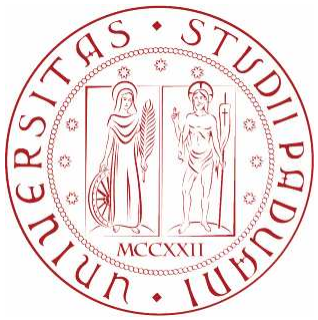


UNIVERSITA' DEGLI STUDI DI PADOVA
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Corso di laurea triennale in Ingegneria Informatica
TESI DI LAUREA

STUDIO SULL'ANNOTAZIONE DI BRANI MUSICALI BASATO SULLA COMBINAZIONE DEL CONTENUTO E DELL'INFORMAZIONE SEMANTICA

Laureando: Paolo Cadorin matricola 543956-IF

Relatore: Dott. Giorgio Maria Di Nunzio

31 marzo 2011

Anno Accademico 2010-2011

alla mia famiglia
ai miei colleghi
ai miei amici

Indice

Indice delle figure

Sommario

Le informazioni che riguardano un brano musicale o un artista sono molteplici e possono essere rintracciate da qualsiasi utente in modo facile e veloce nel Web. Molti siti Internet specializzati nella musica o nell'informazione riportano notizie, discussioni e recensioni dove chiunque può avviare delle ricerche. I siti più conosciuti e utilizzati per questo tipo di navigazione sono:

- Wikipedia che contiene biografie e notizie su canzoni e artisti;
- Last.fm che raccoglie i tag inseriti dagli utenti, le immagini, la discografia degli artisti e le informazioni sulle singole canzoni;
- Amazon che raduna le recensioni dei brani, degli album e degli artisti;
- Youtube che rende possibile la visualizzazione dei videoclip musicali, dei video dei concerti e delle interviste agli artisti, oltre che i commenti che qualsiasi utente può annotare;
- altre fonti quali Blog e siti di discussioni che gli utenti del web intrattengono e sono riferiti ad un singolo artista, un album, un brano o ad una corrente musicale; siti di proprietà delle riviste di musica e degli artisti stessi nei quali è possibile consultare biografie, blog, immagini e video musicali.

Lo scopo di questo progetto è quello di ricercare, partendo da una lista predefinita di canzoni e autori, informazioni sia di carattere strettamente musicale, sia di carattere biografico e di rielaborarle per trovare quante più uguaglianze e relazioni per accomunare canzoni e artisti e nello stesso tempo trovare quali siano le informazioni che rende univoco e identificabile tra tutti l'artista o la canzone.

L'obiettivo di questo elaborato è quindi quello di ricercare, scaricare e rielaborare queste informazioni, spesso non eterogenee, in modo da trovare le possibili relazioni e le disuguaglianze tra i vari artisti e tra le canzoni.

Le fasi principali del progetto sono:

- studio della letteratura e delle pubblicazioni che trattano questo argomento. Conoscenza e indagine delle strutture interne e di formazione delle pagine Web nelle quali sono presenti le informazioni;
- raccolta dal Web dei dati ritenuti importanti per il nostro l'elaborato;
- estrapolazione delle porzioni di testo di nostro interesse;
- elaborazione dei singoli file e dell'insieme dei file di testo;
- studio dei risultati ottenuti e conclusioni finali.

La lista di partenza, sotto forma di elenco canzone-autore è stata estratta dalla classifica che viene stilata periodicamente dalla rivista Rolling Stone e che riguarda le migliori canzoni mai scritte e pubblicate, aggiornate al 2010. Il nostro studio prende in considerazione le prime 500 canzoni della suddetta lista che risulta molto vasta musicalmente e spazia dai Beatles, Queen e Abba fino ai più moderni Jay-Z, 50 Cent e Rihanna.

Capitolo 1

Introduzione

Nel Web si trovano milioni di notizie riguardanti ormai una infinità di argomenti, tra i quali la musica in generale, gli autori, i compositori e ad altre informazioni associate alla musica. Quando si cerca un artista o un singolo brano musicale la precisione della richiesta, basata sui metadati, ci permette di avere una buona o una pessima risposta alla domanda. Se si ottengono molte soluzioni alla richiesta significa che la nostra pretesa era troppo vasta per trovare quello che si stava realmente cercando. Trovando le relazioni e gli aspetti singolari riferiti ai brani e agli artisti si potrà così ottimizzare la ricerca e il risultato finale sarà migliore.

Lo scopo di questo elaborato è quello di costruire un componente di un sistema automatico di reperimento di informazioni musicali che raccolga dai siti Web di informazione musicale alcune notizie, siano esse biografiche o relative alla musica, elaborandole in maniera automatica permettendo così di trovare relazioni che comprendono più cose o aspetti peculiari che rendono un brano musicale o un artista diverso ed identificabile univocamente rispetto ad altri.

1.1 Sistemi di reperimento delle informazioni

Verranno di seguito descritte le tecniche utilizzate nell'informatica per affrontare il problema del reperimento delle informazioni sia riguardanti la musica che di qualsiasi altro tipo. Nella sezione 1.2 verrà descritto l'Information Retrieval, nella sezione 1.3 i suoi modelli, nella sezione 1.4 gli sviluppi attuali e futuri, mentre nelle sezioni 1.5 e 1.6 viene definito il Music Information Retrieval e le sue tecniche.

1.2 Che cosa è l'Information Retrieval

L'Information Retrieval (IR) **[1]** è l'insieme delle tecniche utilizzate per il recupero mirato dell'informazione in formato elettronico. Per "informazione" si intendono tutti i documenti, i metadati e i file presenti all'interno di banche dati o nel web. Il termine è stato coniato da Calvin Mooers **[2]** alla fine degli anni '40 del Novecento, ma oggi è usato quasi esclusivamente in ambito informatico.

L'IR si occupa della rappresentazione, memorizzazione e organizzazione dell'informazione, al fine di rendere agevole all'utente il soddisfacimento dei propri bisogni informativi. Data una collezione di documenti e un bisogno informativo dell'utente, lo scopo di un sistema di IR è di trovare informazioni che potrebbero essere utili, o rilevanti, per l'utente stesso. Rispetto alla teoria classica delle basi di dati (DBMS), l'aspetto principale non è la ricerca di dati ma la ricerca di informazioni.

Il settore dell'Information Retrieval ha avuto dall'inizio degli anni 1990 la sua esplosione grazie alla diffusione del Web, moltiplicandone l'interesse. Il Web infatti non è altro che un'enorme collezione di documenti, sui quali gli utenti vogliono ricercare le informazioni.

Uno dei problemi dei sistemi di IR è la rappresentazione dei bisogni informativi degli utenti. Un sistema di IR gestisce raramente dati con una struttura ben definita e per tale motivo potrebbe restituire, in risposta ad una richiesta degli utenti, oggetti non propriamente esatti.

Per recuperare l'informazione, i sistemi IR usano i linguaggi di interrogazione quasi sempre basati su comandi testuali. Due concetti sono di fondamentale importanza: query ed oggetto:

- Le query sono stringhe di parole che rappresentano l'informazione richiesta. Vengono digitate dall'utente in un sistema IR;
- Un oggetto è un'entità che mantiene al suo interno le informazioni. Un documento di testo, è un oggetto di dati.

Una tipica ricerca di IR ha come input un comando dell'utente. La sua query viene così messa in relazione con gli oggetti presenti nell'indice costruito dal sistema IR. Spesso i documenti stessi non sono mantenuti o conservati direttamente nel sistema IR, ma vengono rappresentati da loro surrogati.

Formalmente un modello di IR è una quadrupla (D, Q, F, R), dove:

- D è un insieme di viste logiche dei documenti della collezione;
- Q è un insieme di query dei bisogni informativi dell'utente;
- F è un sistema per modellare documenti, query e le relazioni tra loro;
- $R(qi, dj)$ è una funzione di ranking che associa un numero reale ad una query qj e un documento dj, definendo un ordinamento tra i documenti con riferimento alla query qj.

1.3 Modelli di Information Retrieval

Sono presenti diversi tipi di IR che in ogni caso derivano direttamente da questi modelli classici:

- il modello booleano
- il modello vettoriale
- il modello probabilistico

1.3.1 Modello booleano

E' il modello più semplice e si basa sulla teoria degli insiemi e dell'algebra booleana. Storicamente è stato il primo ad essere utilizzato ed è stato il più impiegato per decenni. In questo modello i documenti vengono rappresentati come insiemi di termini mentre le query vengono dichiarate come espressioni booleane cioè come un elenco di termini connessi dagli operatori booleani AND, OR e NOT. La strategia di ricerca è basata su un criterio di decisione binario, senza alcuna nozione di grado di rilevanza. Con questa tecnica un documento viene considerato rilevante o non rilevante.

1.3.2 Modello vettoriale

In questo modello viene assegnato un peso, indicato con un numero reale, ad ogni termine e ad ogni query. I documenti e le query vengono quindi rappresentati come vettori in uno spazio n-dimensionale, dove n è il numero di termini indicizzati. La ricerca viene svolta calcolando il grado di similarità tra il vettore che rappresenta la query e i vettori che rappresentano ogni singolo documento.

Una metrica di similarità è una funzione che calcola il grado di similitudine tra due vettori. Grazie all'uso di una metrica di similarità tra la query ed ogni documento e' possibile ordinare i documenti dal più rilevante al meno, impostando una soglia al di sopra della quale respingere i documenti. In questo modo i documenti con maggior grado di similarità con la query hanno più probabilità di essere rilevanti per l'utente.

Il grado di similarità viene quantificato utilizzando il coseno dell'angolo tra i due vettori, in modo tale da dare un risultato migliore alle coppie con maggiore somiglianza.

