Diogenes: build Corpus from Social Network

Sandro Stancampiano[1]

[1]Istat – sandro.stancampiano@istat.it

Abstract[1]

This paper presents "Diogenes": a new Text Mining support software. It automates the preliminary steps of a Text Mining process.
In order to be able to study texts and carry out analysis we must create corpus of sufficiently large. The exponential growth of social network has caused an increase in potentially available texts to conduct studies and research. With this in mind we are developing this software.
Diogenes performs Data Wrangling with the aim to extract and organise textual data from platform as TripAdvisor. It automates, for example, the process of finding reviews on restaurants, points of interest or hotels on the web and collating them in a single corpus.
Currently designed to read reviews on TripAdvisor in the 28 languages supported by the platform. This software developed in Java runs on any hardware which has the Java Virtual Machine (JVM), therefore is platform independent. It follows design principles such as extensibility and scalability; these characteristics allow future developments of additional modules. Diogenes to parse the web page uses the API (Application Programming Interface) made available by the Jsoup (Java HTML Parser) library.
Once the data scraped from the target site are stored in the Database it is possible build the corpus: currently, using Diogenes you can create corpus properly formatted for IRaMuTeQ and TaLTaC2.
Therefore, it is possible to use this software as a complementary tool to more specific software dedicated to textual analysis. Avoiding manually and time-consuming operations during the preliminary stages prior the text mining research.
As far as we know this is the first software that build a Corpus setting few parameters as the Url and the numbers of the page of reviews to download.

Keywords: Diogenes, text mining, Java, Jsoup, IRaMuTeQ, Web Scraping, Data Wrangling.

Abstract

In questo documento presentiamo il software « Diogenes » progettato per supportare le analisi di Text Mining. Il software consente di creare corpus di grandi dimensioni scaricando dati testuali dal web. Attualmente il software scarica e salva nella base dati relazionale automaticamente i contenuti del sito TripAdvisor e consente agli utenti di creare corpus formattati per TaLTaC2 e IRaMuTeQ. Le caratteristiche principali sono la modularità e il basso accoppiamento, peculiarità che consentono di integrare ulteriori funzioni e nuove sorgenti dati per contribuire alla estrazione dei dati dal web al fine di creare conoscenza.
La presenza di questo software è innovativa e promettente in quanto si tratta del primo software che utilizzando pochi parametri consente al ricercatpre di creare corpus di dati testuali da analizzare in pochi passaggi per effettuare studi e ricerche.

Parole Chiave: Diogenes, text mining, Java, IRaMuTeQ, Web Scraping, Data Wrangling.

[1] Le opinioni espresse in questo contributo sono sotto la responsabilità dell'autore e non riflettono necessariamente le politiche dell'ISTAT (Istituto Nazionale di Statistica).

1. Data Wrangling - Il progetto Diogenes

Il progetto Diogenes è stato sviluppato per prelevare e organizzare i dati necessari alla realizzazione dell'analisi testuale: questo software è stato realizzato utilizzando il linguaggio Java e la metodologia di progettazione Object Oriented (Larman, 2005).

Si tratta di un *software* che consente di scaricare e organizzare dati non strutturati dal web: con l'espressione *data wrangling* intendiamo il processo di trasformazione e mappatura dei dati al fine di raggiungere l'obiettivo prefissato.

Il software è composto da tre moduli indipendenti caratterizzati dalla massima coesione interna e dal minimo accoppiamento tra loro. I moduli, difatti comunicano mediante interfacce (Figura 1). Questo approccio consente di modificare i singoli componenti senza pregiudicare il funzionamento complessivo del programma (Gamma et al., 1995). Inoltre è possibile aggiungere ulteriori funzioni al nucleo attuale per soddisfare ulteriori necessità.
La filiera prevede i seguenti passaggi: il primo modulo si occupa di stabilire le connessioni verso il sito obiettivo, il secondo modulo gestisce le operazioni di scrittura e lettura della base di dati, il terzo modulo gestisce la formattazione del corpus e la sua scrittura sul *file system* in formato UTF8.

