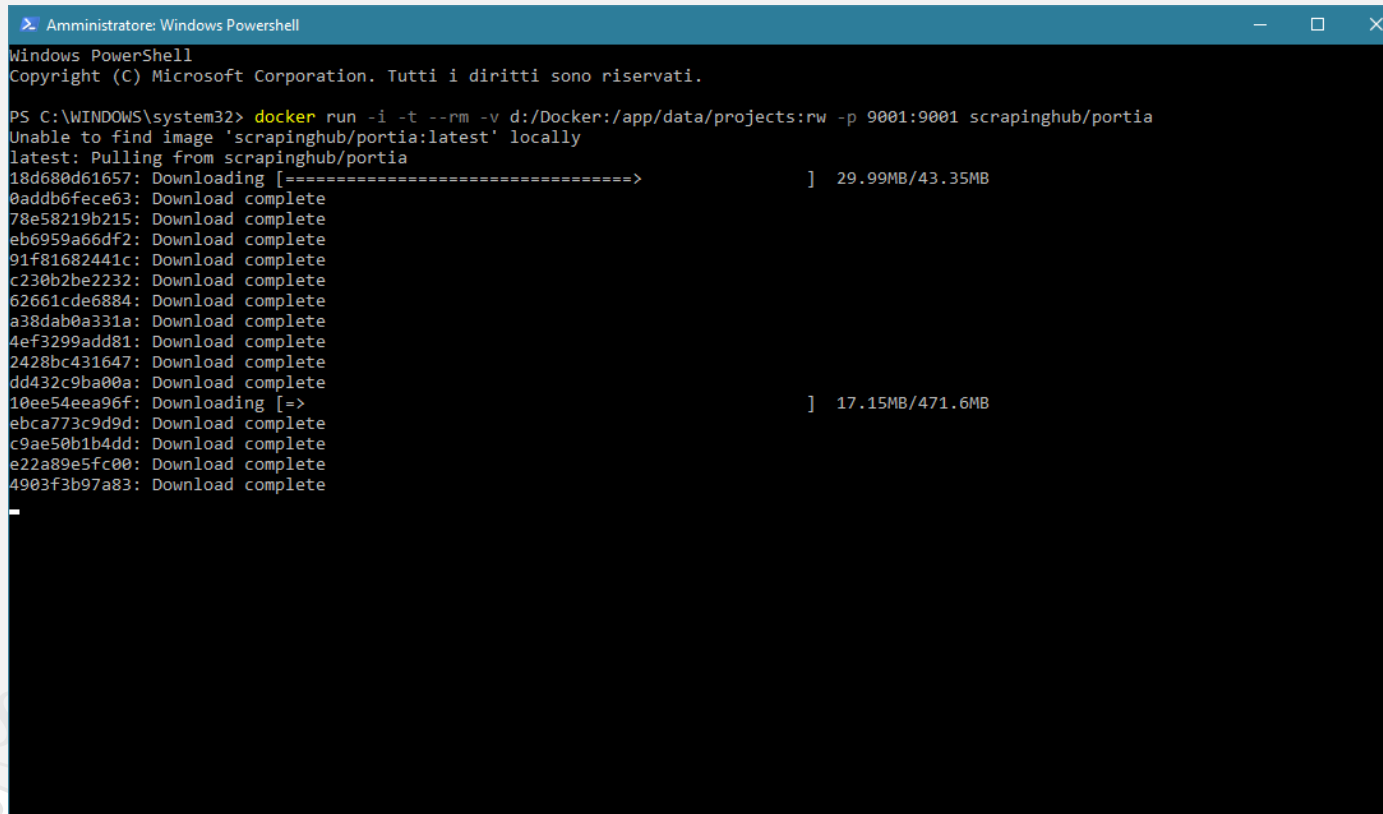
Per estrarre i dati dal sito web di interesse, una volta configurato lo spider:

- `docker run -i -t --rm -v <PROJECT_FOLDER>:/app/data/projects:rw -v <OUTPUT_FOLDER>:/mnt:rw -p 9002:9001 scrapinghub/portia portiacrawl /app/data/projects/<PROJECT_NAME> <SPIDER_NAME> -t <OUTPUT_FORMAT> -o /mnt/<SPIDER_NAME>.<FILE_EXT_OUT_FORMAT>`



- Portia Web Scraper – Scrapinghub su Docker

- `docker run -i -t --rm -v d:/Docker:/app/data/projects:rw -p 9001:9001 scrapinghub/portia`



```

Amministratore: Windows PowerShell
Windows PowerShell
Copyright (C) Microsoft Corporation. Tutti i diritti sono riservati.

PS C:\WINDOWS\system32> docker run -i -t --rm -v d:/Docker:/app/data/projects:rw -p 9001:9001 scrapinghub/portia
Unable to find image 'scrapinghub/portia:latest' locally
latest: Pulling from scrapinghub/portia
18d680d61657: Downloading [=====] 29.99MB/43.35MB
0addb6fece63: Download complete
78e58219b215: Download complete
eb6959a66df2: Download complete
91f81682441c: Download complete
c230b2be2232: Download complete
62661cde6884: Download complete
a38dab0a331a: Download complete
4ef3299add81: Download complete
2428bc431647: Download complete
dd432c9ba00a: Download complete
10ee54eea96f: Downloading [=] 17.15MB/471.6MB
ebca773c9d9d: Download complete
c9ae50b1b4dd: Download complete
e22a89e5fc00: Download complete
4903f3b97a83: Download complete
  