1.3.3 Modello probabilistico

Data una query *q* e un documento *d*, il modello stima la probabilità che l'utente consideri il documento rilevante come conseguenza solamente dalla query e del

modo in cui il documento è rappresentato. La specifica della query definisce le caratteristiche della risposta ideale, in modo tale che data una query esista sempre un insieme di documenti che costituiscono tale risposta ideale. Il problema principale da affrontare è quello di capire quali sono tali caratteristiche che riguardano la distribuzione dei termini.

1.4 Sviluppi dell'Information Retrieval

L'Information Retrieval è nata per gestire collezioni statiche e ben conosciute: OPAC, enciclopedie ecc. Quando la collezione di riferimento diventa il Web, le cose cambiano completamente in quanto la collezione diventa dinamica e molto variabile nel tempo, le dimensioni sono enormi, i documenti sono protetti e accessibili tramite password.

Negli ultimi anni sta emergendo la necessità di applicare tecniche di IR anche a dati semistrutturati come i documenti XML, con la tecnica Structured Information Retrieval (SIR). Oltre a questo si sta lavorando per assecondare alle richieste nate in questi ultimi anni come il problema della geolocalizzazione, delle mappe e dei dispositivi mobili. Molto attuale è il reperimento delle informazioni multimediali che verrà trattato nella sezione 1.5.

1.5 Music Information Retrieval

Music Information Retrieval (MIR) **[3]** è la scienza che si occupa del recupero di informazioni musicali. Essa comprende la ricerca delle similarità tra brani musicali avvalendosi della tecnica di pattern matching; l'identificazione automatica e il riconoscimento musicale, la classificazione, il clustering e la modellazione della musica.

Negli ultimi anni si sta sviluppando velocemente questo campo di ricerca, cercando di superare i metodi utilizzati in precedenza che si basavano sulla semplice indicizzazione e ricerca dei metadati testuali. Le difficoltà riscontrate in questo campo riguardano tra le altre la complessità della rappresentazione musicale, la difficoltà di acquisizione dei diritti d'autore per le varie composizioni e l'assenza di standard di riferimento tra i gruppi di lavoro.

1.6 Tecniche di Music Information Retrieval

I metodi di Music Information Retrieval possono essere suddivisi nei seguenti modi:

- Analytic/Production systems [AP]
- Locating systems [LS]

1.6.1 Metodo analityc/production system (AP)

Questo metodo è caratterizzato da un alto grado di complessità ed è stato pensato per utenti specialisti del settore quali musicologi, compositori e trascrittori che utilizzano tale sistema per specifici compiti di analisi teorica e produzione di brani musicali. Lo sviluppo di questo metodo viene generalmente condotto riducendo il problema della rappresentazione dell'informazione musicale a problematiche già note per le quali esistono soluzioni consolidate applicabili al Music Retrieval. Si cerca di modellare la rappresentazione musicale ad un linguaggio di programmazione, sviluppando appositi linguaggi funzionali e di ricondurre tutto ad espressioni regolari tramite modelli di codifica ad hoc che comunque hanno una elevata complessità concettuale ed implementativa.

1.6.2 Metodo locating system (LS)

Questo procedimento è utilizzato specialmente nei sistemi per la localizzazione e l'identificazione di brani musicali ed è rivolto ad utenti non esperti. Questa tecnica viene utilizzata per risolvere semplici interrogazioni come trovare tutte le opere di un autore o dato un testo trovare l'opera che lo contiene. In questo modo l'utente generico può trovare le informazioni che cerca in maniera facile e veloce. Nel caso in cui la richiesta sia più complessa, del tipo data una linea melodica trovare l'opera corrispondente deve essere risolta tramite la tecnica dell'Incipit Index ovvero una rappresentazione testuale dell'inviluppo melodico della parte iniziale del tema principale. Tale sistema risulta incompleto dal punto di vista della rappresentazione in quanto la ricerca si basa sul solo tema principale. Ha comunque buoni effetti nell'applicazione pratica avendo una ridotta complessità di formulazione della query da parte dell'utente. E' completamente trascurata l'intonazione delle note nella melodia, dando una probabilità minima di errore nella formulazione della query. La scelta di privilegiare l'inviluppo melodico su altri aspetti della rappresentazione, per esempio l'aspetto ritmico, è giustificata da studi in ambito psicoacustico.

Capitolo 2

Analisi dei requisiti

La fase iniziale del mio elaborato si è focalizzata sulla ricerca e sullo studio delle informazioni riguardanti alcune parti del vasto mondo del Music Information Retrieval. In questo stadio ho dovuto approfondire la letteratura che sta alla base di questo argomento ed ho trovato grande aiuto nel consultare le pubblicazioni relative al tema, in particolare **[4]** nella quale vengono trattati approfonditamente molti di questi aspetti.

2.1 Siti Web per la ricerca delle informazioni musicali

Durante il lavoro di analisi dei requisiti ho esaminato i principali siti di informazione musicale presenti nel Web per valutare il tipo di contenuti e quali informazioni erano più utili per la successiva fase operativa di ricerca descritta nel capitolo 3 dedicato all'implementazione.

2.1.1 Wikipedia

E' un'enciclopedia multilingue online e gratuita nata dal progetto intrapreso da una organizzazione statunitense chiamata Wikimedia Foundation **[5]**. È pubblicata in oltre 270 lingue differenti, di cui 180 attive, e contiene voci sia sugli argomenti propri di una tradizionale enciclopedia che su argomenti di attualità.

Lo scopo per cui è nata, e tuttora presente, è quello di creare e distribuire un'enciclopedia libera e ricca di contenuti, nel maggior numero di lingue possibili.

Questa sua prerogativa l'ha resa uno dei dieci siti più visitati al mondo con circa 60 milioni di accessi totali al giorno.

La versione più completa e aggiornata è quella inglese seguita dalla versione italiana, tedesca e spagnola. Uno dei principi alla base di Wikipedia **[6]** è il suo punto di vista neutrale, secondo il quale le opinioni riportate su personaggi e opere letterarie vengono descritte senza tentare di determinarne una verità oggettiva.

Le varie edizioni di lingua diversa sono sviluppate indipendentemente l'una dall'altra e non sono vincolate ai contenuti presenti nelle altre ma sono tenute unicamente al rispetto delle linee guida generali del progetto. Tuttavia le voci e i contenuti multimediali sono spesso condivisi tra le varie edizioni, i primi grazie alle traduzioni, i secondi grazie progetto condiviso chiamato Wikimedia Commons **[7]**.

2.1.2 Last.fm

Last.fm **[8]** è una radio su internet e un social network, la cui caratteristica principale è quella di costruire un dettagliato profilo per ogni utente, gruppo, artista, album o canzoni che si vanno a creare. Le statistiche sono aggiornate in tempo reale: in particolare, il sito prevede statistiche di artisti e tracce più ascoltate dell'ultima settimana, degli ultimi tre, sei o dodici mesi.

Le etichette discografiche e gli artisti stessi sono incoraggiati a promuovere la loro musica su Last.fm, perché in questo modo verrà proposta agli utenti che hanno espresso preferenze simili, grazie al suo sistema di raccomandazioni. Last.fm ha una collezione di 100.000 canzoni e rende disponibili demo di 30 secondi per ogni brano.

Per gli abbonati sono disponibili anche playlist e tag. Le playlist permettono di selezionare elenchi di canzoni da ascoltare, senza limiti temporali, nell'ordine stabilito. Il tag radio si basa invece sulle "etichette" che l'utente ha assegnato ad artisti, album o specifiche canzoni

2.1.3 Amazon

Amazon.com **[9]** è una compagnia di commercio elettronico statunitense ed è stata tra le prime a vendere merci su Internet dalla fine degli anni Novanta. Amazon.com iniziò come libreria online, ampliando ben presto la gamma dei prodotti venduti a DVD, CD musicali, software, videogiochi, prodotti elettronici e altro offrendo una scelta molto maggiore di qualsiasi grande negozio di vendita per corrispondenza. Amazon ha creato poi altri siti in Canada, Regno Unito, Germania, Austria, Francia, Italia, Cina e Giappone e spedisce i prodotti in tutto il mondo.

Su Amazon.com ricercando l'album o una pubblicazione di un artista è possibile leggerne i commenti, che possono essere sia positivi che negativi, rilasciati dagli utenti dopo aver acquistato on-line tali opere.

2.1.4 Youtube

E’ un sito web che consente la condivisione di filmati caricati dagli utenti e fa uso della tecnologia Adobe Flash per la riproduzione dei suoi contenuti. Il suo scopo è quello di ospitare solamente video realizzati direttamente da chi li carica, malgrado ciò contiene materiale di terze parti caricate senza autorizzazioni come spettacoli televisivi, video musicali e parti di film. Consente l'incorporazione dei propri video all'interno di altri siti web, e si occupa di generare il codice HTML necessario.

All’interno di Youtube **[10]**, l’utente potrà visualizzare i filmati, leggere i commenti relativi a tali video e inserirne di propri, indicando il proprio grado di apprezzamento. La durata dei video è stata bloccata a 15 minuti, con un massimo di 2 GB di memoria disponibile. Dal 9 dicembre 2010, Youtube ha concesso ai soli utenti che hanno caricato video senza aver mai infranto le regole sul copyright l'upload illimitato, che era riservato ai solo soci partner.

2.1.5 Altre fonti

Tra le altre fonti non citate fino ad ora trovano particolare spazio i blog. Può essere descritto come un luogo virtuale dove gli utenti possono dialogare assieme e in questo modo esprimere liberamente la propria opinione. È un sito, gestito in modo autonomo, dove si tiene traccia dei pensieri avvicinandosi molto ad una sorta di diario personale.

