

Scraping

Rachele Sprugnoli – rachele.sprugnoli@unicatt.it

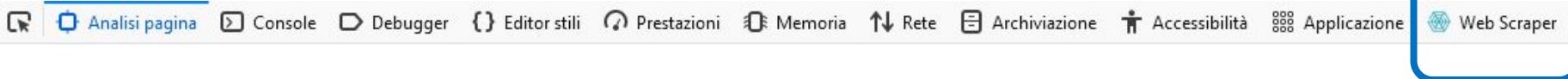
Centro Interdisciplinare di Ricerche per la Computerizzazione
dei Segni dell'Espressione (CIRCSE)



UNIVERSITÀ
CATTOLICA
del Sacro Cuore

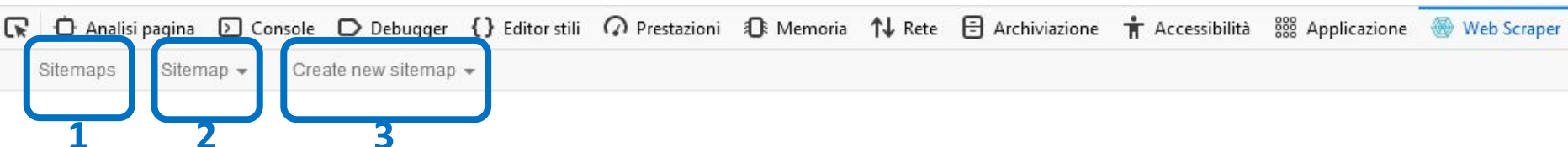
WEB SCRAPER

- Documentazione online: <https://www.webscraper.io/documentation>
- Video tutorial: <https://www.webscraper.io/tutorials>
- Aprire Web Scraper
 - aprire Firefox: l'icona dell'estensione deve apparire in alto a destra
 - l'estensione fa parte degli strumenti per sviluppatori, un'area separata che si apre sullo schermo come segue:
Windows, Linux: Ctrl+Shift+I oppure F12
Mac: Cmd+Opt+I
 - una volta aperta l'area degli strumenti per sviluppatori, cliccare su "Web Scraper"



WEB SCRAPER

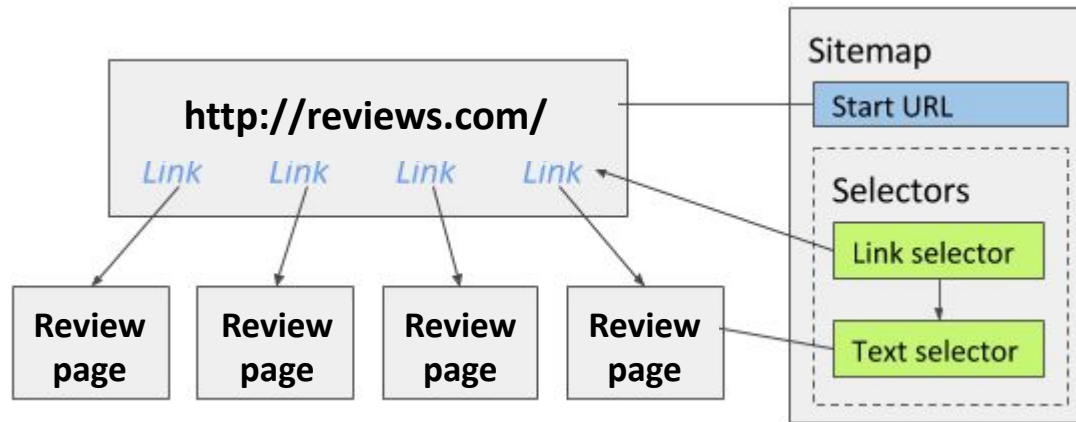
- Cliccando su “Web Scraper” sulla barra degli strumenti per sviluppatori, appare una sotto-barra con 3 opzioni:



- 1) Sitemaps: lista di sitemap (mappa della pagina web da cui fare scraping) create
- 2) Sitemap: lista di opzioni relativa ad una singola sitemap
- 3) Create new sitemap: lista di opzioni per creare una nuova sitemap
 - creazione da zero
 - importazione

WEB SCRAPER - TREE STRUCTURE

- Ragioniamo avendo in mente una struttura ad albero!