```

- Portia Web Scraper – Scrapinghub su Docker

```

Amministratore: Windows PowerShell
Windows PowerShell
Copyright (C) Microsoft Corporation. Tutti i diritti sono riservati.

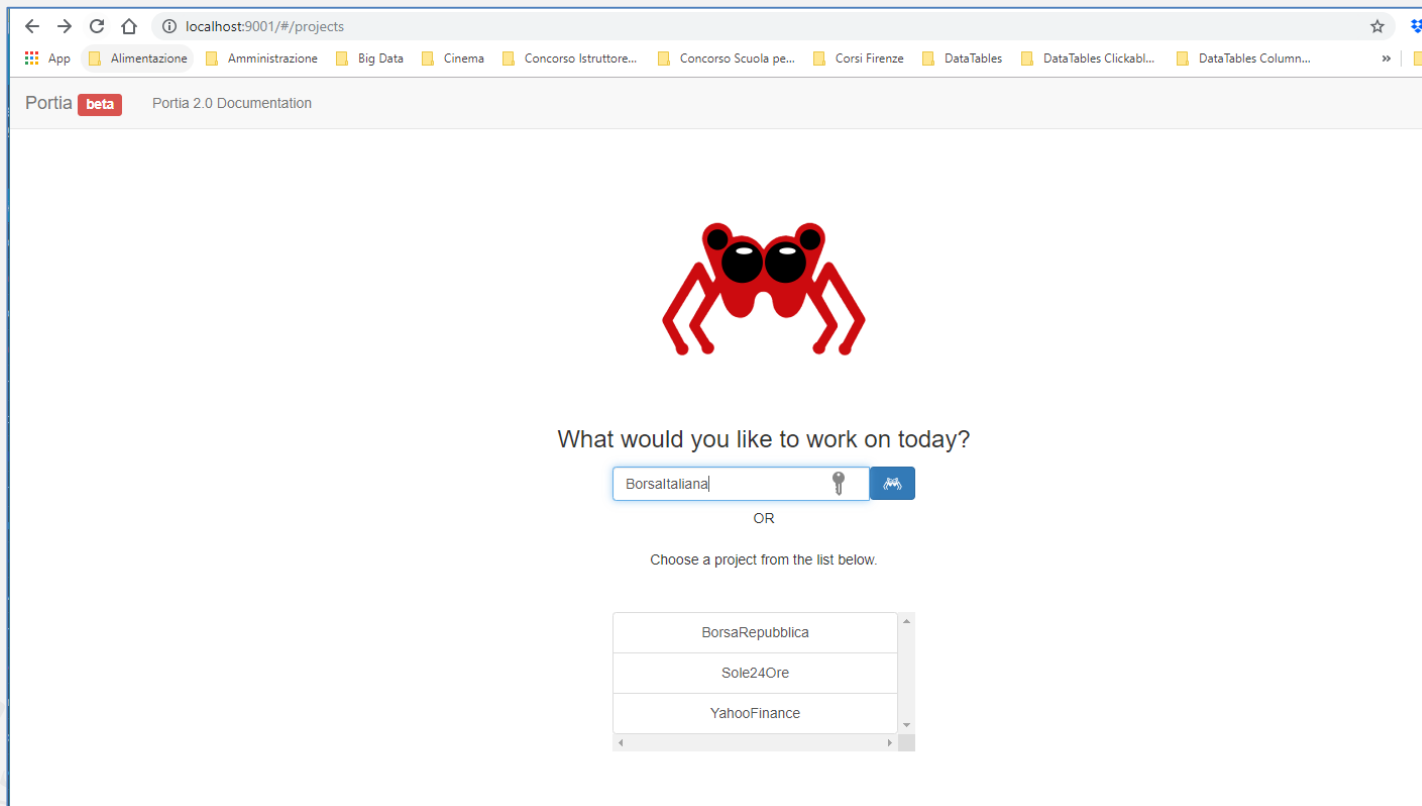
PS C:\WINDOWS\system32> docker run -i -t --rm -v d:/Docker:/app/data/projects:rw -p 9001:9001 scrapinghub/portia
Unable to find image 'scrapinghub/portia:latest' locally
latest: Pulling from scrapinghub/portia
18d680d61657: Pull complete
0addbfece63: Pull complete
78e58219b215: Pull complete
eb6959a66df2: Pull complete
91f81682441c: Pull complete
c230b2be2232: Pull complete
62661cde6884: Pull complete
a38dab0a331a: Pull complete
4ef3299add81: Pull complete
2428bc431647: Pull complete
dd432c9ba00a: Pull complete
10ee54eea96f: Pull complete
ebca773c9d9d: Pull complete
c9ae50b1b4dd: Pull complete
e22a89e5fc00: Pull complete
4903f3b97a83: Pull complete
Digest: sha256:a2acdf0c4a243deeea289f267feff06171c473c7ce1c0b992ae130b7cf486fa6
Status: Downloaded newer image for scrapinghub/portia:latest
+ action=
+ shift
+ '[' -z '' ']'
+ _run
+ service nginx start
+ _set_env
+ path=/app/portia_server:/app/slyd:/app/slybot
+ export PYTHONPATH=/app/portia_server:/app/slyd:/app/slybot
+ PYTHONPATH=/app/portia_server:/app/slyd:/app/slybot
+ echo /app/portia_server:/app/slyd:/app/slybot
/app/portia_server:/app/slyd:/app/slybot
+ /app/portia_server/manage.py runserver
+ /app/slyd/bin/slyd -p 9002 -r /app/portiaui/dist
2019-09-24 10:31:12+0000 [-] Log opened.
2019-09-24 10:31:12.396526 [-] Splash version: 3.2
2019-09-24 10:31:12.435039 [-] Qt 5.9.1, PyQt 5.9, WebKit 602.1, sip 4.19.3, Twisted 19.2.1, Lua 5.2
2019-09-24 10:31:12.435605 [-] Python 3.5.2 (default, Nov 12 2018, 13:43:14) [GCC 5.4.0 20160609]
2019-09-24 10:31:12.436183 [-] Open files limit: 1048576
2019-09-24 10:31:12.436613 [-] Can't bump open files limit
2019-09-24 10:31:12.561461 [-] Xvfb is started: ['Xvfb', ':979710311', '-screen', '0', '1024x768x24', '-nolisten', 'tcp']
QStandardPaths: XDG_RUNTIME_DIR not set, defaulting to '/tmp/runtime-root'
Watching for file changes with StatReloader
Performing system checks...

System check identified no issues (0 silenced).
September 24, 2019 - 10:31:15
Django version 2.2.3, using settings 'portia_server.settings'
Starting development server at http://127.0.0.1:8000/
Quit the server with CONTROL-C.
2019-09-24 10:31:15.556213 [-] Site starting on 9002
2019-09-24 10:31:15.556347 [-] Starting factory <slyd.server.Site object at 0x7f5de06f01d0>

```

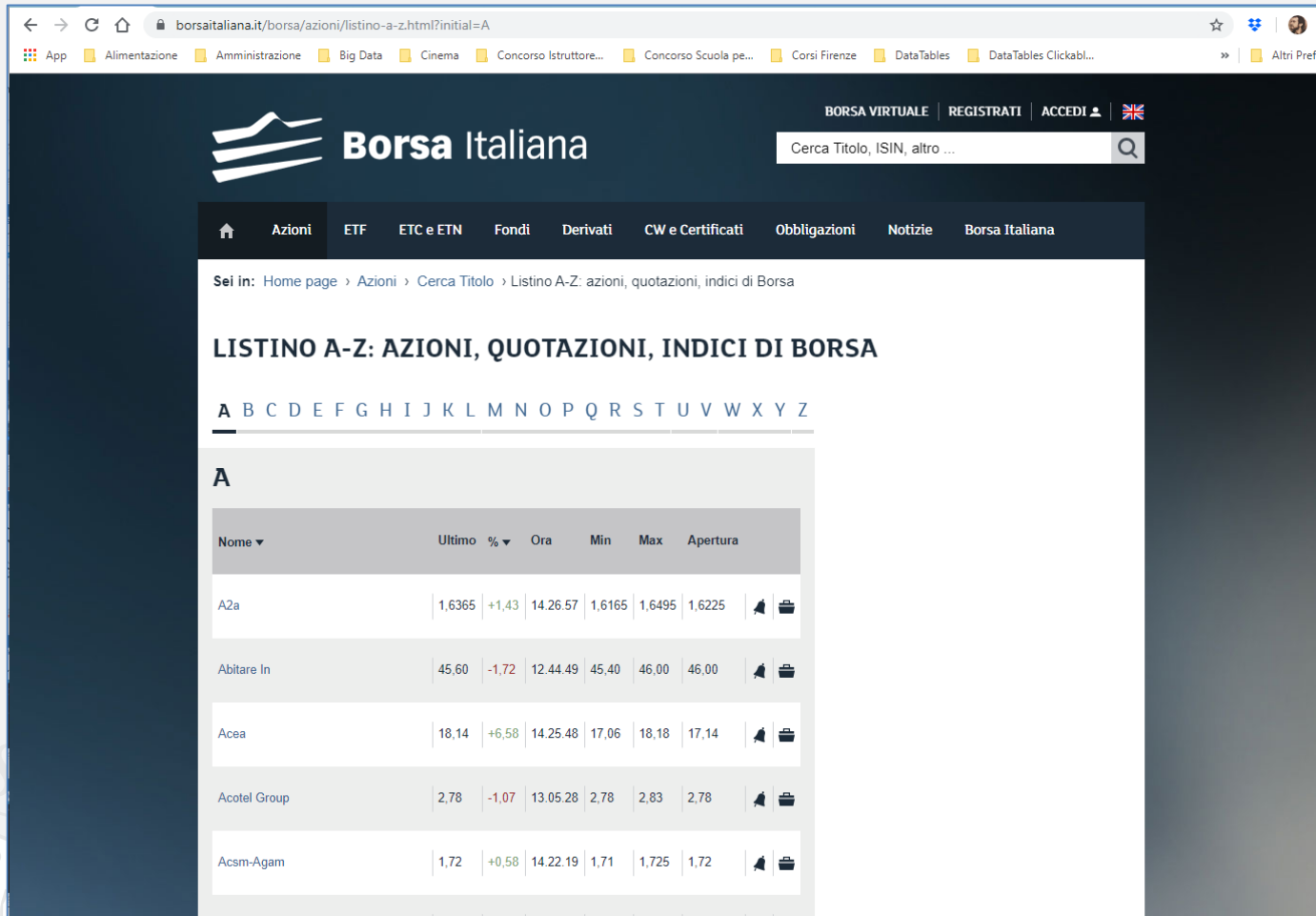
- Portia Web Scraper – Creare un progetto

- Portia sarà quindi accessibile localmente aprendo il browser web al seguente indirizzo: `http://localhost:9001`



- Portia Web Scraper – Creare un progetto

- Obiettivo: estrarre dati finanziari da pagine web



The screenshot shows the Borsa Italiana website interface. The main navigation bar includes 'Borsa Italiana' and a search bar. Below the navigation, there is a breadcrumb trail: 'Sei in: Home page > Azioni > Cerca Titolo > Listino A-Z: azioni, quotazioni, indici di Borsa'. The main heading is 'LISTINO A-Z: AZIONI, QUOTAZIONI, INDICI DI BORSA'. Below this, there is a navigation bar for letters A through Z, with 'A' selected. The main content area displays a table of stock data for the letter 'A'.