In questo luogo si possono pubblicare notizie, informazioni e storie di ogni genere, aggiungendo, se si vuole, anche dei link di proprio interesse. Tramite il blog si viene in contatto con persone lontane fisicamente ma spesso vicine alle proprie idee e ai propri punti di vista. Con esse si condividono i pensieri e le riflessioni su diversi temi.

2.2 Siti utilizzati

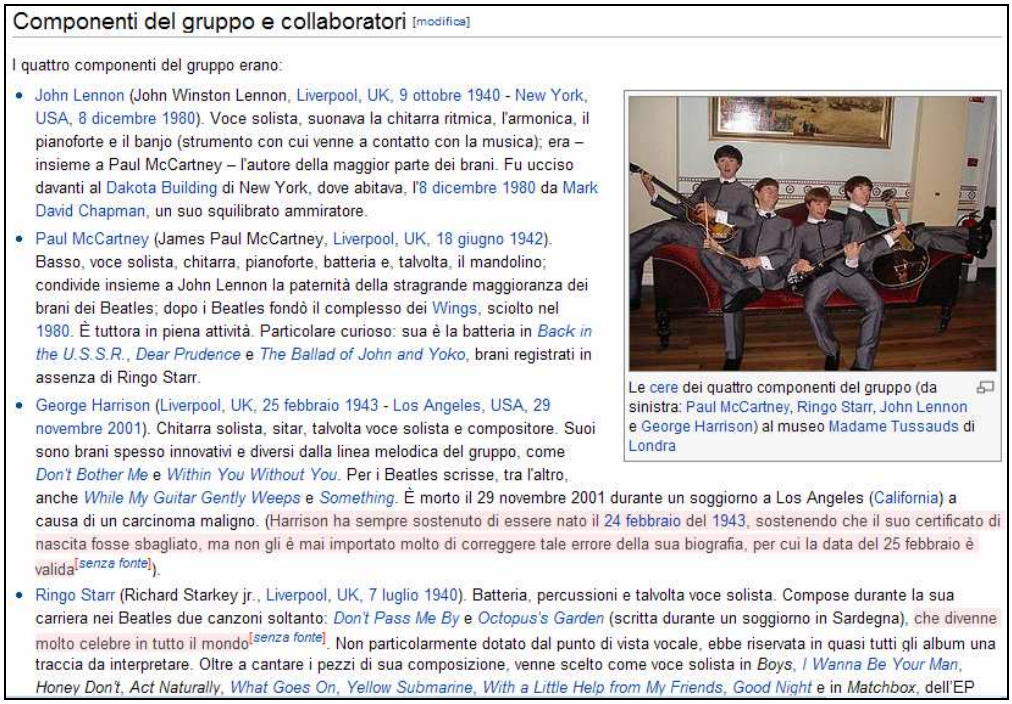
Dopo una riflessione che considerava sia la parte riguardanti le informazioni presenti che la successiva fase implementativi sono giunto alla conclusione che i due siti da utilizzare per il download fossero Wikipedia (sezione 2.2.1) e Last.fm (sezione 2.2.2). Ho scelto Wikipedia in quanto la priorità era quella di ottenere grandi quantità di dati da elaborare sia per gli artisti che per le canzoni della lista e questo sito mi dava garanzie a riguardo. La scelta di Last.fm è stata fatta, oltre che per la possibilità di trovare una vasta quantità di notizie, anche e soprattutto per il fatto di poter reperire, diversamente da altri siti, i tag. Queste etichette, associate a canzoni e ad artisti sono in grado di descrivere in una o al massimo in poche parole il contenuto delle opere o dei personaggi a cui sono associati.

2.2.1 Wikipedia

La struttura di Wikipedia è costruita in modo da avere una sua struttura ripetitiva. Le varie pagine cominciano con una breve descrizione dell'artista o della canzone come nella **Figura 1**.

Figura 1: pagina di Wikipedia dei "The Beatles"
URL: http://it.wikipedia.org/wiki/The_Beatles

Si prosegue con la biografia o le biografie dei componenti delle band che hanno composto un determinato brano o del periodo storico nel quale una canzone è stata scritta come riportato in **Figura 2.**



Componenti del gruppo e collaboratori [modifica]

I quattro componenti del gruppo erano:

- John Lennon (John Winston Lennon, Liverpool, UK, 9 ottobre 1940 - New York, USA, 8 dicembre 1980). Voce solista, suonava la chitarra ritmica, l'armonica, il pianoforte e il banjo (strumento con cui venne a contatto con la musica); era – insieme a Paul McCartney – l'autore della maggior parte dei brani. Fu ucciso davanti al Dakota Building di New York, dove abitava, l'8 dicembre 1980 da Mark David Chapman, un suo squilibrato ammiratore.
- Paul McCartney (James Paul McCartney, Liverpool, UK, 18 giugno 1942). Basso, voce solista, chitarra, pianoforte, batteria e, talvolta, il mandolino; condivide insieme a John Lennon la paternità della stragrande maggioranza dei brani dei Beatles; dopo i Beatles fondò il complesso dei Wings, sciolto nel 1980. È tuttora in piena attività. Particolare curioso: sua è la batteria in *Back in the U.S.S.R.*, *Dear Prudence* e *The Ballad of John and Yoko*, brani registrati in assenza di Ringo Starr.
- George Harrison (Liverpool, UK, 25 febbraio 1943 - Los Angeles, USA, 29 novembre 2001). Chitarra solista, sitar, talvolta voce solista e compositore. Suoi sono brani spesso innovativi e diversi dalla linea melodica del gruppo, come *Don't Bother Me* e *Within You Without You*. Per i Beatles scrisse, tra l'altro, anche *While My Guitar Gently Weeps* e *Something*. È morto il 29 novembre 2001 durante un soggiorno a Los Angeles (California) a causa di un carcinoma maligno. (Harrison ha sempre sostenuto di essere nato il 24 febbraio del 1943, sostenendo che il suo certificato di nascita fosse sbagliato, ma non gli è mai importato molto di correggere tale errore della sua biografia, per cui la data del 25 febbraio è valida[*senza fonte*]).
- Ringo Starr (Richard Starkey jr., Liverpool, UK, 7 luglio 1940). Batteria, percussioni e talvolta voce solista. Compose durante la sua carriera nei Beatles due canzoni soltanto: *Don't Pass Me By* e *Octopus's Garden* (scritta durante un soggiorno in Sardegna), che divenne molto celebre in tutto il mondo[*senza fonte*]. Non particolarmente dotato dal punto di vista vocale, ebbe riservata in quasi tutti gli album una traccia da interpretare. Oltre a cantare i pezzi di sua composizione, venne scelto come voce solista in *Boys*, *I Wanna Be Your Man*, *Honey Don't*, *Act Naturally*, *What Goes On*, *Yellow Submarine*, *With a Little Help from My Friends*, *Good Night* e in *Matchbox*, dell'EP

Le cere dei quattro componenti del gruppo (da sinistra: Paul McCartney, Ringo Starr, John Lennon e George Harrison) al museo Madame Tussauds di Londra

Figura 2: biografia dei singoli componenti dei "The Beatles"
URL: http://it.wikipedia.org/wiki/The_Beatles

La pagina viene conclusa con la discografia completa dell'artista o con la lista delle canzoni dell'album nel quale è presente il singolo brano come riportato in **Figura 3**.

Discografia [modifica]

La discografia ufficiale si basa sulle edizioni inglesi degli album (spesso venivano modificati e rititolati per l'uscita in USA), che sono alla base delle riedizioni in compact disc. Data la rarità di apparecchi stereofonici, i Beatles e il loro produttore George Martin si applicarono tardi a produrre master stereofonici dei brani. Così i primi quattro album furono pubblicati in mono, e fino al 2009 anche i CD da essi ricavati sono monofonici.

Molti singoli contengono brani di grande importanza e fama non usciti su album. La EMI ha provveduto a rendere reperibili tutti i singoli su CD con due raccolte. Al catalogo ufficiale si aggiungono alcune raccolte che si distinguono dalle altre (mere ricompilazioni di brani già editi) per alcune caratteristiche particolari. Vanno ricordati i due doppi album: 1962-1966 (noto come *The Red Album*) e 1967-1970 (noto come *The Blue Album*) a cui vanno aggiunti i due album *Past Masters, Volume One* e *Past Masters, Volume Two*. In questo modo, con gli album "inglesi" si hanno a disposizione tutte le canzoni dei Beatles non pubblicate su questi.

Il 9 settembre 2009 l'intero catalogo dei Beatles è stato riproposto in versione CD in seguito a un processo di rimasterizzazione digitale durato quattro anni.[168] Le edizioni stereo di tutti i dodici album originali (versione inglese), *Magical Mystery Tour* e una coppia di CD dei Past Masters sono stati riproposti sia individualmente sia in forma di raccolta. Una seconda raccolta comprende tutte le tracce mono.[169]

La mela, logo della Apple

Studio [modifica]

Nella lista degli album inglesi si comprende per tradizione il doppio EP *Magical Mystery Tour*, che in USA uscì come album con l'aggiunta di brani già pubblicati su singolo: tale versione è alla base dell'edizione su compact disc.

Tutti i dischi fino a *Magical Mystery Tour* uscirono su etichetta *Parlophone*. Dal *White Album* in poi uscirono su etichetta *Apple*, di proprietà degli stessi Beatles, distribuita dalla EMI.

- *Please Please Me* - 22 marzo 1963
- *With the Beatles* - 22 novembre 1963
- *A Hard Day's Night* - 10 luglio 1964

Figura 3: discografia del Gruppo "The Beatles"
URL: http://it.wikipedia.org/wiki/The_Beatles

La descrizione è corredata da foto e tabelle riguardanti gli artisti e le canzoni.