WEB SCRAPER - IMPORTAZIONE SITEMAP

- Nel menu a tendina “Create new sitemap”, scegliere l’opzione “Import sitemap”
- Aprire amazon-reviews.json con un editor di testo, selezionare tutto il testo e copiarlo
- Incollare il testo appena copiato su Web Scraper nello campo “Sitemap JSON” e salvare cliccando su “Impost sitemap”: appariranno i selettori creati per fare scraping delle review di Amazon

ID	Selector	type	Multiple	Parent selectors	Actions			
review	div.a-section.review	SelectorElement	yes	_root, next	Element preview	Data preview	Edit	Delete
next	li.a-last a	SelectorLink	no	_root, next	Element preview	Data preview	Edit	Delete

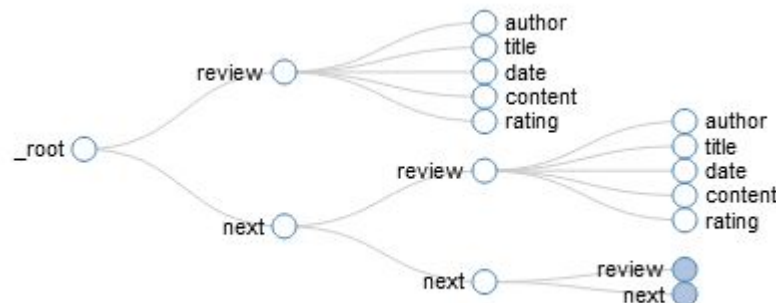
P.S. Potete seguire lo stesso procedimento anche per il sitemap contenuto nel file tripadvisor-reviews.json

WEB SCRAPER - IMPORTAZIONE SITEMAP

- Cliccando su il selector chiamato “review” ne appaiono altri: struttura ad albero!

ID	Selector	type	Multiple	Parent selectors	Actions			
author	span.a-profile-name	SelectorText	no	review	Element preview	Data preview	Edit	Delete
title	a.a-size-base.review-title	SelectorText	no	review	Element preview	Data preview	Edit	Delete
date	span.a-size-base.a-color-secondary	SelectorText	no	review	Element preview	Data preview	Edit	Delete
content	div.a-row.review-data span.a-size-base	SelectorText	no	review	Element preview	Data preview	Edit	Delete
rating	span.a-icon-alt	SelectorText	no	review	Element preview	Data preview	Edit	Delete

- Nel menu a tendina sotto “Sitemap Amazon” cliccare su “Selector graph” per vedere la struttura generale:



WEB SCRAPER - SCRAPING

- Andare sulla pagina web di cui si vuole fare lo scraping: nel menu a tendina “Sitemap amazon”, cliccare su “Edit metadata”, copiare la URL nel campo “Start URL” e incollarla nel campo di ricerca del browser

N.B. Se si vuole applicare la stessa sitemap ad un'altra pagina è sufficiente cambiare la URL nel campo “Start URL” di “Edit metadata”

ATTENZIONE: per scaricare correttamente tutte le recensioni è necessario usare la pagina “Recensioni clienti” e non la pagina principale del prodotto

Ad es. andare in fondo alla pagina principale del prodotto

(https://www.amazon.it/citt%C3%A0-invisibili-Oscar-opere-Calvino-ebook/dp/B008FHSP3Y/ref=cm_cr_ar_p_d_bdcrb_top?ie=UTF8) e cliccare su

“Visualizza tutte le recensioni”: in cima alla pagina con le recensioni

appare: Le città invisibili (Oscar opere di Italo Calvino Vol. 11) > Recensioni clienti

WEB SCRAPER - SCRAPING

- Nel menu a tendina “Sitemap amazon” cliccare su “Scrape” e poi su “Start scraping”



Request interval (ms)

2000

Page load delay (ms)

2000

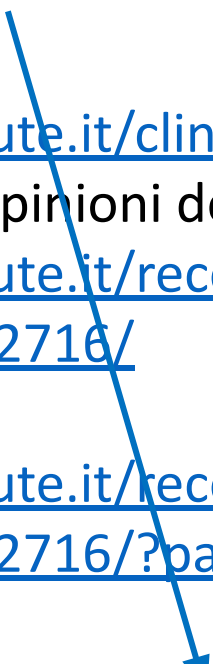

Start scraping

- Si apre una nuova finestra: lo scraping si conclude quando questa finestra si chiude e appare per qualche secondo un avviso in basso a destra
- Cliccare su “refresh”: appaiono i dati scaricati in formato tabellare, controllare che sia stato scaricato ciò che si voleva
- Nel menu a tendina “Sitemap amazon” cliccare su “Export data as CSV” e poi su “Download now” → il CSV si apre come foglio di calcolo, il separatore è una virgola