Nome ▼	Ultimo	% ▼	Ora	Min	Max	Apertura		
A2a	1,6365	+1,43	14.26.57	1,6165	1,6495	1,6225		
Abitare In	45,60	-1,72	12.44.49	45,40	46,00	46,00		
Acea	18,14	+6,58	14.25.48	17,06	18,18	17,14		
Acotel Group	2,78	-1,07	13.05.28	2,78	2,83	2,78		
Acsm-Agam	1,72	+0,58	14.22.19	1,71	1,725	1,72		

- Portia Web Scraper – Creare un progetto

- Verificare sempre eventuali disclaimer del sito sorgente dei dati!



The screenshot shows the Borsa Italiana website. At the top, there is a navigation bar with links for "BORSA VIRTUALE", "REGISTRATI", "ACCEDI", and a flag icon. Below this is a search bar with the text "Cerca Titolo, ISIN, altro ...". The main navigation menu includes "Azioni", "ETF", "ETC e ETN", "Fondi", "Derivati", "CW e Certificati", "Obbligazioni", "Notizie", and "Borsa Italiana". The current page is titled "DISCLAIMER" and contains the following text:

DISCLAIMER

Tutti i contenuti pubblicati su questo sito (www.borsaitaliana.it), compresi informazioni, testi, dati, software, fotografie, video, grafica, musica, loghi, icone, immagini, clip audio e qualsiasi altro materiale e servizi presenti su questo sito (il "Contenuto"), sono di esclusiva proprietà di Borsa Italiana S.p.A., e/o delle altre società facenti parte del gruppo London Stock Exchange (il "Gruppo"), e/o di terze parti in qualità di licenziatari, e sono protetti da copyright e altre leggi sulla proprietà intellettuale.

I Contenuti possono essere utilizzati, in tutto o in parte, solo per scopo personale (ad esempio, per motivi di studio e di ricerca) e non per finalità commerciali.

Solo per uso personale è consentito scaricare in modo occasionale, stampare, conservare temporaneamente i Contenuti su disco fisso (ma non su server o altro supporto di conservazione di dati connessi a un network). Non è consentito scaricare in modo sistematico e/o massivo i Contenuti e includere i Contenuti, in tutto o in parte, in elaborati o pubblicarli in qualsiasi forma.

I Contenuti, o parte di essi, possono essere riprodotti solo con il consenso preliminare ed esplicito di Borsa Italiana S.p.A. (il quale può essere richiesto mediante e-mail all'indirizzo info@borsaitaliana.it) e, se del caso, delle terze parti licenziatarie.

La copia, la riproduzione, la modifica, la distribuzione, la trasmissione, la ripubblicazione, lo scaricamento ("download") la diffusione, in qualsiasi modo e in tutto o in parte, a terzi, dei Contenuti per finalità commerciali sono severamente vietati.

Borsa Italiana S.p.A. si riserva il diritto di modificare in qualsiasi momento i Contenuti e le funzionalità di questo sito nonché dei servizi offerti. Né Borsa Italiana S.p.A. (i.e. i propri amministratori, revisori contabili, dirigenti, dipendenti, agenti e consulenti), né le altre società del Gruppo, né alcun licenziatario saranno responsabili per reclami o perdite di qualsiasi natura, derivanti direttamente o indirettamente (i) dall'uso dei dati o di materiale di questo sito o (ii) di accesso non autorizzato a questo sito o (iii) altrimenti derivanti in qualunque modo. Borsa Italiana S.p.A., le altre società del Gruppo e le terze parti licenziatarie non garantiscono né assicurano l'accuratezza, la completezza o la tempestività dei Contenuti presenti su questo sito web.

Per ragioni di traffico o sicurezza, Borsa Italiana si riserva il diritto di impedire temporaneamente l'accesso a questo sito anche qualora dovesse rilevare il download sistematico e/o massivo dei Contenuti.

On the right side of the page, there is a box titled "RIPRESE FOTOGRAFICHE E/O RIPRESE VIDEO" with the text: "Richiesta autorizzazione all'esecuzione di riprese fotografiche e/o di riprese video di Palazzo Mezzanotte (file pdf - 128 KB)".

- Portia Web Scraper – Configurare lo Spider

- Inserire l'URL di partenza;

The screenshot shows the Portia web scraper interface. On the left, there is a sidebar with the following sections:


- PROJECT**: Shows a folder named "Borsaitaliana" with a "Show all projects" link.
- SPIDERS**: Indicates "This project has no spiders" and provides a tip: "To create a spider first visit a web page that you would like to start crawling from."
- DATA FORMATS**: Indicates "This project has no data formats".

The main area of the interface features a large red spider icon and the text "What would you like to scrape?". Below this, a yellow input field contains the URL: `https://www.borsaitaliana.it/borsa/azioni/listino-a-z.html?initial=A`.



- Portia Web Scraper – Configurare lo Spider

- Portia offre una modalità di visualizzazione/navigazione della pagina corrispondente all'URL inserito.



Portia **beta** Portia 2.0 Documentation

Last saved 4 minutes ago

PROJECT [Show all projects](#)

- Borsaitaliana

SPIDERS [+](#)

This project has no spiders

To create a spider first visit a web page that you would like to start crawling from.

DATA FORMATS [+](#)

This project has no data formats

Browser address: <https://www.borsaitaliana.it/borsa/azioni/listino-a-z.html?initial=A>

Borsa Italiana

BORSA VIRTUALE | REGISTRATI | ACCEDI

Cerca Titolo, ISIN, altro ...

Azioni | ETF | ETC e ETN | Fondi | Derivati | CW e Certificati | Obbligazioni | Notizie | Borsa Italiana

Sei in: Home page > Azioni > Cerca Titolo > Listino A-Z: azioni, quotazioni, indici di Borsa