2.2.2 Last.fm

In Last.fm si possono trovare informazioni relative a musica, artisti e canzoni. Tali argomenti sono organizzati in sezioni strutturate e omogenee, diversamente da Wikipedia. Durante la ricerca di qualsiasi informazione bisogna tener conto che la parte del protagonista assoluto è attribuita all'artista come si nota nella **Figura 4**.

Figura 4: sezione di Last.fm dedicata ai The Beatles
URL: http:www.lastfm.it/music/The+Beatles

Ad ogni musicista vengono associate alcune sottosezioni, in maniera sistematica dagli utenti dotati di login o dalle stesse etichette discografiche. Cliccando sulle linguette presenti a sinistra del testo principale è possibile accedere ad alcune parti relative all'artista selezionato in precedenza.

In questo modo accedendo ad una di queste sezioni si è certi di trovare quello che si cercava, siano esse immagini, video o notizie. Nella parte dedicata alle immagini quindi si troveranno solamente immagini e nella parte attribuita alla biografia si troveranno solamente le biografie dell'artista o degli artisti che compongono la band.

Alcuni esempi delle sezioni riferite alla band “The Beatles”, sono:

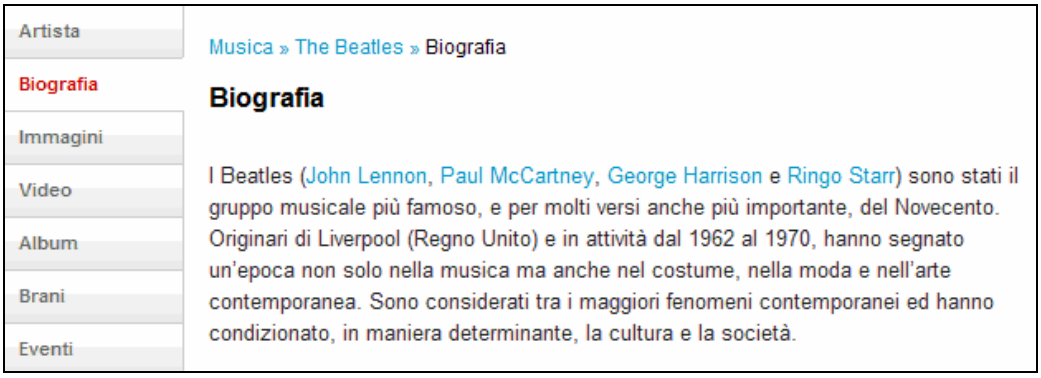


Figura 5: sezione di Lastfm contenente la biografia dei "The Beatles"
URL: http:www.lastfm.it/music/The+Beatles/+wiki

Figura 6: sezione di Lastfm contenente le immagini dei "The Beatles"
URL: http:www.lastfm.it/music/The+Beatles/+images

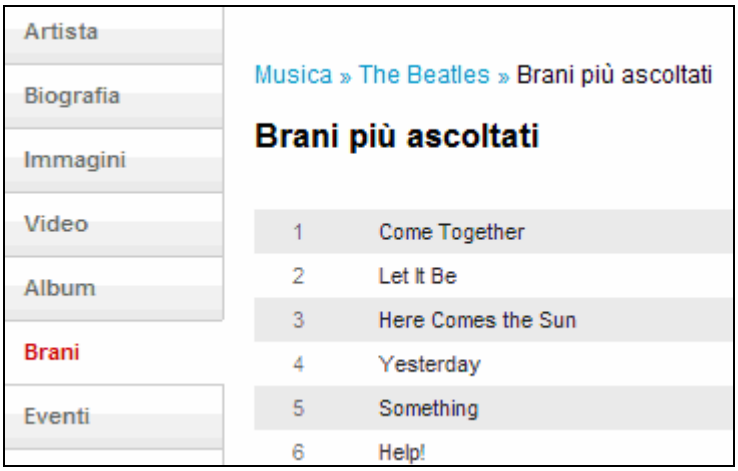


Figura 7: sezione di Lastfm contenete i brani più ascoltati dei "The Beatles"
URL: http:www.lastfm.it/music/The+Beatles/+tracks

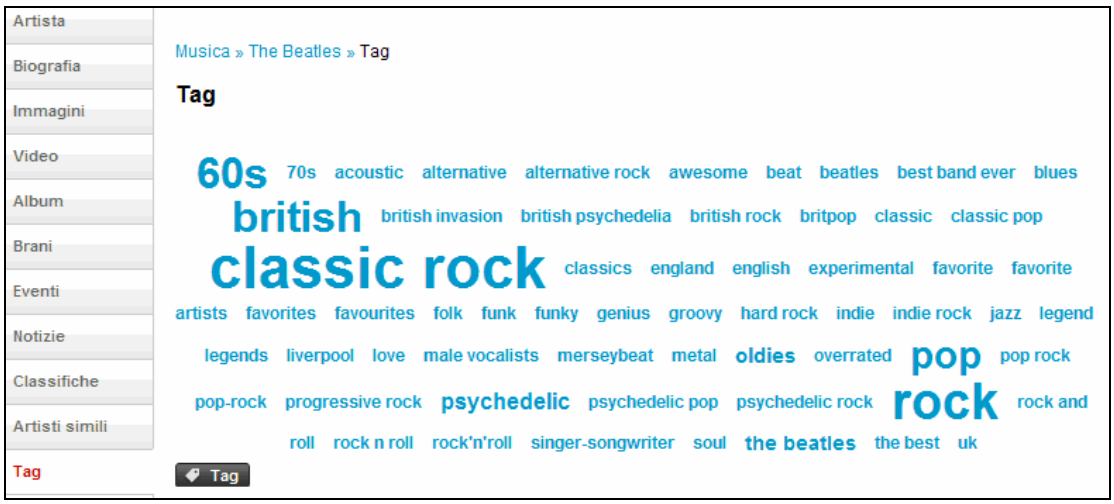


Figura 8: sezione di Lastfm contenete i tag riferiti ai "The Beatles"

URL: http:www.lastfm.it/music/The+Beatles/+tags

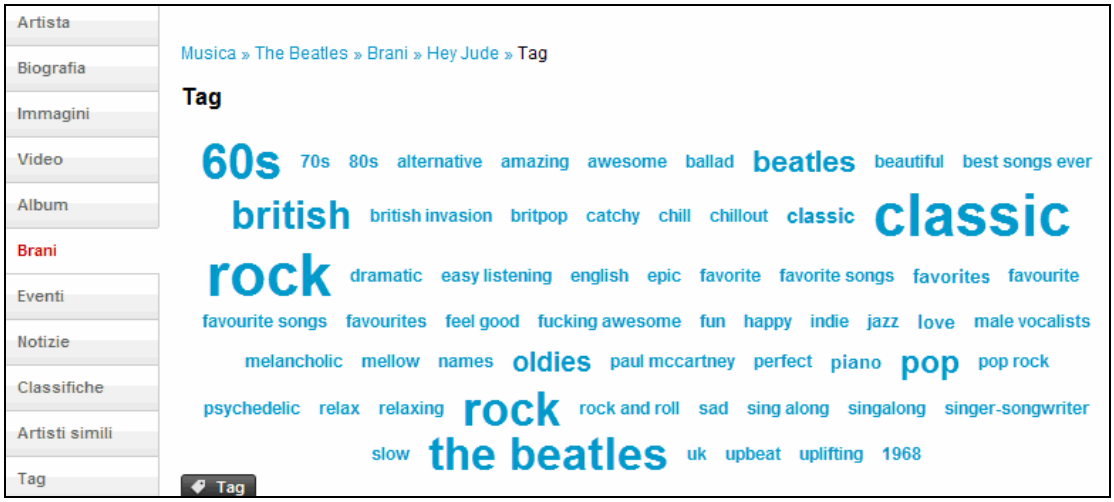


Figura 9: sezione di Lastfm contenente i tag del brano "Hey Jude" dei "The Beatles"

URL: http:www.lastfm.it/music/The+Beatles/_/Hey+Jude/+tags

La parte più interessante riguarda comunque quella riferita ai tag, ossia alle etichette che gli utenti voglio associare all'artista o alla canzone. Queste etichette sono molto utili per la ricerca di artisti e canzoni in Last.fm.

Indicando sulla casella di ricerca un tag è così possibile trovare quali sono le canzoni o gli artisti che assecondano in modo certo la richiesta. Come si può vedere in **Figura 8** e in **Figura 9** i tag sono delle parole che vengono inserite e votate dagli utenti. Un tag cresce di grandezza in relazione ai voti che riceve dagli utenti.

Per questo elaborato è stata fatta la scelta di utilizzare sia le versioni in lingua italiana che quelle in lingua inglese per entrambi i siti utilizzati nella fase di download. In questo modo è possibile valutare le differenze in termini di informazioni presenti tra le diverse versioni.

2.3 Software utilizzato

In questa sezione vengono descritti i principali tipi di software che permettono di scaricare in modo automatico le informazioni presenti nel web. Nella **sezione 2.3.1** viene descritto cos'è un Web Crawler, mentre nella **sezione 2.3.2** cos'è un parser.

2.3.1 Web Crawler

E' un software **[11]** che analizza in un modo metodico e automatizzato la rete in genere per conto di un motore di ricerca. Solitamente acquisiscono una copia testuale di tutti i documenti visitati e le inseriscono in un indice. Un uso estremamente comune dei crawler è nel Web. Sul Web, il crawler si basa su una lista di URL da visitare fornita dal motore di ricerca o da un utente.

Durante l'analisi di un URL, identifica tutti gli hyperlink presenti nel documento e li aggiunge alla lista di URL da visitare. Il processo può essere concluso manualmente o dopo che un determinato numero di collegamenti.

2.3.2 Parser

E' un programma **[12]** che automaticamente analizza un flusso continuo di codice in input per determinare la sua struttura grammaticale rispetto ad una data grammatica formale.

