Confronto tra diverse tecniche di web scraping nella rilevazione dei prezzi al consumo per i pacchetti vacanza internazionali e nazionali.

CNS14 Conferenza Nazionale di Statistica

GIUSEPPINA NATALE - Istat natale@istat.it | MASSIMILIANO AMARONE - Istat amarone@istat.it | RICCARDO GIANNINI - Istat rigianni@istat.it

1. La rilevazione dei prezzi al consumo dei pacchetti vacanza

1.1 Campo di osservazione

I **pacchetti vacanza** rientrano nel paniere dei prodotti a rilevazione centralizzata e partecipano all'indagine mensile sui prezzi al consumo, rientrano nella divisione di spesa **ECOICOP 09 (Ricreazione, spettacoli cultura)**. La rilevazione dei prezzi al consumo per i pacchetti vacanza comprende due aggregati di prodotto:

- Pacchetti Vacanza Internazionali;
- Pacchetti Vacanza Nazionali

1.2 Dettaglio I pacchetti vacanza internazionali e nazionali

La rilevazione dei prezzi si basa sulla consultazione mensile dei listini/cataloghi on-line offerti dai tour operator presenti nel campione. La quantità dei prezzi da rilevare mensilmente è molto elevata, 282 pacchetti internazionali stratificati per destinazioni estere e macroaree geografiche internazionali e 126 nazionali stratificati per macroaree geografiche nazionali e tipologie di viaggio. I tour operator sono i più rappresentativi del settore turistico e sono soci *Astoi Confindustria Viaggi (Associazione aderente a Federturismo/Confindustria)*

L'inserimento dei prezzi per questi aggregati è molto impegnativo si è pensato di studiare quindi la possibilità di estrarre i prezzi tramite web scraping partecipando alla call del *LabInn Laboratorio Innovazione Istat*, tale progetto è stato approvato dal *comitato scientifico* e sviluppato all'interno del Laboratorio.

2. Progetto di ricerca sviluppato all'interno di LabInn – Laboratorio Innovazione Istat

Tale progetto è stato sviluppato all'interno del LabInn:
Confronto tra le tecnologie di Web Scraping per l'acquisizione dei prezzi; considerando 4 campioni

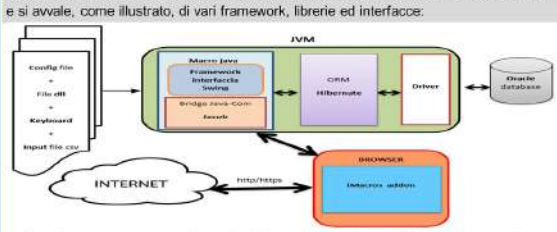
- 2 per i Pacchetti Vacanza Internazionali (campione fisso/variabile Last Minute)
- 2 per i Pacchetti Vacanza Nazionali (campione fisso/variabile Last Minute).

Confronto, sperimentazione, utilizzo, innovazione, tutto il progetto ruota intorno a queste quattro parole, Confronto tra due diverse tecnologie di web scraping, sperimentazione rilevando i prezzi last minute offerti dai tour operator coinvolti come suggerito da Eurostat, utilizzo del web scraping per la rilevazione mensile per entrambi gli aggregati, innovazione sostituzione della rilevazione manuale con il web scraping. I campioni sono due per ogni singolo aggregato uno fisso ed uno variabile, formato dai last minute e dalle offerte proposte dai tour operator.

L'introduzione dei Pacchetti Vacanza «Last Minute» è suggerito da **Eurostat** nel manuale metodologico «Harmonised Index of Consumer (HICP)» novembre 2018; capitolo 12.5 Flights and package holidays.

2.1 Tecnologia iMacros

Architettura interna della tecnologia iMacros Ogni singola macro basata su iMacros può essere utilizzata sia su personal computer che Server Windows, è realizzata in Java e si avvale, come illustrato, di vari framework, librerie ed interfacce:



- **Macro java**: questo programma è eseguito all'interno della **Java Virtual Machine**, oltre a coordinare tutte le operazioni all'interno del browser, si occupa anche di eseguire il cleaning dei dati letti e comandare la scrittura su Db attraverso l'ORM Hibernate.
- **Browser**: Mozilla Firefox è il browser all'interno del quale avviene il webscraping.
- **Jacob:** questa è una libreria core per il webscraping e rappresenta un bridge JAVA-COM che consente di chiamare i componenti di automazione COM da Java. La macro java interagisce con il browser passando ad esso i comandi di webscraping, poi eseguiti utilizzando l' addon di iMacros.
- **iMacros addon:** il browser grazie a questa estensione dialoga con il modulo Jacob della macro Java.
- **Framework Swing:** è utilizzato per elevare la *user experience* dell'utilizzatore grazie alle GUI.
- **ORM Hibernate**: attraverso questo modulo software si semplifica l'interfacciamento del mondo ad oggetti Java con quello relazionale del Db Oracle utilizzato.
- **Config** file: è un file testo letto all'avvio della macro, qui sono scolpiti tutti i comandi iMacros.
- **Input** file csv: qui risiedono i dati utilizzati della macro per svolgere la rilevazione.

2.2 Tecnologia Java

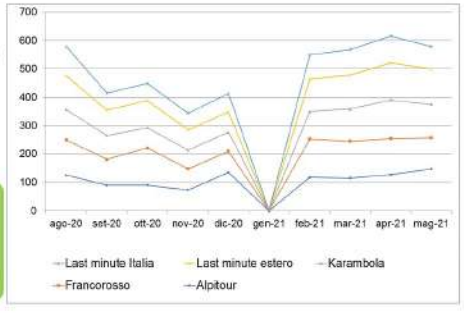
Portali da scansionare I portali da scansionare delle offerte e dei last minute relativi ai viaggi sono i seguenti: Alpitour, Bravo Club, Eden Viaggi, Francorosso, Karambola, Presstour, Swantour, Settemari, Veratour, Viaggi Idea.

Tecnologia utilizzata La tecnologia utilizzata è Java con l'estensione delle librerie Selenium necessarie per l'automazione dei browser web. Al centro di Selenium c'è il WebDriver, un'interfaccia per scrivere set di istruzioni che possono essere eseguite in modo intercambiabile in molti browser.

Descrizione del progetto Il progetto si compone di tre processi principali. Il primo estrae e scrive su file le informazioni relative alle offerte e ai last minute dai portali specificati. Questo processo avvia il browser con il set di istruzioni personalizzate per ottenere le informazioni ricercate. Inoltre è' automatizzato e schedulato per essere eseguito quotidianamente su server Linux. Il secondo processo legge i file appena scritti, ne normalizza le classificazioni e li carica su database Oracle in forma strutturata. Il terzo ed ultimo legge le informazioni salvate nel database per calcolare alcuni indicatori relativi alle statistiche sul turismo.

3 Risultati

Per il campione fisso è stato impossibile catturare i prezzi tramite web scraping per la grande quantità di informazioni presenti nei pdf, i quali necessitano di una continua manutenzione dal punto di vista informatico, invece per il campione variabile la scelta del web scraping è risultata ottimale. I dati sono stati rilevati da luglio 2020 fino a ottobre 2021, il numero di quotazioni raccolte è molto elevato, i prezzi sono disponibili in due file distinti *last minute* ed *offerte* racchiuse in un DB Oracle, i tour operator coinvolti sono 10, oltre ai prezzi sono stati rilevati tutte le informazioni aggiuntive presenti, come ad esempio hotel, trattamento vacanza, volo. Le destinazioni per l'Italia riguardano le regioni meridionali, quelle estere le località di mare come ad esempio Grecia, Egitto. Sono stati calcolati gli indici di Laysperes con base di calcolo luglio 2020. L'andamento di questi indici (sia Italia sia Estero) rispecchia quello tipico dei pacchetti vacanza influenzati dalla stagionalità. Questo progetto analizzato durante la pandemia mondiale ha messo in evidenza alcune difficoltà nell'utilizzare i prezzi in quanto erano disponibili nei cataloghi ma non acquistabili dagli acquirenti, invece questi prezzi sono disponibili ed utilizzabili, tuttavia la scelta di un integrazione con il campione variabile richiede adeguati aggiustamenti di qualità.



Tecnologia iMacros

- Raggiunge il risultato per last minute/ offerte, richiede una maggiore manutenzione rispetto alla tecnologia Java, poiché prima di lanciare le singole macro è necessario predisporre i file di input da aggiornare ad ogni lancio.

Tecnologia Java

- Raggiunto il risultato per last minute/ offerte, catturando tutti i giorni i prezzi presenti senza creare nessun file aggiuntivo oltre quello iniziale, con uno sforzo minimo ottenendo una grande quantità di informazioni.