Figura 1: I principali componenti del sistema

2. Le fasi del progetto

La filiera di estrazione e organizzazione dei dati si suddivide nelle seguenti fasi:

1. *web scraping*;
2. preparazione del dataset in funzione dell'obiettivo della ricerca;
3. creazione del corpus da analizzare.

Le fasi sono state gestite mediante i moduli indipendenti applicando la logica del *divide et impera* per consentire la scalabilità del *software*.
La progettazione e realizzazione dell'applicativo di *data wrangling* (Figura 2) ha compreso inoltre tutte le fasi relative alla definizione di una base di dati che è stata gestita per mezzo del DBMS (*Database Management System*) MySQL (Elmasri e Navathe, 2011).
La procedura prevede l'aggiornamento automatico della base di dati ogni 48 ore (valore configurabile dall'utente) per consentire l'allineamento quasi in tempo reale tra i contenuti pubblicati dagli utenti sul sito obiettivo e quelli presenti nella base di dati che rappresentano la sorgente per la realizzazione del corpus. Attualmente il *software* è in grado di scaricare contenuti dalla sito TripAdvisor e supporta tutte le 28 lingue presenti sulla piattaforma. Di conseguenza è possibile creare corpus in lingue differenti ed effettuare studi comparati utilizzandolo in coppia con IRaMuTeQ per applicare tecniche di studio supportate dal software di scuola francese (Stancampiano, 2019a; Stancampiano, 2019b).

