PROGETTI CONCLUSI II CALL

Titolo progetto	Confronto delle tecniche di web scraping nella rilevazione dei prezzi al consumo per i pacchetti vacanza internazionali e nazionali
Descrizione	Nella rilevazione dei prezzi al consumo rientrano otto aggregati di consumo della filiera turistica che vengono monitorati mensilmente tramite consultazioni di listini o cataloghi on line. ***La rilevazione dei pacchetti vacanza internazionali*** si basa su un campione di 282 pacchetti offerti da 21 tour operator nazionali, stratificati per 12 macroaree geografiche estere e 41 destinazioni estere. ***La rilevazione dei pacchetti vacanza nazionali*** utilizza un campione di 126 pacchetti proposti da 8 tour operator nazionali, stratificati per 4 macroaree geografiche nazionali (nord ovest, nord est, centro, sud e isole) e 4 diverse tipologie di viaggio (mare, monti, arte e benessere). L'idea principale è quella di confrontare diverse tecniche di "web scraping" per l'acquisizione dei prezzi di due aggregati turistici: "pacchetti vacanza internazionali e pacchetti vacanza nazionali" estendendo l'utilizzo di queste tecniche sia al campione attuale delle due rilevazioni sia ad un campione parallelo costruito considerando le offerte (variabili) last minute proposte mensilmente dai principali tour operator nazionali, seguendo in tal modo i suggerimenti determinati da Eurostat nella bozza "TFQI15.2018/05 recommendation on the treatment of tangible services purchased in advance and/or priced flexibly".
Obiettivi	**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 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.

Metodologia	Tecniche di web scraping: **Software iMacros** - Architettura interna della tecnologia iMacros. Ogni singola macro, basata sulla tecnologia proprietaria 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 e può eventualmente beneficiare, nel ciclo di vita del software, del "Customer Support Channels" del produttore. **Software scritto in Java con utilizzo delle API della libreria HtmlUnit** 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.
Risultati ottenuti specificare l'impatto sulla produzione statistica	Per il campione fisso è stato impossibile catturare i prezzi tramite web scraping per la grande quantità di informazioni presenti nei vari 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 (Covid -19) ha messo in evidenza alcune difficoltà nell'utilizzare i prezzi, in quanto, erano disponibili nei cataloghi o nei listini ma non acquistabili dagli acquirenti, invece questi prezzi preposti dai tuor operator come last minute / offerte rilevati tramite web scraping sono disponibili ed utilizzabili poiché considerano soltanto le destinazioni dove era concesso viaggiare, tuttavia la scelta di un'integrazione*

	aggiungendo il campione variabile a quello fisso richiede adeguati aggiustamenti di qualità. Di seguito alcune considerazioni confrontando tra le due tecnologie di web scarping coinvolte nel progetto: *Tecnologia iMacros* Raggiunge il risultato per last minute/ offerte, essendo realizzato a livello prototipale 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* Raggiunge 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.
Membri del team	Giuseppina Natale - Massimiliano Amarone – Riccardo Giannini
Nome cognome e indirizzo e mail	Giuseppina Natale natale@istat.it Massimiliano Amarone amarone@istat.it Riccardo Giannini rigianni@istat.it