WEB SCRAPER - NUOVA SITEMAP

- Creare una sitemap da zero: il primo passo è sempre l'analisi della pagina web. Le recensioni sono su una sola pagina? Se sono su più pagine, la URL come cambia? Quali informazioni vogliamo scaricare?
 - Pagina iniziale:
<https://www.qsalute.it/clinica-citt%C3%A0-studi-milano/>
 - “Guarda tutte le opinioni degli utenti”:
https://www.qsalute.it/recensioni/health-ratings/clinica-citt%C3%A0-studi-milano_l2716/
 - Pagina 2:
https://www.qsalute.it/recensioni/health-ratings/clinica-citt%C3%A0-studi-milano_l2716/?page=2
 - Pagina 3:
https://www.qsalute.it/recensioni/health-ratings/clinica-citt%C3%A0-studi-milano_l2716/?page=3

WEB SCRAPER - NUOVA SITEMAP

- Creare una sitemap da zero: il primo passo è sempre l'analisi della pagina web. Le recensioni sono su una sola pagina? Se sono su più pagine, la URL come cambia? Quali informazioni vogliamo scaricare?
 - Pagina iniziale:
<https://www.qsalute.it/clinica-citt%C3%A0-studi-milano/>
 - “Guarda tutte le opinioni degli utenti”:
https://www.qsalute.it/recensioni/health-ratings/clinica-citt%C3%A0-studi-milano_l2716/
 - Pagina 2:
https://www.qsalute.it/recensioni/health-ratings/clinica-citt%C3%A0-studi-milano_l2716/?page=2
 - Pagina 3:
https://www.qsalute.it/recensioni/health-ratings/clinica-citt%C3%A0-studi-milano_l2716/?page=3
- 
- 

WEB SCRAPER - NUOVA SITEMAP

- Procedimento:
 1. andare sulla prima pagina delle recensioni e aprire web scraper
 2. “Create new sitemap” → “Create sitemap”
 3. “Sitemap name”: qsalute, “Start URL”:
[https://www.qsalute.it/recensioni/health-ratings/clinica-citt%C3%A0-studi-milano_l2716/?page=\[1-4\]](https://www.qsalute.it/recensioni/health-ratings/clinica-citt%C3%A0-studi-milano_l2716/?page=[1-4]) da pagina 1 a pagina 4 (l’ultima)
 4. salvare cliccando su “Create sitemap”
 5. si apre la sezione dei selectors: per scegliere i selettori bisogna ragionare a scatole cinesi, dalla struttura più grande alle sue sotto-parti

WEB SCRAPER - NUOVA SITEMAP

Silvia Peterlongo

05 Marzo, 2021


Encomio equipe Otorinolaringoiatria

Arrivata in PS per sospetto ascesso peritonsillare, mia figlia è stata presa in carico con estrema tempestività, professionalità e gentilezza dal dott. Vittorio Saginario e dall'equipe ORL dell'ospedale. Sottoposta a un intervento di incisione e drenaggio, è stata seguita successivamente con diversi controlli fino a completa guarigione. Ringrazio il dottore Saginario, l'intera equipe di otorini e il reparto, che in periodo Covid hanno curato con grande professionalità mia figlia.

**Patologia
trattata**

Ascesso peritonsillare.

Voto medio	★★★★★	5.0
Competenza	★★★★★	5.0
Assistenza	★★★★★	5.0
Pulizia	★★★★★	5.0
Servizi	★★★★★	5.0

 Commenti (0)



WEB SCRAPER - NUOVA SITEMAP

1: **elemento principale** che contiene vari sotto-elementi testuali → selector type: Element

Silvia Peterlongo

05 Marzo, 2021

Encomio equipe Otorinolaringoiatria

Arrivata in PS per sospetto ascesso peritonsillare, mia figlia è stata presa in carico con estrema tempestività, professionalità e gentilezza dal dott. Vittorio Saginario e dall'equipe ORL dell'ospedale. Sottoposta a un intervento di incisione e drenaggio, è stata seguita successivamente con diversi controlli fino a completa guarigione. Ringrazio il dottore Saginario, l'intera equipe di otorini e il reparto, che in periodo Covid hanno curato con grande professionalità mia figlia.

Patologia trattata

Ascesso peritonsillare.