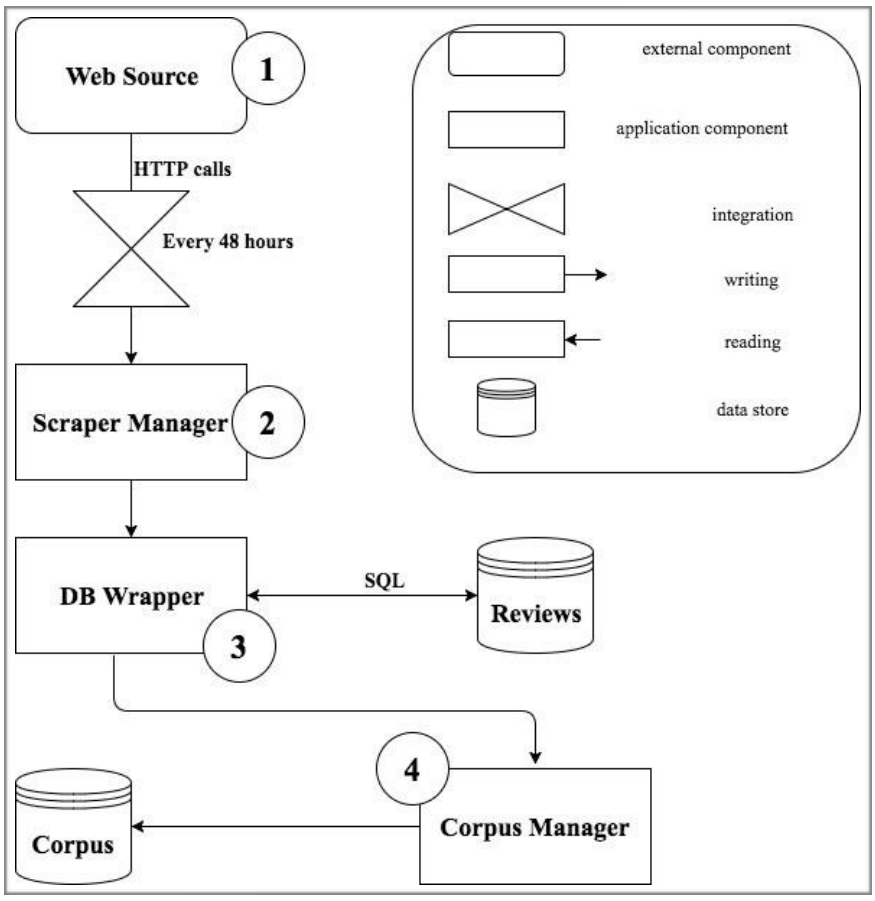


Figura 2: Il processo di data wrangling

3. Web scraping

Le informazioni prelevate tramite la fase di *scraping* sono successivamente memorizzate sulla base di dati. Utilizzare un *database* relazionale consente di effettuare ulteriori manipolazioni dei dati per aderire alle esigenze specifiche dei progetti cui fornire supporto, adottando un approccio *risk driven* (Sommerville, 2011).

Il passaggio dei dati dalla sorgente esterna, fuori dal controllo del *software* Diogenes, alla base dati interna al sistema, permette di evitare i principali rischi tipici delle tecniche di *scraping*: eventuali modifiche alla struttura delle pagine web che ospitano i dati da scaricare, servizi di rete poco affidabili e *policy* adottate dai gestori dei siti per bloccare le connessioni in entrata[2]. Ogni pagina web è un documento HTML[3] organizzato in una gerarchia rappresentata dalla struttura ad albero nel DOM[4], i nodi dell'albero rappresentano gli elementi

[2] Il software Diogenes risolve questo problema simulando il comportamento umano, facendo trascorrere 1500 ms tra le richieste di connessione al sito obiettivo.

[3] Hyper Text Markup Language (HTML), è un linguaggio di *markup* utilizzato per la creazione e la formattazione di documenti destinati al web.

[4] Document Object Model (DOM) è lo standard ufficiale del W3C (World Wide Web Consortium) per la rappresentazione di documenti strutturati (xml, xhtml, html, ecc.) in maniera da essere neutrali sia per la lingua sia per la piattaforma.

(tag html) come <body> o <p>.
Nelle figure 3 e 4 vediamo un semplice documento HTML e la sua raffigurazione come modello a oggetti. La struttura del DOM è costituita da un insieme di nodi collegati da archi che impongono un rapporto di parentela tra i nodi e definisce proprietà e metodi per attraversare e manipolare il documento. Diogenes per effettuare il *parsing* della pagina web utilizza le API (*Application Programming Interface*[5]) messe a disposizione dalla libreria *jsoup*[6]. La libreria permette l'analisi e la visita del documento per estrarre le informazioni di interesse, in questo caso le recensioni presenti sul sito TripAdvisor.

```
<html>
  <head>
    <title> titolo del doc </title>
  </head>
  <body>
    <h1> header del doc </h1>
    <p>questo è un paragrafo</p>
  </body>
</html>
```