LISTINO A-Z: AZIONI, QUOTAZIONI, INDICI DI BORSA

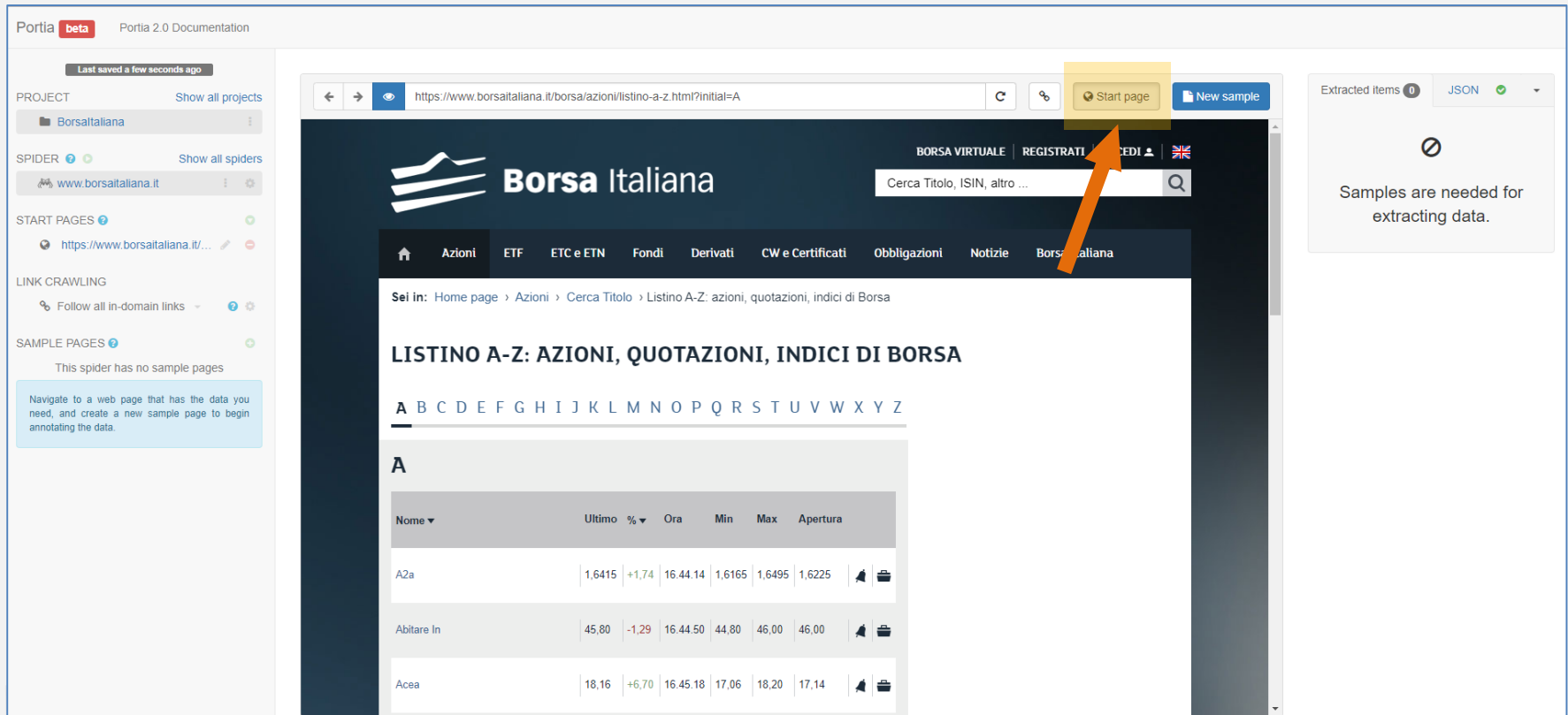
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

A

Nome ▼	Ultimo	% ▼	Ora	Min	Max	Apertura		
A2a	1,6415	+1,74	16.44.14	1,6165	1,6495	1,6225	🔔	📁
Abitare In	45,80	-1,29	16.44.50	44,80	46,00	46,00	🔔	📁
Acea	18,16	+6,70	16.45.18	17,06	18,20	17,14	🔔	📁

- Portia Web Scraper – Configurare lo Spider

- E' ora possibile associare e inizializzare un nuovo Spider e associare la pagina inserita come pagina iniziale (Start Page).



Portia **beta** Portia 2.0 Documentation

Last saved a few seconds ago

PROJECT Show all projects
Borsaitaliana

SPIDER Show all spiders
www.borsaitaliana.it

START PAGES
https://www.borsaitaliana.it/...

LINK CRAWLING
Follow all in-domain links

SAMPLE PAGES
This spider has no sample pages

Navigate to a web page that has the data you need, and create a new sample page to begin annotating the data.

Extracted Items 0 JSON

Samples are needed for extracting data.

Start page

New sample

Borsa Italiana

BORSA VIRTUALE REGISTRATI CEDI

Cerca Titolo, ISIN, altro ...

Azioni ETF ETC e ETN Fondi Derivati CW e Certificati Obbligazioni Notizie Borsa Italiana

Sei in: Home page > Azioni > Cerca Titolo > Listino A-Z: azioni, quotazioni, indici di Borsa

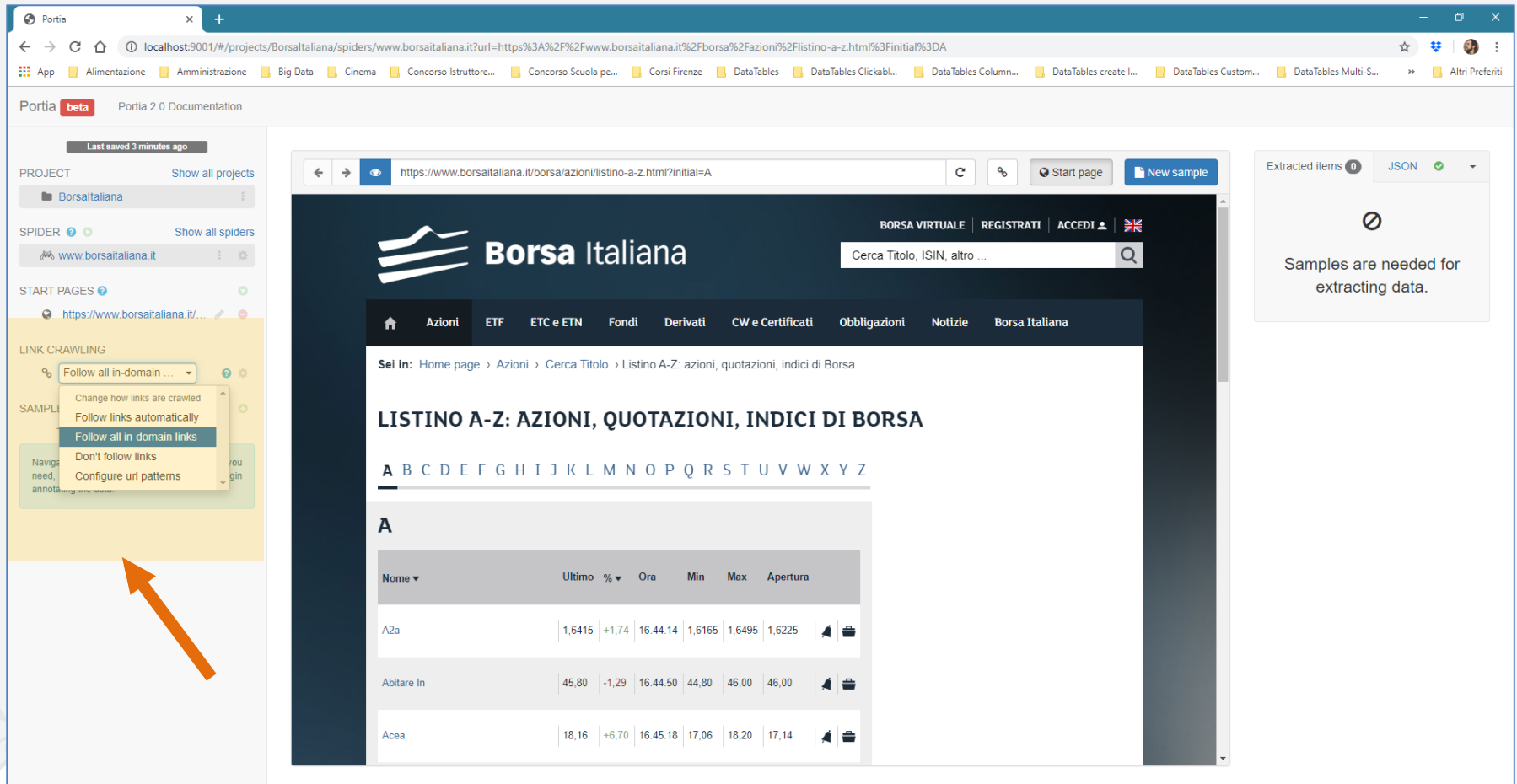
LISTINO A-Z: AZIONI, QUOTAZIONI, INDICI DI BORSA

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Nome	Ultimo	%	Ora	Min	Max	Apertura
A2a	1,6415	+1,74	16.44.14	1,6165	1,6495	1,6225
Abitare In	45,80	-1,29	16.44.50	44,80	46,00	46,00
Acea	18,16	+6,70	16.45.18	17,06	18,20	17,14