Commenti (0)

Voto medio  5.0

Competenza  5.0

Assistenza  5.0

Pulizia  5.0

Servizi  5.0

WEB SCRAPER - NUOVA SITEMAP

2: **sotto-elementi** testuali: → selector type: Text

Silvia Peterlongo 05 Marzo, 2021

Encomio equipe Otorinolaringoiatria

Arrivata in PS per sospetto ascesso peritonsillare, mia figlia è stata presa in carico con estrema tempestività, professionalità e gentilezza dal dott. Vittorio Saginario e dall'equipe ORL dell'ospedale. Sottoposta a un intervento di incisione e drenaggio, è stata seguita successivamente con diversi controlli fino a completa guarigione. Ringrazio il dottore Saginario, l'intera equipe di otorini e il reparto, che in periodo Covid hanno curato con grande professionalità mia figlia.

Patologia trattata Ascesso peritonsillare.

Voto medio ★★★★★ 5.0

Competenza ★★★★★ 5.0

Assistenza ★★★★★ 5.0

Pulizia ★★★★★ 5.0

Servizi ★★★★★ 5.0

Commenti (0)

WEB SCRAPER - NUOVA SITEMAP

- Procedimento (continua):

6. sotto “Id” dare un nome al selettore, ad es. recensione
7. sotto “Type” scegliere Element
8. sotto “Selector” cliccare su “Select”: la pagina web diventa un ambiente cliccabile, le aree selezionabili vengono evidenziate passandoci sopra con il mouse → cliccare sull’area del riquadro della prima recensione che diventerà rossa poi cliccare sull’area del riquadro della seconda recensione → tutti gli altri riquadri simili verranno evidenziati di rosso (vedi prossima slide)
9. cliccare su “Done selecting”



10. selezionare l’opzione “Multiple”: in una pagina ci sono più recensioni
11. cliccare su “Save selector”

WEB SCRAPER - NUOVA SITEMAP

Silvia Peterlongo 05 Marzo, 2021

Encomio equipe Otorinolaringoiatria

Arrivata in PS per sospetto ascesso peritonsillare, mia figlia è stata presa in carico con estrema tempestività, professionalità e gentilezza dal dott. Vittorio Saginario e dall'equipe ORL dell'ospedale. Sottoposta a un intervento di incisione e drenaggio, è stata seguita successivamente con diversi controlli fino a completa guarigione. Ringrazio il dottore Saginario, l'intera equipe di otorini e il reparto, che in periodo Covid hanno curato con grande professionalità mia figlia.

Patologia trattata Ascesso peritonsillare.

Voto medio	★★★★★	5.0
Competenza	★★★★★	5.0
Assistenza	★★★★★	5.0
Pulizia	★★★★★	5.0
Servizi	★★★★★	5.0

Commenti (0)

Silvia Peterlongo 05 Marzo, 2021

Encomio equipe Otorinolaringoiatria

Arrivata in PS per sospetto ascesso peritonsillare, mia figlia è stata presa in carico con estrema tempestività, professionalità e gentilezza dal dott. Vittorio Saginario e dall'equipe ORL dell'ospedale. Sottoposta a un intervento di incisione e drenaggio, è stata seguita successivamente con diversi controlli fino a completa guarigione. Ringrazio il dottore Saginario, l'intera equipe di otorini e il reparto, che in periodo Covid hanno curato con grande professionalità mia figlia.

Patologia trattata Ascesso peritonsillare.

Voto medio	★★★★★	5.0
Competenza	★★★★★	5.0
Assistenza	★★★★★	5.0
Pulizia	★★★★★	5.0
Servizi	★★★★★	5.0

Commenti (0)

Silvia Peterlongo 05 Marzo, 2021

Encomio equipe Otorinolaringoiatria

Arrivata in PS per sospetto ascesso peritonsillare, mia figlia è stata presa in carico con estrema tempestività, professionalità e gentilezza dal dott. Vittorio Saginario e dall'equipe ORL dell'ospedale. Sottoposta a un intervento di incisione e drenaggio, è stata seguita successivamente con diversi controlli fino a completa guarigione. Ringrazio il dottore Saginario, l'intera equipe di otorini e il reparto, che in periodo Covid hanno curato con grande professionalità mia figlia.

Patologia trattata Ascesso peritonsillare.