Figura 3: Un documento HTML

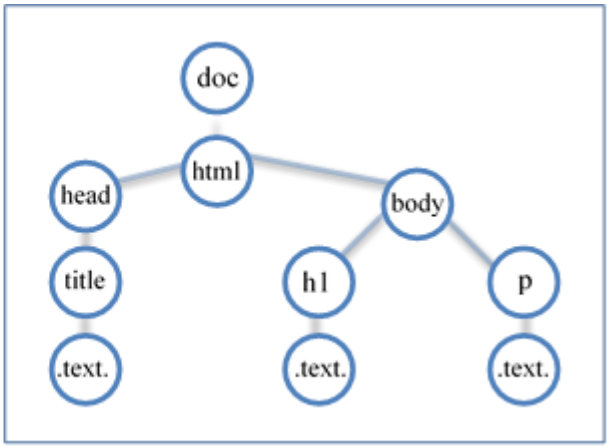


Figura 4: Rappresentazione ad albero

I nodi rappresentano la chiave per accedere alle informazioni visibili all'utente, in alcuni casi si utilizzano anche nodi con informazioni non visibili ma necessarie al processo di *scraping*.
La lettura e il salvataggio dei dati avviene in due passaggi successivi: prima si leggono **n** pagine web, si valorizza una collezione di oggetti che vengono inseriti nella base dati mediante il modulo « DB Wrapper »; successivamente si torna sulla base dati per leggere l'URL[7] di dettaglio, si stabilisce una nuova connessione HTTP[8] per ognuna delle recensioni e si aggiorna ogni record inserito nella fase precedente valorizzando il campo *text full* che contiene il testo della recensione.
Nella Figura 5 vediamo il *System Sequence Diagram* relativo al processo di acquisizione delle recensioni nella base dati di Diogenes.
La classe GrabTAEntryPoint gestisce l'intero processo utilizzando, tramite apposita interfaccia, la classe DBWrapper che si occupa di interagire con la base di dati e la classe TAGrabber che a sua volta si occupa della interazione con la sorgente dati esterna effettuando le connessioni HTTP necessarie a valorizzare la lista con i dati di interesse.

[5] Le API sono un insieme di procedure messe a disposizione per espletare compiti specifici.

[6] Per maggiori dettagli visitare https://jsoup.org/

[7] Uniform Resource Locator (URL) è una sequenza di caratteri che identifica l'indirizzo di una risorsa in internet. Per i dettagli consultare https://tools.ietf.org/html/rfc3986#section-1.1.3.

[8] HyperText Transfer Protocol è un protocollo a livello applicativo usato per la trasmissione di dati sul web tipicamente in architettura client-server.

La classe principale carica le risorse dalla base dati in cui vengono scritti i dati, comprensivi di URL da contattare, nella fase di setup dell'applicazione.