Tipicamente viene utilizzato per riferirsi al riconoscimento di una grammatica e alla conseguente costruzione di un albero sintattico, che mostra le regole utilizzate durante il riconoscimento dall'input; l'albero sintattico viene poi visitato durante l'esecuzione di un compilatore. L'analisi sintattica propriamente detta riceve in input la sequenza dei "*token*" e controlla che essi formino espressioni valide.

Questo lavoro è svolto basandosi su una grammatica libera dal contesto, che ricorsivamente definisce i componenti che determinano un'espressione e l'ordine in cui devono comparire. La fase finale è l'analisi semantica, che trova gli effetti nell'espressione appena validata ed esegue le azioni derivanti.

In questo elaborato è stato utilizzato un parser STAX **[13]**. Il parser STAX è stato creato per risolvere le limitazioni presenti nelle due API di analisi più diffuse, SAX e DOM. L'obiettivo primario delle API STAX è quello di dare al programmatore l'analisi di controllo oltre al fatto di consente la memorizzazione dello stato in modo procedurale.

Capitolo 3

Implementazione

Conclusa la fase di analisi dei requisiti ho cominciato a implementare i programmi che secondo la precedente fase di studio mi avrebbero permesso di ottenere dei risultati. Questo è possibile indicando l'URL della pagina che si vuole visualizzare.

L'impostazione accurata dei parametri all'interno dello STAX parser, ha permesso di cominciare allo scarico automatico delle pagine. La struttura e il codice sorgente di Wikipedia e di Lastfm mi hanno permesso di notare quali erano le parti fisse e le parti variabili degli URL e del testo HTML da scaricare.

3.1 Il parser di Wikipedia

La struttura di Wikipedia, speculare nelle versioni italiana e inglese, è molto intuitiva e utilizza codice HTML puro in tutte le sue parti. La prima parte dell'URL **http://it.wikipedia.org/wiki/** per la versione italiana e **http://en.wikipedia.org/wiki/** per quella inglese rimane costante per tutte le pagine mentre la seconda parte varia in base alla sezione da visualizzare.

Per fare in modo che le pagine siano ritrovate correttamente, qualora presenti, la parte variabile dell'URL deve essere formattato nel modo seguente. Tutti gli spazi sono sostituiti con "_" e le punteggiature cambiate con la relativa codifica ASCII.

Per trovare la pagina relativa al cantante Michael Jackson deve essere utilizzata quindi la stringa "Michael_Jackson" il cui URL risulta essere **http://it.wikipedia.org/wiki/Michael_Jackson** come riportato in **Figura 10**.

Figura 10: pagina di Wikipedia versione italiana del cantante Michael Jackson

Il parser utilizzato per Wikipedia scorre tutto il codice sorgente presente negli URL che vengono indicati ma solamente quello contenuto tra i comandi <p> del codice HTML viene elaborato realmente come riportato in **Figura 11**.

```
<div class="thumbcaption">
<div class="magnify"><a href="/wiki/File:Quincy_Jone
<a href="/wiki/Quincy_Jones">Quincy Jones</a></div>
</div>
</div>
<p>Nel 1978, Michael Jackson interpretò lo <a href="
tesso Jackson, artisti del calibro di <a href="/wiki
</p><i>Off the Wall</i> ottenne un ottimo successo,
>Grammy per il miglior album R&amp;B</a> e ben 8 <a
<p>Dopo la parentesi da solista, Michael ritornò in
```

Figura 11: parte di codice HTML con selezionato la parte di testo che viene analizzata

Le stringhe di testo con i nomi delle canzoni o degli artisti possono essere presenti, assenti o incomplete all'interno del sito. Se le stringhe di testo trovano corrispondenza con le pagine di Wikipedia allora vengono trascritte in un file di testo che conterrà la lista di tutte le pagine presenti.

Nel caso in cui la pagina non fosse presente viene segnalato a video un messaggio nel quale viene indicato che non è stata trovata alcuna corrispondenza e manualmente bisognerà verificare se tale errore è dovuto ad una mancanza reale della pagina o ad un errore nella scrittura dell'URL. Nel primo caso si procederà ad eliminare dal file con gli URL la riga incriminata, mentre nel secondo si correggerà l'URL riportando quello esatto.

Nel caso di più pagine con in comune lo stesso nome ma con vari significati, Wikipedia visualizza una pagina nella quale ci sono i vari significati che quella stringa può avere come riportato in **Figura 12**. Nell'esempio della figura si cercava la canzone "One" interpretata della band U2. Come riportato di seguito la stringa "One" può essere interpretata come un titolo di un film, un codice aeroportuale, un

album dei Bee Gees o un singolo degli U2. Una volta scoperto l'URL esatto bisognerà modificare manualmente il file contenete l'imprecisione indicando "One_(U2)" .

Figura 12: pagina di chiarimento delle ambiguità di Wikipedia versione italiana

Alla fine delle quattro ricerche, due per gli artisti e due per le canzoni nelle due versioni di Wikipedia dovrebbero risultare 1560 file separati, 1000 per i brani e 560 per gli interpreti. I risultati ottenuti riguardanti il numero di pagine realmente trovato sono riportati nella **Tabella 1** .

Tabella 1: numero pagine di Wikipedia (IT e EN) di artisti e canzoni trovate

	Teorico	Wikipedia (IT)	Wikipedia.(EN)	Totale pp. trovate
Artisti	562	182	241	423
Canzoni	1000	356	468	824

3.2 Il parser di Last.fm

La ricerca su Lastfm è stata più complicata, non rispetto alla struttura degli URL ma riguardo la scrittura del codice sorgente presente delle pagine web. Questo sito non è scritto in maniera sintatticamente corretta per un file XML mentre lo è per il formato HTML e ciò non permette al parser, così come utilizzato per Wikipedia, di trovare in automatico le pagine identificate dall'URL di Last.fm.

Il Parser segnala ad ogni tentativo di lettura del codice sorgente che alcune entità presenti in tale codice sono state utilizzate senza essere però preventivamente dichiarate. Per ovviare a questo inconveniente ho dovuto utilizzare del codice che aveva la funzione di scaricare alcune parti delle pagine di Lastfm in locale.

Anche l'URL di Lastfm è formato da una parte fissa e da una parte variabile per indicare le varie sezioni. La versione italiana aveva come parte fissa **http://www.lastfm.it/music/**, mentre quella inglese era **http://www.last.fm/music/**.

Come per Wikipedia, anche in questo caso il file di ingresso al parser di Last.fm deve essere modificato per essere compatibile con la formattazione richiesta dall'URL. In questo caso gli spazi devono essere sostituiti con dei "+"e le punteggiature cambiate con la relativa codifica ASCII, come per Wikipedia.

In Lastfm si possono trovare e visualizzare i tag relativi alle canzoni e agli artisti ed è quindi possibile, indicando nell'URL la pagina che si desidera.

Di seguito vengono indicati gli URL che permettono di visualizzare le notizie relative a John Lennon e al brano "Imagine" nella versione italiana di Last.fm:

- **http://www.lastfm.it/music/John+Lennon** (**Figura 13**)

Figura 13: sezione relativa alla biografia di "John Lennon"

- **http://www.lastfm.it/music/John+Lennon/+tags** (**Figura 14**)

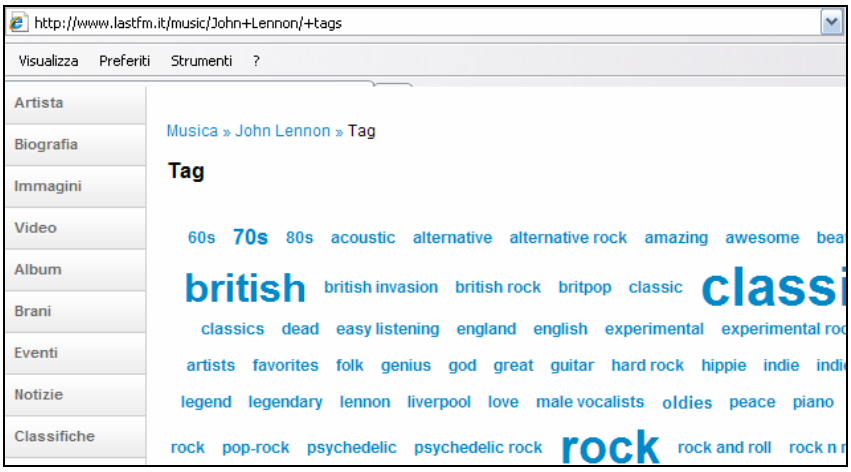


Figura 14: sezione relativa ai tag riferiti a "John Lennon"

- **http://www.lastfm.it/music/John+Lennon/_/Imagine** (**Figura 15**)

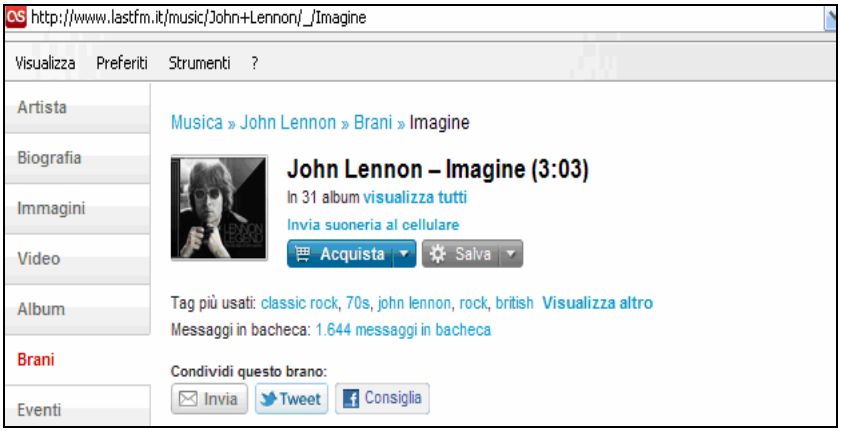


Figura 15: sezione relativa alla notizie relative al brano "Imagine"