Voto medio	★★★★★	5.0
Competenza	★★★★★	5.0
Assistenza	★★★★★	5.0
Pulizia	★★★★★	5.0
Servizi	★★★★★	5.0

Commenti (0)

Alessandro 26 Febbraio, 2021

Neurochirurgia: revisione artrodesi lombare

Fin dal pre-ricovero è filato tutto liscio, considerando il periodo di pandemia in corso di Covid-19. Ho poi trovato personale infermieristico gentile e preparato durante la degenza. Intervento di revisione artrodesi in Alif a livello I5-s1 riuscito al meglio, col primario di neurochirurgia Dott. Flavio Tandoni che è riuscito a rimediare agli errori fatti dal precedente chirurgo, ed anche in modo poco invasivo. Posso solo ringraziare di cuore tutto il reparto di neurochirurgia, dove ho trovato umanità e professionalità. Sono molto soddisfatto, grazie.

Patologia trattata Disco di I5-s1 collassato.

Voto medio	★★★★	4.5
Competenza	★★★★	5.0
Assistenza	★★★★	5.0
Pulizia	★★★★	4.0
Servizi	★★★★	4.0

Commenti (0)

Francesca 25 Febbraio, 2021

Ecodoppler arto inferiore

Nel mese di Febbraio 2021 sono stata fortunata: grazie alla gentilezza e competenza delle addette al Cup del centro prenotazioni, mi hanno trovato disponibilità per una visita con un'angiologa, la dottoressa Luisella Troyer. La visita è stata effettuata, con ecodoppler un po' frettoloso ma completo. Il mio più grande problema era l'aver da un bel po' di tempo il piede e il polpaccio sinistri gonfi. Chiedendo e spiegando visibilmente cosa potessi fare, lei mi ha risposto che non conoscendomi non poteva darmi alcun consiglio, e di rivolgermi al mio dottore di base. Molto scostante e soccia. Di molte poche parole. Mi dispiace, in quanto la struttura Città Studi è per me una delle migliori, ma difficilmente prenoterò ancora con questa specialista. Grazie.

Patologia trattata Ecodoppler per arto gonfio.

Voto medio	★★★	3.5
Competenza	★★★	3.0
Assistenza	★★	1.0
Pulizia	★★★★	5.0
Servizi	★★★★	5.0

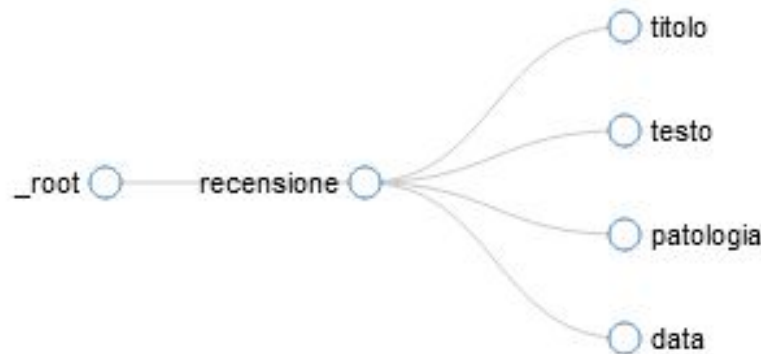
Commenti (0)

WEB SCRAPER - NUOVA SITEMAP

- Procedimento (continua):
 12. cliccare sul selettore appena creato per creare una gerarchia tra elemento principale ed sotto-elementi testuali
 13. cliccare su “Add new selector”
 14. sotto “Id” dare un nome al selettore, ad es. titolo
 15. sotto “Type” scegliere Text
 16. sotto “Selector” cliccare su “Select”: cliccare sul testo del primo titolo e poi su “Done selecting”, l’HTML corrispondente sarà h4
 17. NON selezionare “Multiple”: ogni recensione ha un solo titolo
 18. sotto “Parent Selectors” deve essere evidenziato il nome dell’elemento superiore, “recensione”
 19. cliccare su “Save selector”

WEB SCRAPER - NUOVA SITEMAP

- Procedimento (continua):
20. ripetere lo stesso procedimento per ogni sotto-elemento testuale che si vuole estrarre: ad esempio, il testo della recensione, la patologia trattata e la data
 21. per controllare di aver selezionato la porzione di pagina web corretta, cliccare su “Data preview”
 22. il selector graph corrispondente sarà come quello che segue:



23. procedere allo scraping e al download dei dati

WEB SCRAPER - SCROLL DOWN

- Alcune pagine web caricano il testo man mano che l'utente scorre verso il basso: meccanismo di scroll down
 - esempio: commenti su Youtube e su Instagram