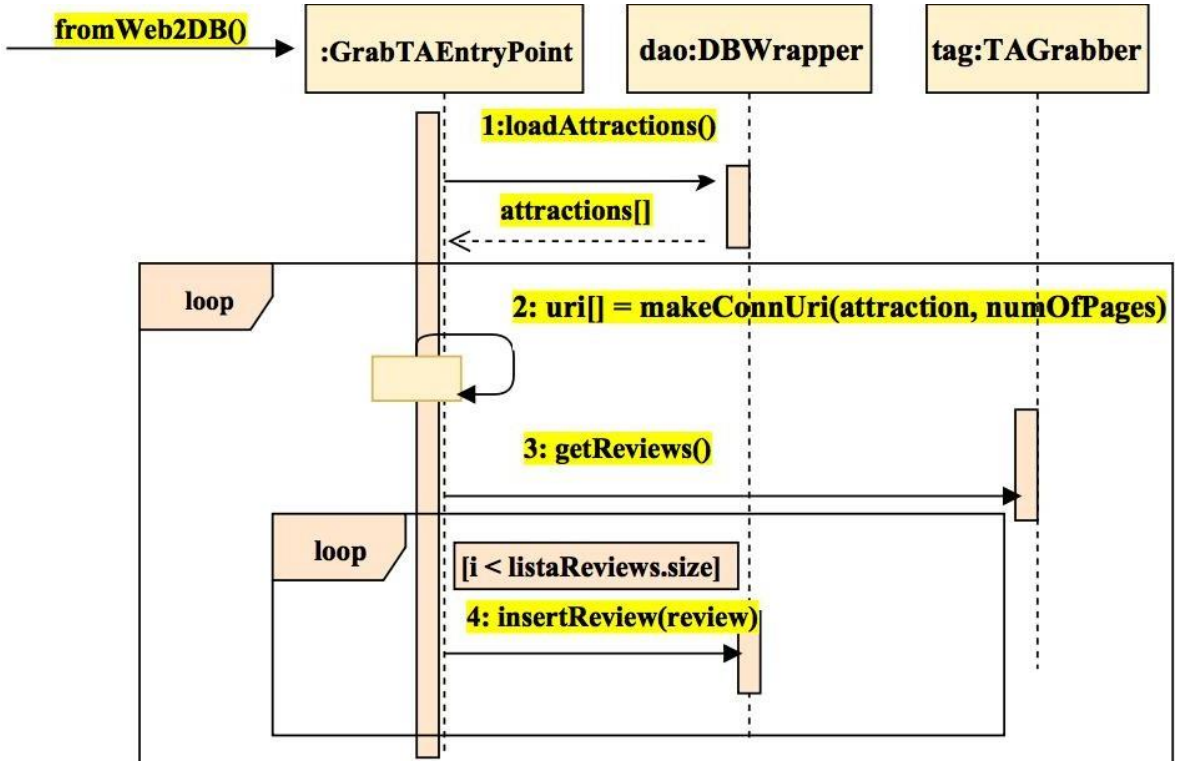


Figura 5: SSD - Scraping: dal web alla base dati

4. Preparazione del dataset

I dati salvati nella macro fase di *scraping* necessitano adeguata preparazione per la ricerca di Text Mining, a questo scopo è stata creata la tabella denominata CORPUS.
La procedura di valorizzazione è gestita da un modulo dedicato del software Diogenes. Il sottosistema « Corpus Manager » mediante opportune interrogazioni e trasformazioni inserisce le recensioni nella tabella CORPUS. In questa tabella le informazioni sono già nella forma adatta alla creazione del corpus ovvero in formato stringa.
Il file di testo con il corpus da studiare viene creato con una semplice interrogazione SQL[9] scegliendo il tipo di formato adatto alla fase successiva che può essere condotta con IRaMuTeQ o TalTaC2.
La preparazione e la valorizzazione della tabella CORPUS oltre a rendere più efficiente la procedura di creazione del corpus, dal momento che consente di evitare *join* per estrarre i dati, permette di mantenere separata la logica del modulo di *data scraping* da quella del modulo deputato alla gestione del corpus.
Il modulo « Corpus Manager » legge i dati dalle tabelle valorizzate tramite la procedura di *scraping* e inserisce solo ed unicamente i dati di interesse per lo studio che si vuole realizzare.
I dati grezzi scaricati dal web sono sottoposti alla procedura di pulizia necessaria prima della creazione del corpus. Il diagramma di attività in Figura 6 sintetizza le operazioni sopra descritte.

[9] Structured Query language (SQL) linguaggio per interrogare basi di dati relazionali.