- **http://www.lastfm.it/music/John+Lennon/_/Imagine/+tags**

(Figura 16)

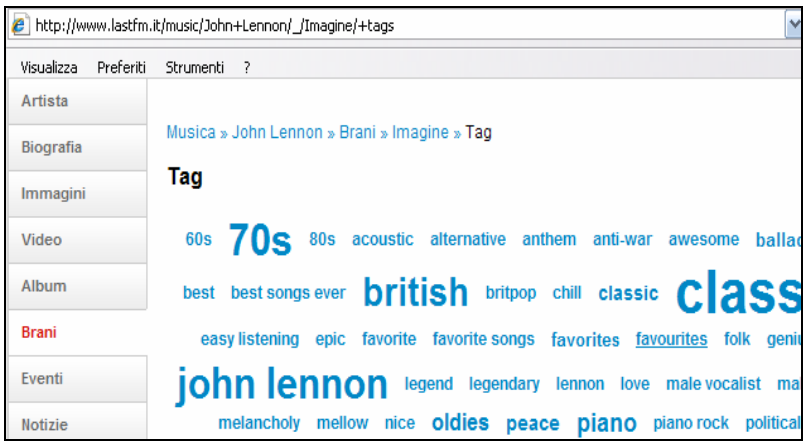


Figura 16: sezione relativa ai tag riferiti al brano "Imagine"

Questi quattro URL, modificati con gli artisti e le canzoni della lista ci permetteranno di visualizzare tutte le pagine e le relative informazioni utili. Alla fine di queste operazioni avremo tutti gli URL che si riferiscono alle descrizioni delle canzoni, degli artisti e i tag riferiti ad entrambi.

Capitolo 4

Elaborazione dei dati

Dopo l'implementazione inizia la fase di elaborazione dei dati nei seguenti modi:

- segmentazione delle parole delle pagine web in singoli termini
- frequenza dei termini nelle singole pagine
- frequenza dei termini dei documenti della collezione
- aggregazione delle parole con la stessa radice
- confronto tra varie sorgenti

4.1 Dati di partenza

I dati utilizzati, estratti dal sito internet della rivista "***Rolling Stone***" **[14]**, sono relativi alla classifica delle canzoni più famose di tutti i tempi. Per questo elaborato sono state utilizzate le prime 500 che qui di seguito riportate.

Per prima cosa ho dovuto separare la lista in due sotto liste contenenti da una parte le sole canzoni (**Figure 17, 18 e 19**) e dall'altra i soli artisti (**Figure 20 e 21**). Dopo la divisione e le relative modifiche nel caso di doppioni il risultato finale è che le canzoni risultavano come era lecito attendersi 500 mentre gli artisti, escludendo quelli con più canzoni, risultavano 281 come riportato in **Tabella 2.**

Tabella 2: numero degli artisti e delle canzoni

Artisti	Canzoni
281	500

Like a Rolling Stone	Stairway to Heaven	Whole Lotta Shakin' Going On	Suspicious Minds
(I Can't Get No) Satisfaction	Sympathy for the Devil	Bo Diddley	Blitzkrieg Bop
Imagine	River Deep - Mountain High	For What It's Worth	I Still Haven't Found What I'm Looking For
What's Going On	You've Lost That Lovin' Feelin'	She Loves You	Good Golly, Miss Molly
Respect	Light My Fire	Sunshine of Your Life	Blue Suede Shoes
Good Vibrations	One	Redemption Song	Great Balls of Fire
Johnny B. Goode	No Woman, No Cry	Jailhouse Rock	Roll Over Beethoven
Hey Jude	Gimme Shelter	Tangled Up in Blue	Love and Happiness
Smells Like Teen Spirit	That'll Be the Day	Crying	Fortunate Son
What'd I Say	Dancing in the Street	Walk On By	Crazy
My Generation	The Weight	Papa's Got a Brand New Bag	You Can't Always Get What You Want
A Change Is Gonna Come	Waterloo Sunset	California Girls	Voodoo Child (Slight Return)
Yesterday	Tutti-Frutti	Superstition	Be-Bop_a_Lula
Blowin' in the Wind	Georgia on My Mind	Summertime Blues	Hot Stuff
London Calling	Heartbreak Hotel	Whole Lotta Love	Living for the City
I Want to Hold Your Hand	Heroes	Strawberry Fields Forever	The Boxer
Purple Haze	All Along the Watchtower	Mystery Train	Mr. Tambourine Man
Maybellene	Bridge Over Troubled Water	I Got You (I Feel Good)	Not Fade Away
Hound Dog	Hotel California	Mr. Tambourine Man	Little Red Corvette
Let It Be	The Tracks on My Tears	You Really Got Me	Brown Eyed Girl
Born to Run	The Message	I Heard It Through the Grapevine	I've Been Loving You Too Long (to Stop Now)
Be My Baby	When Doves Cry	Blueberry Hill	I'm So Lonesome I Could Cry
In My Life	When a Man Loves a Woman	Norwegian Wood (This Bird Has Flown)	That's All Right
People Get Ready	Louie Louie	Every Breath You Take	Up on the Roof
God Only Knows	Long Tall Sally	Crazy	You Send Me
(Sittin' on) the Dock of the Bay	Anarchy in the U.K	Thunder Road	Honky Tonk Women
Layla	Whiter Shade of Pale	Ring of Fire	Take Me to the River
A Day in the Life	Billie Jean	My Girl	Crazy in Love
Help!	The Times They Are AChangin'	California Dreamin'	Go Your Own Way
I Walk the Line	Let's Stay Together	In the Still of the Night	I Want You Back
Young Americans	Heroin	Welcome to the Jungle	Fuck tha Police
I'm Eighteen	Penny Lane	Into the Mystic	Suite: Judy Blue Eyes
Just Like Heaven	The Twist	Where Did Our Love Go	Nuthin' But a 'G' Thang
Under the Boardwalk	Cupid	Do Right Woman - Do Right Man	It's Your Thing
Clocks	Paradise City	How Soon Is Now	Piano Man
I Love Rock 'N Roll	My Sweet Lord	Last Night	Blue Suede Shoes
I Will Survive	Sheena Is a Punk Rocker	I Want to KnowWhat Love Is	William, It Was Really Nothing
Time to Pretend	All Apologies	Sabotage	American Idiot
Ignition (Remix)	Soul Man	Super Freak	Tumbling Dice
Brown Sugar	Kiss	Since U Been Gone	Smoke on the Water
Running on Empty	Rollin' Stone	White Rabbit	New Year's Day
The Rising	Get Ur Freak On	Cry Me a River	Everybody Needs Somebody to Love
Miss You	Respect Yourself	Lady Marmalade	(White Man) In Hammersmith Palais
Buddy Holly	Rain	Mustang Sally	Ain't It a Shame
Shop Around	Standing in the Shadows of Love	Alone Again Or	Midnight Train to Georgia
Runaway	Surrender	Beast of Burden	Ramble On

Figura 17: elenco canzoni (parte uno di tre)

Stand By Me	Earth Angel	Hey Ya!	In My Room
The House of the Rising Sun	Foxey Lady	Green Onions	96 Tears
It's a Man's Man's Man's World	A Hard Day's Night	Save The Last Dance For Me	Caroline, No
Jumpin' Jack Flash	Rave On	The Thrill Is Gone	1999
Will You Love Me Tomorrow	Proud Mary	Please Please Me	Rockin' in the Free World
Shake, Rattle & Roll	The Sounds of Silence	Desolation Row	Your Cheatin' Heart
Changes	I Only Have Eyes for You	Who'll Stop The Rain Creedence Clearwater Revival	Do You Believe in Magic
Rock & Roll Music	(We're Gonna) Rock Around the Clock	I Never Loved a Man (the Way I Love You)	Jolene
Born to Be Wild	Moment of Surrender	Back in Black	Boom Boom
Maggie May	I'm Waiting for the Man	Stayin' Alive	Spoonful
With or Without You	Bring the Noise	Knockin' On Heaven's Door	Walk Away Renee
Who Do You Love	Folsom Prison Blues	Free Bird	Walk On the Wild Side
Won't Get Fooled Again	I Can't Stop Loving You	Rehab	Oh, Pretty Woman
In the Midnight Hour	Nothing Compares 2 U	Wichita Lineman	Dance to the Music
While My Guitar Gently Weeps	Bohemian Rhapsody	There Goes My Baby	Hoochie Coochie Man
Your Song	Fast Car	Peggy Sue	Fire and Rain
Eleanor Rigby	Let's Get It On	Sweet Child O' Mine	Should I Stay or Should I Go
Family Affair	Papa Was A Rollin' Stone	Maybe	Good Times
I Saw Her Standing There	Losing My Religion	Don't Be Cruel	Mannish Boy
Kashmir	Both Sides Now	Hey Joe	Moondance
All I Have to Do Is Dream	99 Problems	Flash Light	Just Like A Woman
Please, Please, Please	Dream On	Loser	Sexual Healing
Purple Rain	Dancing Queen	Bizarre Love Triangle	Only The Lonely
I Wanna be Sedated	God Save The Queen	Come Together	We Gotta Get Out of This Place
Everyday People	Paint It, Black	Positively 4th Street	Paper Planes
Rock Lobster	I Fought The Law	Try a Little Tenderness	I'll Feel A Whole Lot Better
Me and Bobby McGee	Don't Worry Baby	Lean on Me	Everyday
Lust for Life	Free Fallin'	Reach Out, I'll Be There	I Got A Woman
Cathy's Clown	September Girls	Bye Bye Love	Planet Rock
Eight Miles High	Love Will Tear Us Apart	Gloria	I Fall to Pieces
Love Me Tender	Just My Imagination	Sweet Emotion	Complete Control
I Wanna Be Your Dog	Baby I Need Your Loving	Monkey Gone to Heaven	The Letter
Push It	Summer in the City	I Feel Love	Highway 61 Revisited
Pink Houses	Can't Help Falling in Love	Ode to Billie Joe	Unchained Melody
In Da Club	Remember (Walkin' in the Sand)	That Girl Can't Help It	How Deep Is Your Love
Come Go With Me	(Don't Fear) the Reaper	Young Blood	White Room
I Shot the Sheriff	Thirteen	I Can't Help Myself	Personal Jesus
I Got You Babe	Sweet Home Alabama	The Boys of Summer	I'm a Man
Come As You Are	Enter Sandman	Juicy	The Wind Cries Mary
Pressure Drop	Tonight's the Night	Bad Moon Rising	I Can't Explain
Leader of the Pack	Thank You (Falettinme Be Mice Elf Agin)	Sweet Dreams (Are Made of This)	Marquee Moon
Ticket to Ride	C'mon Everybody	Little Wing	Wonderful World
Ohio	Umbrella	Nowhere to Run	Brown Eyed Handsome Man
I Know You Got Soul	Visions of Johanna	Got My Mojo Working	Another Brick in the Wall Part 2
Tiny Dancer	We've Only Just Begun	Killing Me Softly With His Song	Fake Plastic Trees
Roxanne	In Bloom	All You Need Is Love	Maps