- Portia Web Scraper – Configurare lo Spider

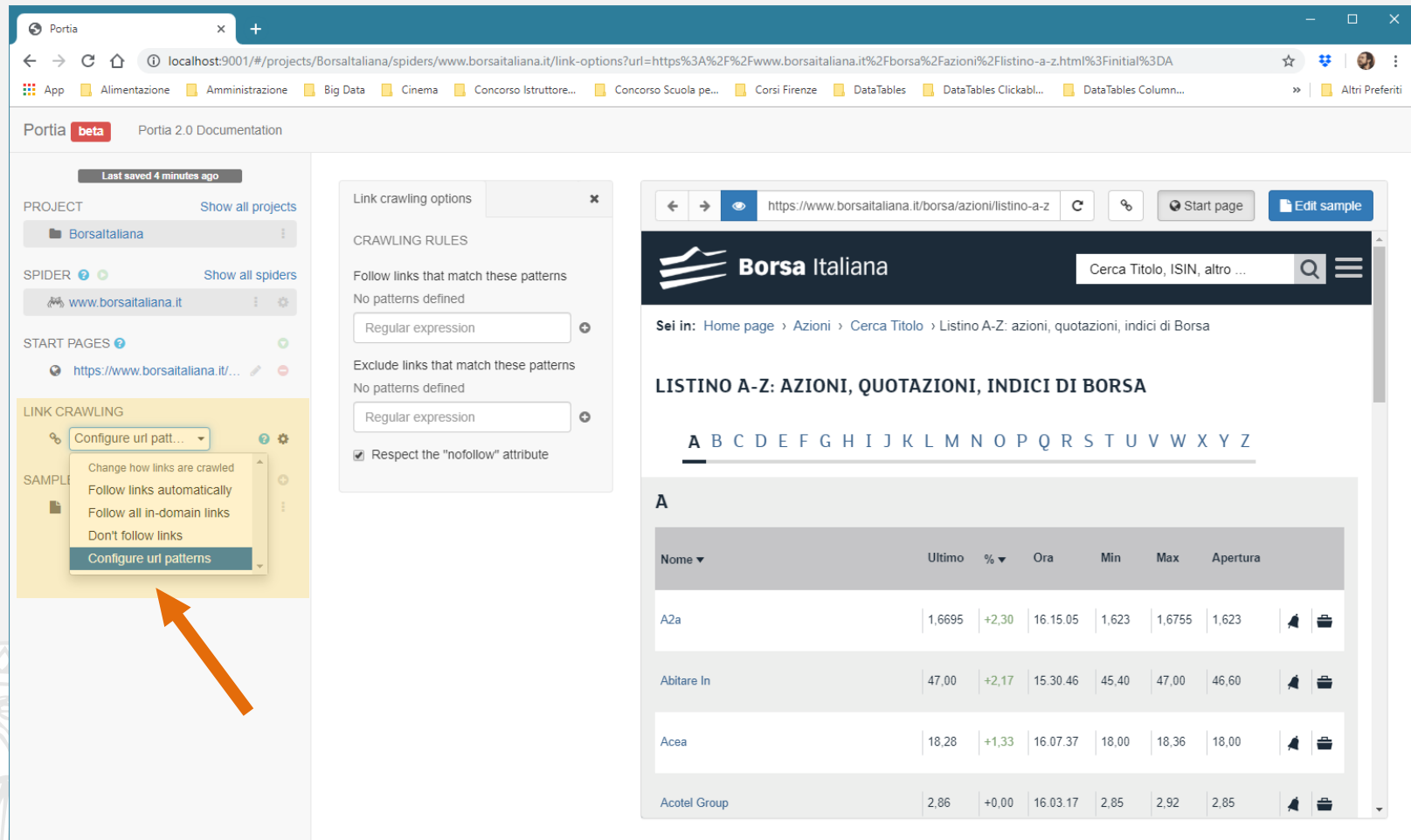
- Si possono definire varie impostazioni, tra cui la politica di navigazione delle pagine:



The screenshot shows the Portia web scraper interface. The central preview window displays the Borsa Italiana website with the URL `https://www.borsaitaliana.it/borsa/azioni/listino-a-z.html?initial=A`. The sidebar on the left contains configuration options for the spider, including 'LINK CRAWLING' settings. An orange arrow points to the 'LINK CRAWLING' section, which is currently set to 'Follow all in-domain links'. Other options include 'Change how links are crawled', 'Follow links automatically', 'Don't follow links', and 'Configure url patterns'. The right-hand panel shows 'Extracted items' in JSON format, with a message indicating that samples are needed for extracting data.

- Portia Web Scraper – Configurare lo Spider

- Si possono definire varie impostazioni, tra cui la politica di navigazione delle pagine:



The screenshot shows the Portia web scraper interface. On the left, the 'LINK CRAWLING' section is highlighted with a yellow background. A dropdown menu is open, showing options: 'Change how links are crawled', 'Follow links automatically', 'Follow all in-domain links', 'Don't follow links', and 'Configure url patterns'. An orange arrow points to the 'Configure url patterns' option.

In the center, the 'Link crawling options' panel is visible, showing 'CRAWLING RULES' with sections for 'Follow links that match these patterns' and 'Exclude links that match these patterns', both currently set to 'No patterns defined'. A checkbox for 'Respect the "nofollow" attribute' is checked.

On the right, a sample of the crawled page is shown. The browser address bar displays 'https://www.borsaitaliana.it/borsa/azioni/listino-a-z'. The page content includes the 'Borsa Italiana' logo, a search bar, and a section titled 'LISTINO A-Z: AZIONI, QUOTAZIONI, INDICI DI BORSA'. Below this is a navigation bar with letters A through Z. The main content area shows a table of stock data for the letter 'A'.

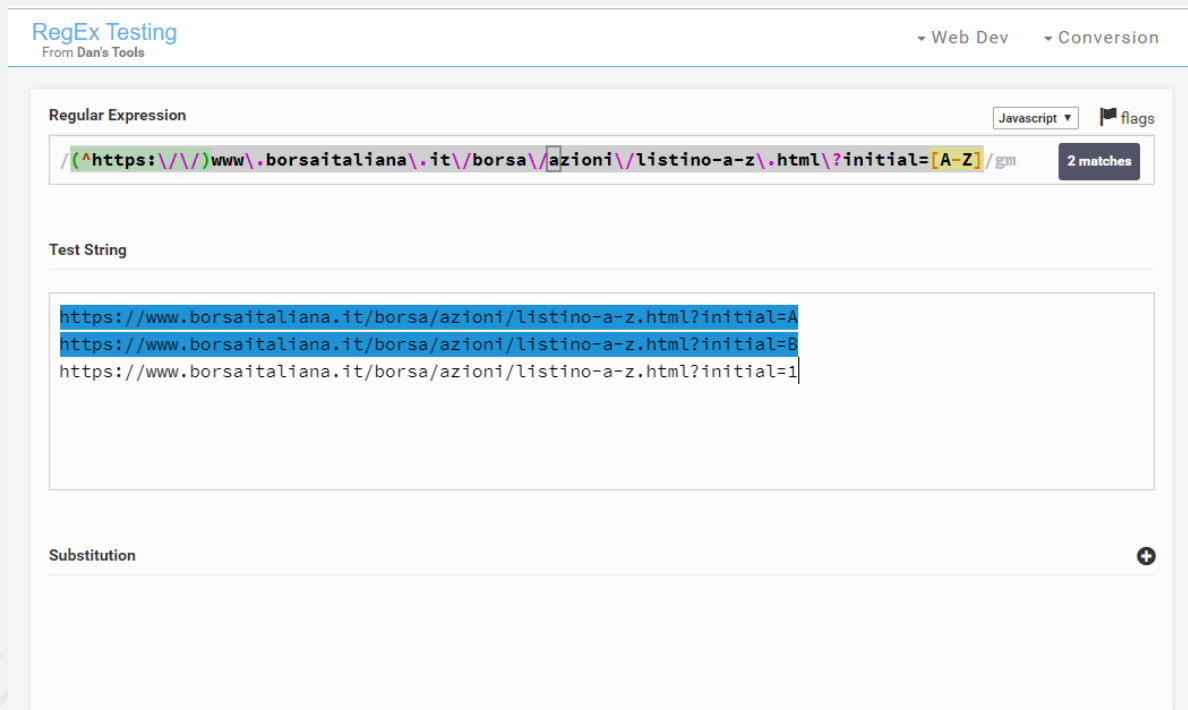
Nome	Ultimo	%	Ora	Min	Max	Apertura
A2a	1,6695	+2,30	16.15.05	1,623	1,6755	1,623
Abitare In	47,00	+2,17	15.30.46	45,40	47,00	46,60
Acea	18,28	+1,33	16.07.37	18,00	18,36	18,00
Acotel Group	2,86	+0,00	16.03.17	2,85	2,92	2,85