Procedimento:

1. andare su un video di Youtube come <https://www.youtube.com/watch?v=Yh2EXENIOz8> e aprire Web Scraper
2. creare una nuova sitemap di nome "youtube"
3. cliccare su "Add new selector"
4. dare un nome al selettore nel campo "Id", ad esempio commento
5. scegliere come "Selector type" "Element scroll down"
6. cliccare su "Select"

WEB SCRAPER - SCROLL DOWN

- Procedimento (continua):
7. selezionare il rettangolo di pagina web che comprende tutte le informazioni sul primo commento, cliccare sullo stesso tipo di rettangolo per il secondo rettangolo: tutti i rettangoli simili diventeranno rossi



WEB SCRAPER - SCROLL DOWN

- Procedimento (continua):
- 8. selezionare il rettangolo di pagina web che comprende tutte le informazioni sul primo commento, cliccare sullo stesso tipo di rettangolo per il secondo rettangolo: tutti i rettangoli simili diventeranno rossi
- 9. selezionare “Multiple”
- 10. salvare il selector
- 11. cliccare sul selector e creare dei selector di tipo Text per le parti che si vogliono scaricare: ad esempio, la data e il testo del commento. Questi selector devono avere il selector “commento” come parent
- 12. fare scraping e salvare i dati

N.B. Rilevato bug! Se lo scraping non va a buon fine: andare sul selector “commento” e cliccare su “Data preview”. Copiare i dati che appaiono e incollarli in un foglio di calcolo

TWINT

- Tool per il download di tweet scritto in Python: non richiede di avere un account Twitter né l'autorizzazione di Twitter per il download dei dati
- Esempi:
<https://github.com/twintproject/twint#cli-basic-examples-and-comb-os>
- Lista completa opzioni:
<https://github.com/twintproject/twint/wiki/Basic-usage>
- Aiuto sul terminale: `twint --help`

TWINT - ESEMPI

- Aprire il terminale, digitare cd, trascinare la cartella twint-master e premere invio: in questo modo entriamo nella cartella di Twint

Esempio di comandi:

- Scrape di tweet da un account:

twint -u username → twint -u SenatoStampa



- Scrape di tweet contenenti una parola (più parole divise da ,)

twint -s parola → twint -s #vaccini

- Scrape di tweet di un account e contenenti una certa parola

twint -u SenatoStampa -s #vaccini

N.B.

1. ricordarsi di premere invio dopo ogni comando per eseguirlo
2. ctrl+c per interrompere l'esecuzione del comando
3. in alcuni Mac bisogna togliere lo spazio tra opzione e valore (e.g. -s#vaccini)

TWINT - ESEMPI CON DATE

- Filtrare in base alle date
 - Scrape di tweet pubblicati a partire da una certa data e contenenti una certa parola
twint -s #JuventusBenevento --since "2021-03-21"
 - Scrape di tweet a partire da una data e un orario e contenenti una certa parola
twint -s #JuventusBenevento --since "2021-03-21 15:00:00"
 - Scrape di tweet a partire da una data e un orario fino ad un'altra data e un altro orario e contenenti una certa parola
twint -s #JuventusBenevento --since "2021-03-21 15:00:00"
--until "2021-03-22 15:00:00"

p.s. Attenzione al formato di date e orari!

TWINT - SALVARE I DATI

- Salvare in un file i dati scaricati
 - salvataggio su un file di testo
twint -u SenatoStampa -s #vaccini -o outVacciniSenato.txt
 - salvataggio su un file formato csv:
ATTENZIONE, il separatore è un tab non una virgola
twint -u SenatoStampa -s #vaccini -o outVacciniSenato.csv --csv

I file txt si possono aprire con un editor di testo, come Sublime Text

I file csv si aprono come fogli di calcolo: Calc, Numbers, Excel, Fogli Google

ALTRI TOOL DI SCRAPING

- Strumenti stand-alone (da scaricare sul proprio pc) commerciali (versione gratuita limitata):
 - Octoparse: <https://www.octoparse.com/>
 - Parsehub: <https://www.parsehub.com/>
- Strumento gratuito ma solo per Mac:
 - Site sucker: <https://ricks-apps.com/osx/sitesucker/index.html>
- Estensioni per browser:
 - Data Scraper: <https://dataminer.io/>



GRAZIE!

Email: rachele.sprugnoli@unicatt.it

Twitter: [@RSprugnoli](https://twitter.com/RSprugnoli)