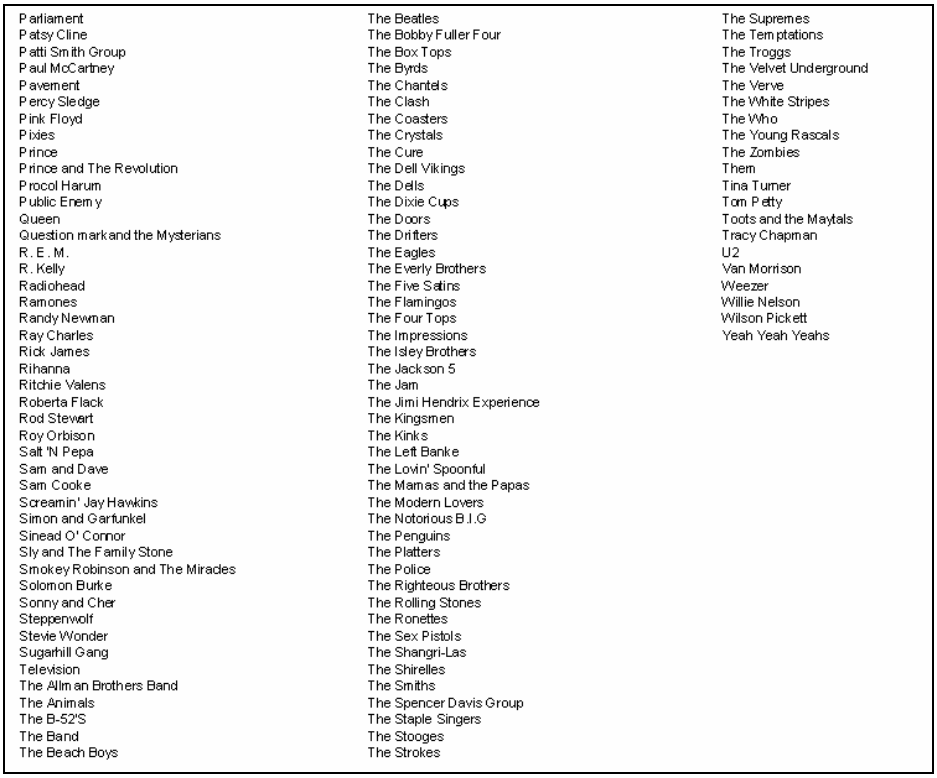
Figura 18: elenco canzoni (parte due di tre)

Son of a Preacher Man	Sunday Bloody Sunday	Get Up, Stand Up
The Wanderer	Jesus Walks	Heart of Gold
Stand!	Roadrunner	Sign 'O' the Times
Rocket Man	He Stopped Loving Her Today	One Way or Another
Love Shack	Sloop John B	Like a Prayer
Gimme Some Lovin	Sweet Little Sixteen	One More Time
(Your Love Keeps Lifting Me) Higher and Higher	Something	Do Ya Think I'm Sexy
The Night They Drove Old Dixie Down	Somebody to Love	Blue Eyes Crying in the Rain
Hot Fun in the Summertime	Born in the U.S.A.	Ruby Tuesday
Rapper's Delight	I'll Take You There	With a Little Help From My Friends
Chain of Fools	Ziggy Stardust	That's Entertainment
Paranoid	Pictures of You	Why Do Fools Fall in Love
Money Honey	Chapel of Love	Lonely Teardrops
Mack the Knife	Ain't No Sunshine	What's Love Got to Do With It
All the Young Dudes	Seven Nation Army	Iron Man
Paranoid Android	You Are the Sunshine of My Life	Wake Up Little Susie
Highway to Hell	Help Me	In Dreams
Heart of Glass	Call Me	I Put a Spell on You
Mississippi	(What's So Funny 'Bout) Peace, Love and Understanding	Comfortably Numb
Wild Thing	Smokestack Lightning	Don't Let Me Be Misunderstood
I Can See For Miles	Summer Babe (Winter Version)	Alison
Oh, What A Night	Walk This Way Run	Wish You Were Here
Hallelujah	Money (That's What I Want)	Many Rivers to Cross
Higher Ground	Can't Buy Me Love	School's Out
Ooo Baby Baby	Stan	Take Me Out
He's a Rebels	She's Not There	Heartbreaker
Sail Away	Train in Vain	Cortez the Killer
Walking in the Rain	Tired of Being Alone	Fight the Power
Tighten Up	Black Dog	Dancing Barefoot
Personality Crisis	Street Fighting Man	Baby Love
Hit the Road Jack	Sweet Jane	Spanish Harley
Pride (In the Name of Love)	Wild Horses	The Great Pretender
Radio Free Europe	Beat It	All Shook Up
Goodbye Yellow Brick Road	Beautiful Day	Tears in Heaven
Tell It Like It Is	Walk This Way	Watching the Detectives
Bitter Sweet Symphony	Maybe I'm Amazed	
Whipping Post	You Keep Me Hangin' On	
Good Lovin'	Baba O' Riley	
Get Up (I Feel Like Being a) Sex Machine	The Harder They Come	
For Your Precious Love	Runaround Sue	
The End	Jim Dandy	
That's the Way of the World	Piece of My Heart	
We Will Rock You	La Bamba	
I Can't Make You Love Me	California Love	
Subterranean Homesick Blues	Candle in the Wind	
Spirit in the Sky	That Lady (Part 1 and 2)	

Figura 19: elenco canzoni (parte tre di tre)

50 Cent	Cream	Jackson Browne
Aaron Neville	Creedence Clearwater Revival	James Brown and His Famous Flames
ABBA	Crosby, Stills and Nash	James Taylor
AC/DC	Daft Punk	Janis Joplin
Aerosmith	David Bowie	Jay-Z
Al Green	Deep Purple	Jeff Buckley
Alice Cooper	Del Shannon	Jefferson Airplane
Amy Winehouse	Depeche Mode	Jerry Butler and the Impressions
Archie Bell and the Drells	Derek and The Dominos	Jerry Lee Lewis
Aretha Franklin	Dion	Jimmy Cliff
B.B. King	Dionne Warwick	Joan Jett and the Blackhearts
Barrett Strong	Dolly Parton	John Cougar Mellencamp
Beastie Boys	Don Henley	John Lee Hooker
Beck	Donna Summer	John Lennon
Bee Gees	Dr. Dre	Johnny Cash
Ben E. King	Dusty Springfield	Joni Mitchell
Beyonce	Earth, Wind and Fire	Joy Division
Big Brother and the Holding Company	Eddie Cochran	Justin Timberlake
Big Joe Turner	Elton John	Kanye West
Big Star	Elvis Costello	Kelly Clarkson
Bill Haley and His Comets	Elvis Presley	Labelle
Bill Withers	Eminem Feat. Dido	Lavern Baker
Billy Joel	Eric B. and Rakim	Led Zeppelin
Black Sabbath	Eric Clapton	Little Richard
Blondie	Eurythmics	Lou Reed
Blue Oyster Cult	Fats Domino	Love
Bo Diddley	Fleetwood Mac	Lynyrd Skynyrd
Bo Diddley	Foreigner	M.I.A.
Bob Dylan	Frankie Lymon and the Teenages	Madonna
Bob Marley and The Wailers	Franz Ferdinand	Martha and the Vandellas
Bobbie Gentry	Gene Vincent and His Blue Caps	Marvin Gaye
Bobby Darin	George Harrison	Metallica
Bonnie Raitt	George Jones	MGMT
Booker T. and the MG's	Gladys Knight and The Pips	Michael Jackson
Bruce Springsteen	Glen Campbell	Missy Elliot
Buddy Holly	Gloria Gaynor	Mott the Hoople
Buddy Holly and The Crickets	Gnarls Barkley	Muddy Waters
Buffalo Springfield	Grandmaster Flash and The Furious Five	N.W.A
Carl Perkins	Green Day	Neil Young
Carpenters	Guns N' Roses	New Order
Cheap Trick	Hank Williams	New York Dolls
Chic	Howlin' Wolf	Nirvana
Chubby Checker	Iggy Pop	Norman Greenbaum
Chuck Berry	Ike and Tina Turner	Otis Redding
Coldplay	Jackie Wilson	OutKast

Figura 20: elenco artisti (parte uno di due)