- Portia Web Scraper – Configurare lo Spider

- Molto comodo l'utilizzo delle **Espressioni Regolari** per identificare pattern di URLs:

<https://www.borsaitaliana.it/borsa/azioni/listino-a-z.html?initial=A>

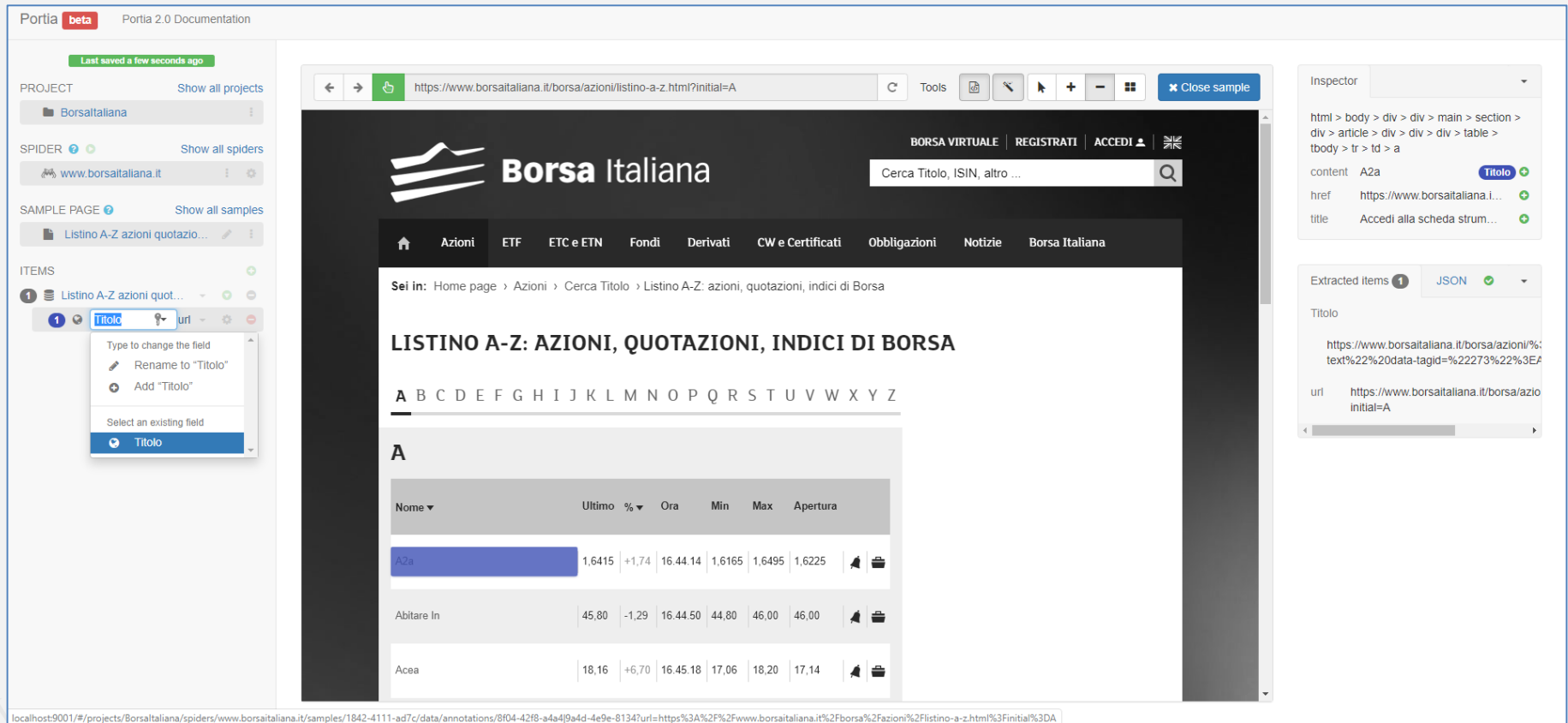
`(^https:\\\\)www\\.borsaitaliana\\.it\\borsa\\azioni\\listino-a-z\\.html\\?initial=[A-Z]`



The screenshot shows a web-based tool titled "RegEx Testing" with the subtitle "From Dan's Tools". It features a "Regular Expression" input field containing the regex `(^https:\\\\)www\\.borsaitaliana\\.it\\borsa\\azioni\\listino-a-z\\.html\\?initial=[A-Z]` with a "2 matches" indicator. Below the input is a "Test String" field containing three URLs: `https://www.borsaitaliana.it/borsa/azioni/listino-a-z.html?initial=A`, `https://www.borsaitaliana.it/borsa/azioni/listino-a-z.html?initial=B`, and `https://www.borsaitaliana.it/borsa/azioni/listino-a-z.html?initial=1`. The first two URLs are highlighted in blue, indicating they are matches. A "Substitution" field is visible at the bottom with a plus sign icon.

- Portia Web Scraper – Configurare lo Spider

- Creare una **Sample Page** per istruire lo spider all'estrazione dei dati di interesse;