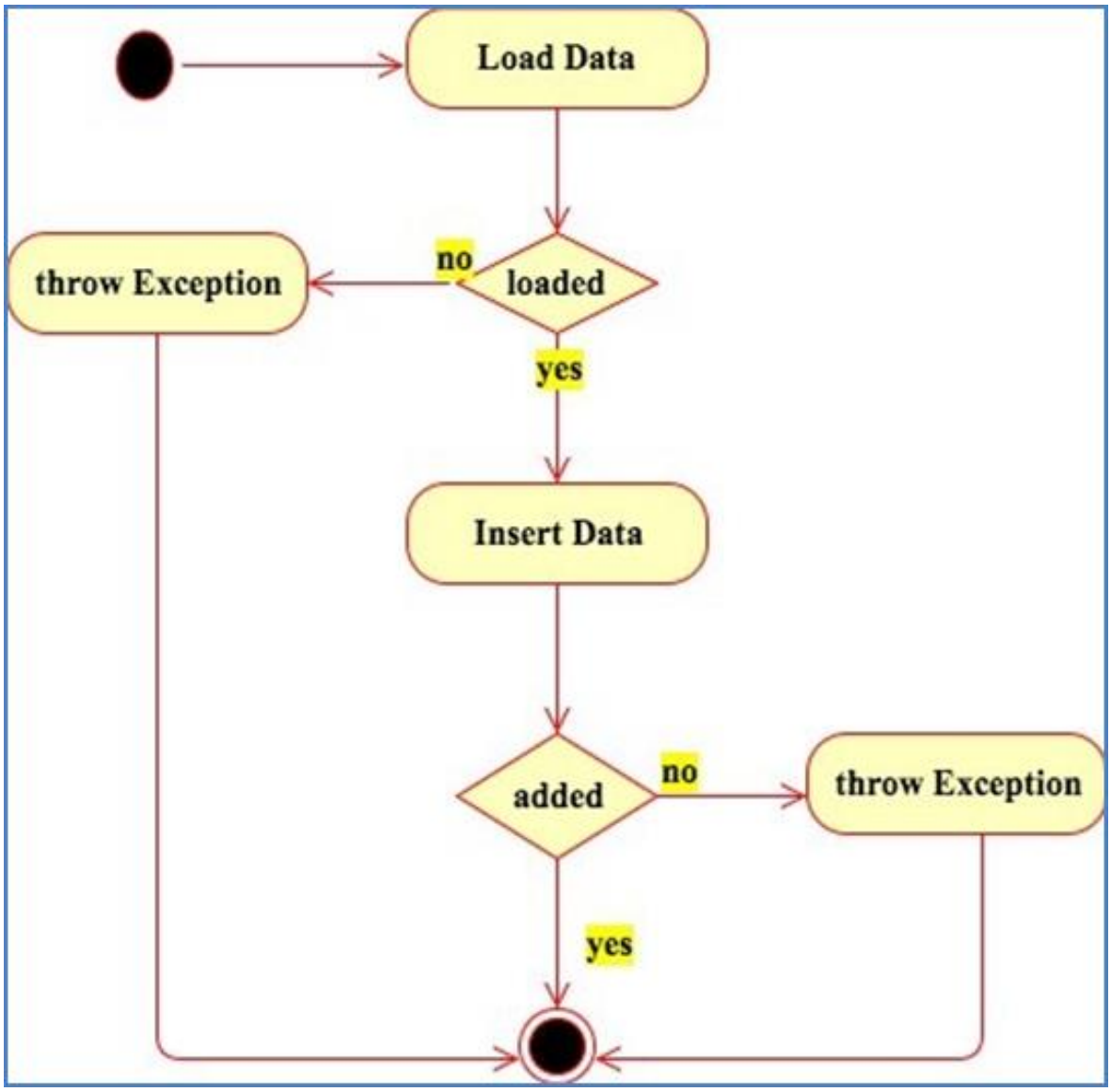


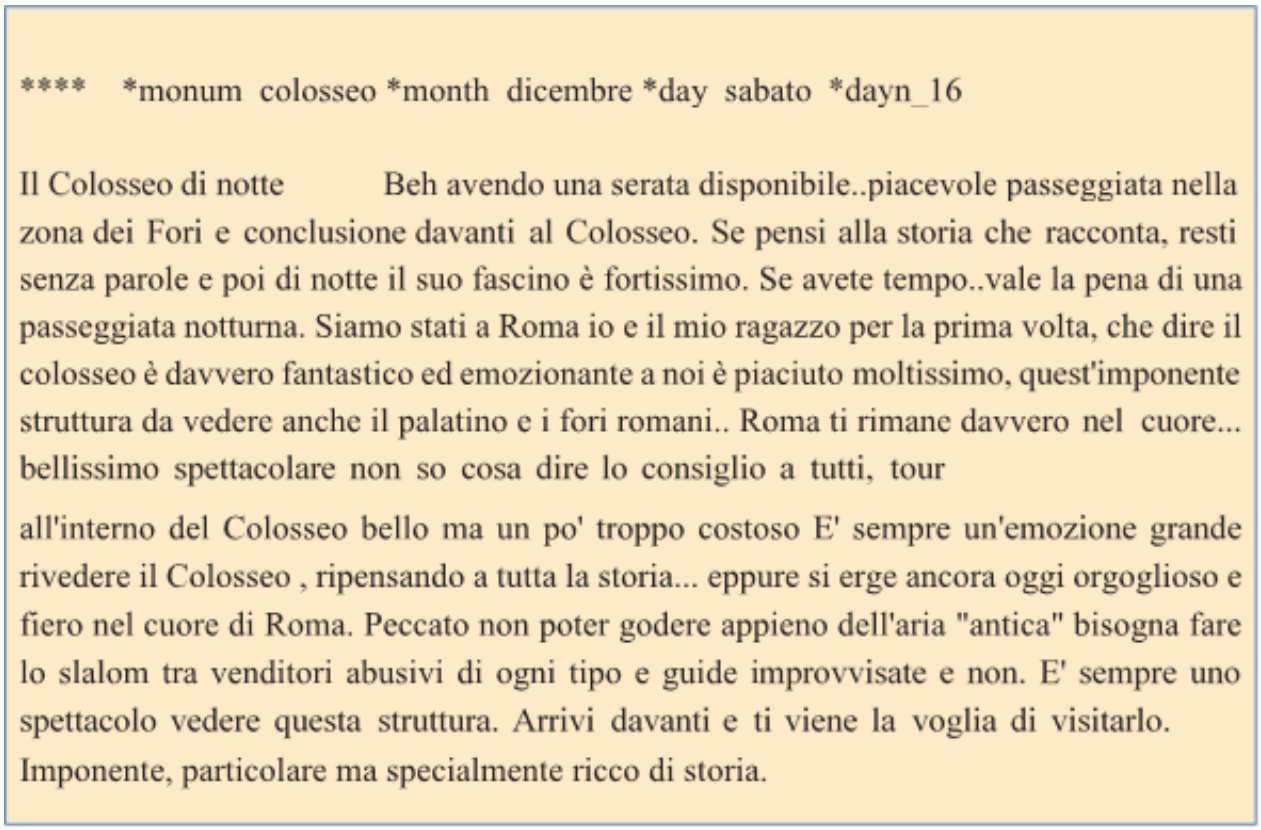
Figura 6 Activity Diagram - preparazione dei dati

5. Creazione del corpus

La terza macro fase si occupa della creazione del corpus secondo le specifiche dettate dall'obiettivo della ricerca, aderendo nel contempo al formato dati richiesto dal software utilizzato.
Prima di passare alla fase di analisi è possibile effettuare ulteriori manipolazioni su tutti i segmenti: nei testi inseriti dagli utenti possiamo trovare riferimenti ai luoghi di cui scrivono come ad esempio "piazza navona" e "fontana di trevi". Tali parole appaiono chiaramente in qualità di toponimi ed è necessario cercare queste sequenze di parole e sostituirle per farle interpretare in modo corretto durante l'analisi lessicale.
Questo processo è stato realizzato utilizzando le espressioni regolari[10] che sono ampiamente

[10] Una espressione regolare è una funzione che prende in ingresso una stringa e restituisce un valore booleano (vero/falso). Se la stringa è conforme al criterio stabilito (pattern) la funzione restituisce vero altrimenti restituisce falso.

supportate in Java. Le espressioni regolari sono di fondamentale importanza nel processo di lessicalizzazione al fine di individuare le occorrenze identificate dall'analista e renderle una sola unità lessicale. Nel modulo di Diogenes deputato alla creazione del corpus è stata implementata una funzionalità per applicare la trasformazione. In questo modo le occorrenze composte da più parole, considerate dall'analista una sola unità lessicale, vengono unificate.