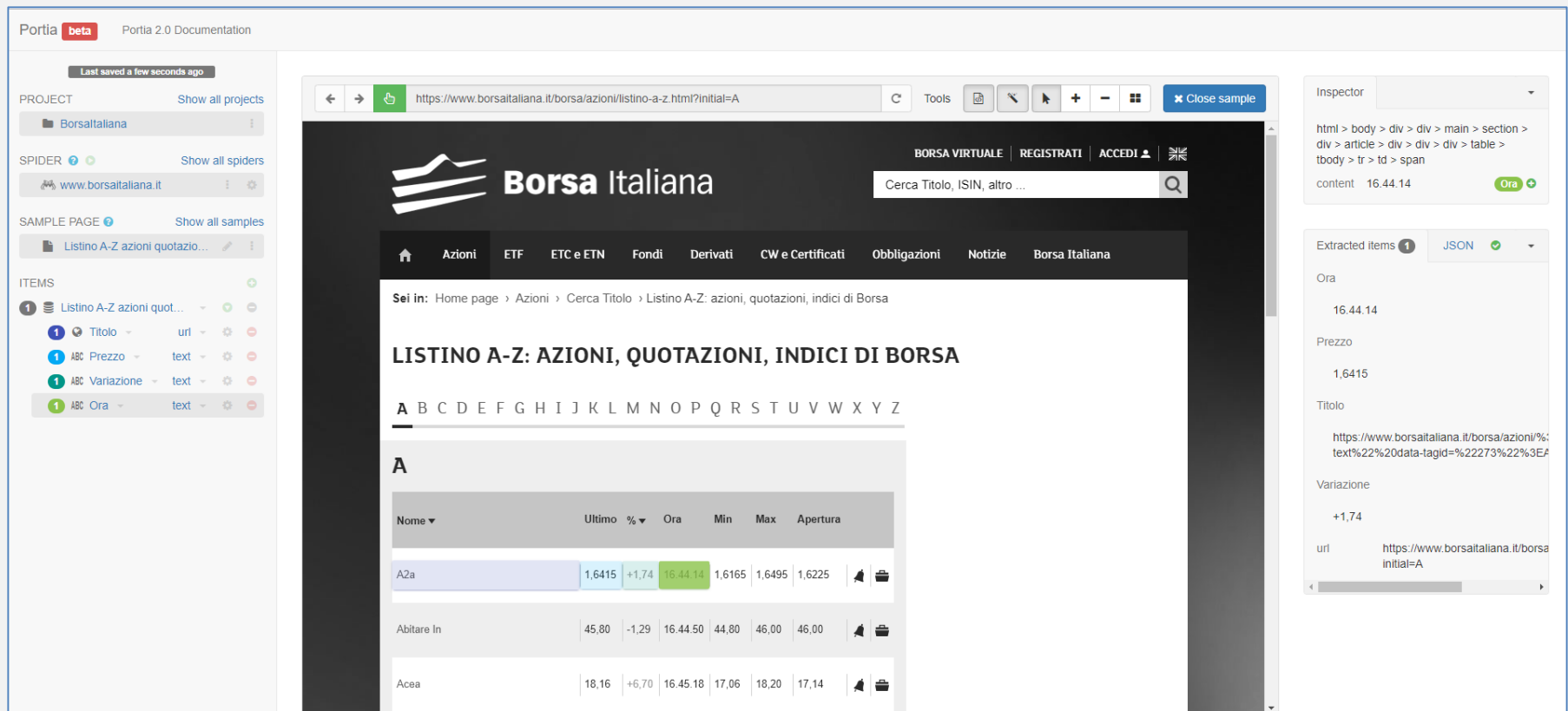


The screenshot shows the Portia web scraper interface. On the left, the 'PROJECT' sidebar shows a project named 'Borsaitaliana'. Under 'SPIDER', the URL 'www.borsaitaliana.it' is listed. Under 'SAMPLE PAGE', a sample page 'Listino A-Z azioni quotazio...' is selected. A dropdown menu is open for the 'ITEMS' section, showing a field named 'Titolo' selected. The main window displays the 'Borsa Italiana' website with the URL 'https://www.borsaitaliana.it/borsa/azioni/listino-a-z.html?initial=A'. The website content shows a table of stock data for 'LISTINO A-Z: AZIONI, QUOTAZIONI, INDICI DI BORSA'. The table has columns for 'Nome', 'Ultimo', '%', 'Ora', 'Min', 'Max', and 'Apertura'. The first row is for 'A2a' with values 1,6415, +1,74, 16.44.14, 1,6165, 1,6495, 1,6225. Other rows include 'Abitare In' and 'Acea'. On the right, the 'Inspector' panel shows the HTML path 'html > body > div > div > main > section > div > article > div > div > div > table > tbody > tr > td > a' and the extracted item 'Titolo' with its href and title. The 'Extracted items' panel shows the JSON output for the 'Titolo' field.

Nome	Ultimo	%	Ora	Min	Max	Apertura
A2a	1,6415	+1,74	16.44.14	1,6165	1,6495	1,6225
Abitare In	45,80	-1,29	16.44.50	44,80	46,00	46,00
Acea	18,16	+6,70	16.45.18	17,06	18,20	17,14

- Portia Web Scraper – Configurare lo Spider

- Creare una **Sample Page** per istruire lo spider all'estrazione dei dati di interesse;



The screenshot shows the Portia web scraper interface. On the left, there are panels for PROJECT, SPIDER, SAMPLE PAGE, and ITEMS. The main window displays a browser view of the Borsa Italiana website. The URL is <https://www.borsaitaliana.it/borsa/azioni/listino-a-z.html?initial=A>. The page content includes the Borsa Italiana logo, navigation tabs (Azioni, ETF, ETC e ETN, Fondi, Derivati, CW e Certificati, Obbligazioni, Notizie, Borsa Italiana), and a table of stock data under the heading "LISTINO A-Z: AZIONI, QUOTAZIONI, INDICI DI BORSA".

The table data is as follows:

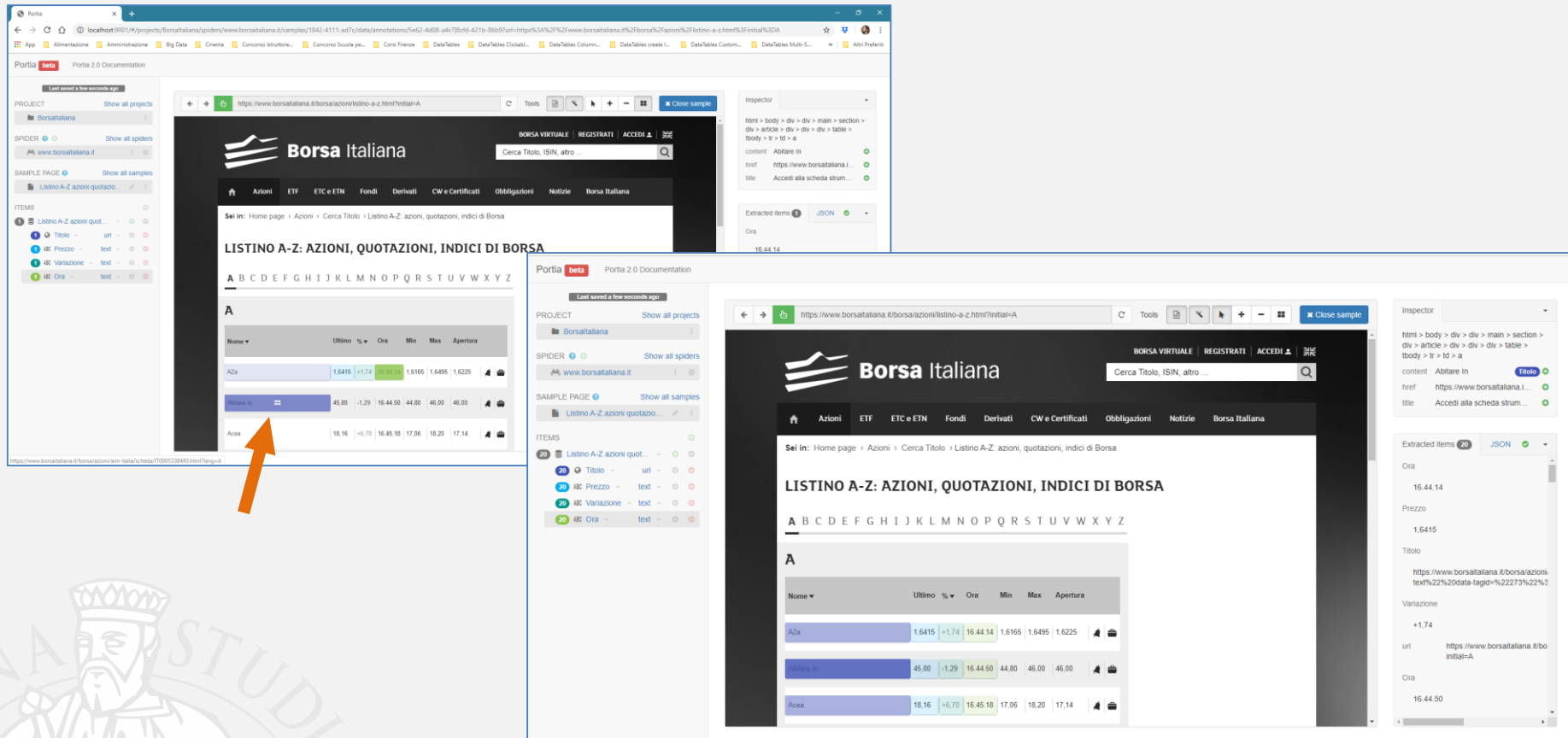
Nome	Ultimo	%	Ora	Min	Max	Apertura
A2a	1,6415	+1,74	16.44.14	1,6165	1,6495	1,6225
Abitare In	45,80	-1,29	16.44.50	44,80	46,00	46,00
Acea	18,16	+6,70	16.45.18	17,06	18,20	17,14

On the right side of the interface, the Inspector panel shows the HTML path: `html > body > div > div > main > section > div > article > div > div > div > table > tbody > tr > td > span`. The content is `16.44.14`. Below the Inspector, the Extracted items panel shows the following data:

- Ora: 16.44.14
- Prezzo: 1,6415
- Titolo: <https://www.borsaitaliana.it/borsa/azioni/text%22%20data-tagId=%22273%22%3E>
- Variazione: +1,74
- url: <https://www.borsaitaliana.it/borsa/azioni/initial=A>

- Portia Web Scraper – Configurare lo Spider

- E' possibile definire semplicemente attraverso l'interfaccia grafica gli elementi di interesse da estrarre, nonché identificare pattern che si ripetono all'interno della pagina.