**** *monum colosseo *month dicembre *day sabato *dayn_16

Il Colosseo di notte Beh avendo una serata disponibile..piacevole passeggiata nella zona dei Fori e conclusione davanti al Colosseo. Se pensi alla storia che racconta, resti senza parole e poi di notte il suo fascino è fortissimo. Se avete tempo..vale la pena di una passeggiata notturna. Siamo stati a Roma io e il mio ragazzo per la prima volta, che dire il colosseo è davvero fantastico ed emozionante a noi è piaciuto moltissimo, quest'imponente struttura da vedere anche il palatino e i fori romani.. Roma ti rimane davvero nel cuore... bellissimo spettacolare non so cosa dire lo consiglio a tutti, tour

all'interno del Colosseo bello ma un po' troppo costoso E' sempre un'emozione grande rivedere il Colosseo , ripensando a tutta la storia... eppure si erge ancora oggi orgoglioso e fiero nel cuore di Roma. Peccato non poter godere appieno dell'aria "antica" bisogna fare lo slalom tra venditori abusivi di ogni tipo e guide improvvisate e non. E' sempre uno spettacolo vedere questa struttura. Arrivi davanti e ti viene la voglia di visitarlo. Imponente, particolare ma specialmente ricco di storia.

Figura 10: Operazioni sui dati: il corpus

In Figura 10 osserviamo, a scopo esemplificativo, una recensione relativa al Colosseo nella forma richiesta da IRaMuTeQ: i metadati presenti nella prima riga possono essere inseriti o meno dall'utente sulla base delle sue esigenze per condurre analisi specifiche su periodi tempo o minumenti differenti.

6. Conclusioni

Diogenes è stato utilizzato in diversi progetti e ricerche, la sua natura modulare consente di inserire ulteriori funzionalità in base alle necessità dei ricercatori (Stancampiano, 2018a ; Stancampiano, 2018b). Inoltre è possibile progettare e realizzare connessioni con software open source già presenti e che danno la possibilità di scaricare dati testuali provenienti da altri *social network* in modo da realizzare corpus più ampi su tematiche specifiche.

References

Elmasri R., Navathe S. B. (2011). *Sistemi di Basi di Dati.* Milano, Torino: Pearson Italia.

Larman C. (2005). *Applicare UML e i Pattern. Analisi e progettazione orientata agli oggetti.* Luca Cabibbo (a cura di), Pearson Education Italia.

Sommerville I. (2011). *Ingegneria del software.* Pearson Italia.

Stancampiano S. (2018a). Gestire i beni culturali con I Big Data. In Domenica F. Iezzi, Livia Celardo and Michelangelo Misuraca (a cura di), Roma, UniversItalia, *JADT' 18. Proceedings of the 14th International Conference on the Statistical Analysis of Textual Data*, pp. 748-754.

Stancampiano S. (2018b). Misurare, monitorare e governare le città con I Big Data. In Enrico Di Bella, Filomena Maggino and Marco Trapani (a cura di), Fiesole (FI), Genova: UniversityPress, *Libro degli abstract. AIQUAV 2018. Atti del V Convegno dell'Associazione Italiana per gli studi sulla Qualità della Vita. Fiesole (FI), 13-15 Dicembre 2018*, pp. 114-116.

Stancampiano S. (2019a). The monitoring of cultural heritage in real time using Social Media. . In Leonardo Salvatore Alaimo, Alberto Arcagni, Enrico Di Bella, Filomena Maggino and Marco Trapani (a cura di), Fiesole (FI), Genova: UniversityPress, *Libro dei contenuti brevi. AIQUAV 2019. Atti del VI Convegno dell'Associazione Italiana per gli studi sulla Qualità della Vita. Fiesole (FI), 12-14 Dicembre 2019, pp. 241-247*.

Stancampiano S. (2019b). Matera 2019 Text Mining dei Social Network. In Leonardo Salvatore Alaimo, Alberto Arcagni, Enrico Di Bella, Filomena Maggino and Marco Trapani (a cura di), Fiesole (FI), Genova: UniversityPress, *Libro dei contenuti brevi. AIQUAV 2019. Atti del VI Convegno dell'Associazione Italiana per gli studi sulla Qualità della Vita. Fiesole (FI), 12-14 Dicembre 2019, pp. 277-284*.