The screenshot displays the Portia Web Scraper interface with the following components:

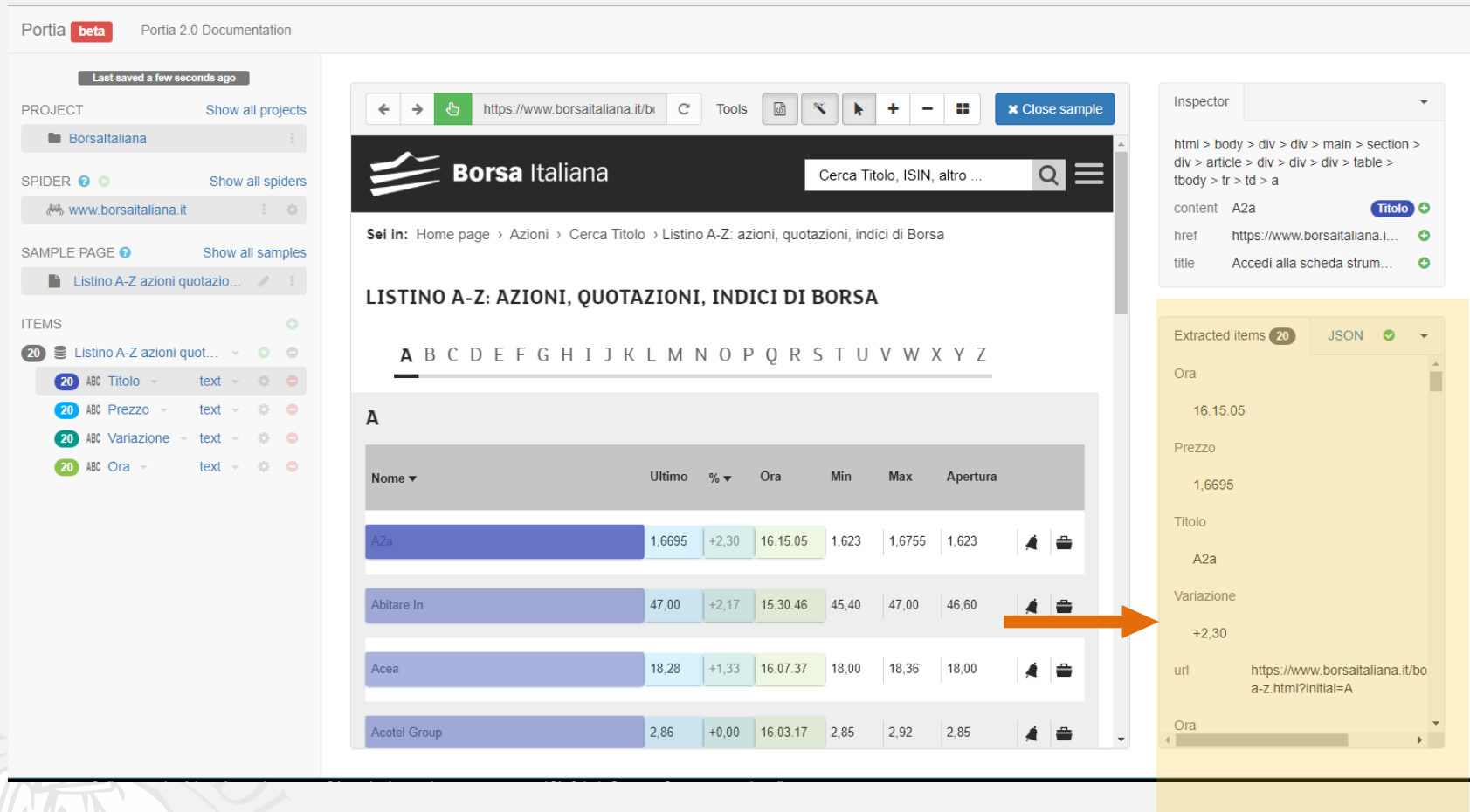
- PROJECT:** Borsa Italiana
- SPIDER:** www.borsaitaliana.it
- SAMPLE PAGE:** Listino A-Z azioni quotazio...
- ITEMS:** Listino A-Z azioni quot... (selected), Titolo, Prezzo, Variazione, Ora.
- Inspector:** Shows the HTML structure of the selected item, including the href and title.
- Extracted items:** Shows the extracted data for the selected item, including the title, price, and time.

An orange arrow points to the 'Variazione' item in the ITEMS list, which is highlighted in blue. The sample page view shows a table of stock data with columns for Name, Last Price, Change, Time, High, Low, and Open.

Nome	Ultimo	%	Ora	Min	Max	Apertura
A2a	1,6415	+1,74	16:44:14	1,6165	1,6495	1,6225
Borsaitalia	45,80	-1,29	16:44:50	44,80	46,00	46,00
Acqa	18,16	+6,70	16:45:18	17,06	18,20	17,14

- Portia Web Scraper – Avviare lo Spider

- Una preview dei risultati dell'estrazione è mostrata nel riquadro a destra dell'interfaccia



The screenshot displays the Portia web scraper interface. On the left, a sidebar shows the project 'Borsaitaliana', the spider 'www.borsaitaliana.it', and a sample page 'Listino A-Z azioni quotazio...'. Below this, a list of items is shown, including 'Listino A-Z azioni quot...', 'ABC Titolo', 'ABC Prezzo', 'ABC Variazione', and 'ABC Ora'. The central browser window shows the 'Borsa Italiana' website with the URL 'https://www.borsaitaliana.it/bi'. The page content includes a search bar, navigation links, and a table titled 'LISTINO A-Z: AZIONI, QUOTAZIONI, INDICI DI BORSA'. The table has columns for 'Nome', 'Ultimo', '%', 'Ora', 'Min', 'Max', and 'Apertura'. The first row is for 'A2a' with values: 1,6695, +2,30, 16.15.05, 1,623, 1,6755, 1,623. An orange arrow points from the 'Ora' column of the 'A2a' row to the 'Extracted Items' panel on the right. This panel shows a list of extracted items in JSON format, including 'Ora' (16.15.05), 'Prezzo' (1,6695), 'Titolo' (A2a), 'Variazione' (+2,30), and 'url' (https://www.borsaitaliana.it/bo-a-z.html?initial=A).

- Portia Web Scraper ~ References

- Documentazione dal sito ufficiale:

<https://portia.readthedocs.io/en/latest/installation.html#installation>

- Guida PDF:

<https://buildmedia.readthedocs.org/media/pdf/portia/latest/portia.pdf>



- References & Link Utili

➤ **NLP**

GATE: <http://gate.ac.uk/>

The Stanford University NLP Group: <https://nlp.stanford.edu/>

Italian NLP Lab: <http://www.italianlp.it/>

➤ **Web Crawlers**

Websphinx: <https://www.cs.cmu.edu/~rcm/websphinx/>

Heritrix: <http://crawler.archive.org/index.html>

Apache Nutch: <http://nutch.apache.org/>

➤ **Semantic Technologies**

The Semantic Web: https://en.wikipedia.org/wiki/Semantic_Web

FOAF (Friend-Of-a-Friend) Ontology: <http://xmlns.com/foaf/spec/>

Time Ontology: <https://www.w3.org/TR/owl-time/>

SKOS (Simple Knowledge Organization System) Ontology: <https://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html>

- Riferimenti

[Winograd, 1971]

Winograd, Terry (1971), *Procedures as a Representation for Data in a Computer Program for Understanding Natural Language*, MAC-TR-84, MIT Project MAC, 1971.

[IBM Watson, 2012]

<https://www-03.ibm.com/innovation/us/watson/>

[IOS Siri, 2010]

<http://www.apple.com/ios/siri/>

[Bellandi et al., 2012]

A. Bellandi, P. Bellini, A. Cappuccio, P. Nesi, G. Pantaleo, N. Rauch, *“Assisted Knowledge Base Generation, Management and Competence Retrieval”*, Int. Journal of Software Engineering and Knowledge Engineering, World Scientific Publishing Company, press, Vol. 32(8), pp.1007-1038, 2012.

[Nesi et al., 2016]

P. Nesi, G. Pantaleo and M. Tenti, *“Geographical localization of web domains and organization addresses recognition by employing natural language processing, Pattern Matching and clustering”*, Engineering Applications of Artificial Intelligence, Vol. 51, pp. 202-211, 2016.

[Massai et al., 2019]

Massai, L., Nesi, P., Pantaleo, G, *“PAVAL: A location-aware virtual personal assistant for retrieving geolocated points of interest and location-based services”*, in Engineering Applications of Artificial Intelligence, vol. 77, pp. 70-85, 2019.