Parliament	The Beatles	The Supremes
Patsy Cline	The Bobby Fuller Four	The Temptations
Patti Smith Group	The Box Tops	The Troggs
Paul McCartney	The Byrds	The Velvet Underground
Pavement	The Chantels	The Verve
Percy Sledge	The Clash	The White Stripes
Pink Floyd	The Coasters	The Who
Pixies	The Crystals	The Young Rascals
Prince	The Cure	The Zombies
Prince and The Revolution	The Dell Vikings	Them
Procol Harum	The Dells	Tina Turner
Public Enemy	The Dixie Cups	Tom Petty
Queen	The Doors	Toots and the Maytals
Question mark and the Mysterians	The Drifters	Tracy Chapman
R.E.M.	The Eagles	U2
R. Kelly	The Everly Brothers	Van Morrison
Radiohead	The Five Satins	Weezer
Ramones	The Flamingos	Willie Nelson
Randy Newman	The Four Tops	Wilson Pickett
Ray Charles	The Impressions	Yeah Yeah Yeahs
Rick James	The Isley Brothers	
Rihanna	The Jackson 5	
Ritchie Valens	The Jam	
Roberta Flack	The Jimi Hendrix Experience	
Rod Stewart	The Kingsmen	
Roy Orbison	The Kinks	
Salt 'N Pepa	The Left Banke	
Sam and Dave	The Lovin' Spoonful	
Sam Cooke	The Mamas and the Papas	
Screamin' Jay Hawkins	The Modern Lovers	
Simon and Garfunkel	The Notorious B.I.G	
Sinead O' Connor	The Penguins	
Sly and The Family Stone	The Platters	
Smokey Robinson and The Miracles	The Police	
Solomon Burke	The Righteous Brothers	
Sonny and Cher	The Rolling Stones	
Steppenwolf	The Ronettes	
Stevie Wonder	The Sex Pistols	
Sugarhill Gang	The Shangri-Las	
Television	The Shirelles	
The Allman Brothers Band	The Smiths	
The Animals	The Spencer Davis Group	
The B-52'S	The Staple Singers	
The Band	The Stooges	
The Beach Boys	The Strokes	

Figura 21: elenco artisti (parte due di due)

4.2 Segmentazione delle parole delle pagine web in singoli termini

In questa fase comincia l'elaborare delle informazioni presenti all'interno delle pagine web, grazie all'utilizzo di un programma che permette di separare i testi in termini.

Per prima cosa, utilizzando la funzione "*lowercase*", tutte le lettere delle parole sono state trasformate in minuscolo, in modo tale da non dover contare con due contatori diversi la stessa parola che aveva, come unica differenza, la prima lettera maiuscola.

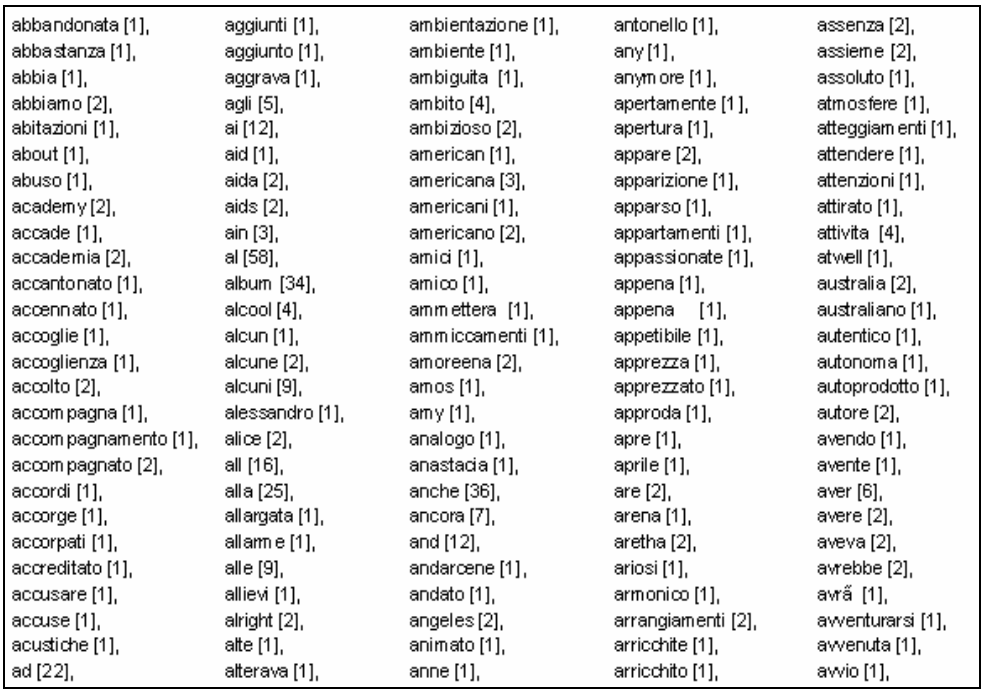
Successivamente, utilizzando la classe **java.io.streamtokenizer** venivano divisi in singoli termini il testo di partenza, eliminando contemporaneamente la punteggiatura, rendendo i singoli termini uniformi tra di loro, tutti minuscoli e senza punteggiatura.

Per questo progetto non si sono potuti mettere dei limiti minimi di lettere per parola in quanto in questi documenti erano presenti molti termini significativi che avevano due o tre caratteri. Imponendo dei limiti sarebbero stati eliminati sicuramente gli articoli e le preposizioni ma anche alcune parole come "U2", "One" e i tag del tipo "60s", molto utili per le elaborazioni successive.

4.3 Frequenza dei termini nelle singole pagine

Una volta separati i testi presenti nelle pagine web in singoli termini, si è proceduto, utilizzando un apposito programma, al conteggio delle occorrenze di ogni termine.

In modo automatico le pagine web venivano così elaborate e salvate in file come riportato nell'esempio di **Figura 22** .



abbandonata [1],	aggiunti [1],	ambientazione [1],	antonello [1],	assenza [2],
abbastanza [1],	aggiunto [1],	ambiente [1],	any [1],	assieme [2],
abbia [1],	aggrava [1],	ambiguita [1],	anymore [1],	assoluto [1],
abbiamo [2],	agli [5],	ambito [4],	apertamente [1],	atmosfere [1],
abitazioni [1],	ai [12],	ambizioso [2],	apertura [1],	atteggiamenti [1],
about [1],	aid [1],	american [1],	appare [2],	attendere [1],
abuso [1],	aida [2],	americana [3],	apparizione [1],	attenzioni [1],
academy [2],	aids [2],	americani [1],	apparso [1],	attirato [1],
accade [1],	ain [3],	americano [2],	appartamenti [1],	attivita [4],
accademia [2],	al [58],	amici [1],	appassionate [1],	atwell [1],
accantonato [1],	album [34],	amico [1],	appena [1],	australia [2],
accennato [1],	alcool [4],	ammettera [1],	appena [1],	australiano [1],
accoglie [1],	alcun [1],	ammiccamenti [1],	appetibile [1],	autentico [1],
accoglienza [1],	alcune [2],	amoreena [2],	apprezza [1],	autonoma [1],
accolto [2],	alcuni [9],	amos [1],	apprezzato [1],	autoprodotto [1],
accompagna [1],	alessandro [1],	amy [1],	approda [1],	autore [2],
accompagnamento [1],	alice [2],	analogo [1],	apre [1],	avendo [1],
accompagnato [2],	all [16],	anastacia [1],	aprile [1],	avente [1],
accordi [1],	alla [25],	anche [36],	are [2],	aver [6],
accorge [1],	allargata [1],	ancora [7],	arena [1],	avere [2],
accorpati [1],	allarme [1],	and [12],	aretha [2],	aveva [2],
accreditato [1],	alle [9],	andarcene [1],	ariosi [1],	avrebbe [2],
accusare [1],	allievi [1],	andato [1],	armonico [1],	avrã [1],
accuse [1],	alright [2],	angeles [2],	arrangiamenti [2],	avventurarsi [1],
acustiche [1],	alte [1],	animato [1],	arricchite [1],	avvenuta [1],
ad [22],	alterava [1],	anne [1],	arricchito [1],	avvio [1],

Figura 22: elaborazione della versione italiana di Wikipedia della biografia di "Elton John"

4.4 Frequenza dei termini in documenti della collezione

Sono stati elaborati, sempre con lo stesso metodo, le parole contenute nelle pagine riferite allo stesso argomento e alla stessa provenienza, come riportato in **Tabella 3**.

Tabella 3: divisione delle elaborazioni secondo argomento e sito di provenienza

Argomento	Provenienza	versione
Artisti	Wikipedia	Italiana
Canzoni	Wikipedia	Italiana
Artisti	Wikipedia	Inglese
Canzoni	Wikipedia	Inglese
Artisti	Last.fm	Italiana
Tag artisti	Last.fm	Italiana
Canzoni	Last.fm	Italiana
Tag canzoni	Last.fm	Italiana
Artisti	Last.fm	Inglese
Tag artisti	Last.fm	Inglese
Canzoni	Last.fm	Inglese
Tag canzoni	Last.fm	Inglese

Dividendo i file in questa maniera si tratterà di elaborare dati dello stesso argomento, come ad esempio tutte le pagine relative ai cantanti provenienti dal sito Wikipedia nella versione italiana come riportato in **Figura 23**.

accedere [1],	accompagnamento [2],	accostò [1],	acquisito [3],
accelerata [1],	accompagnandoli [1],	account [1],	acquista [1],
accelerate [3],	accompagnandosi [1],	accreditano [1],	acquistabile [1],
accendere [1],	accompagnano [2],	accreditare [1],	acquistabili [1],
accenna [1],	accompagnare [1],	accreditata [2],	acquistando [1],
accennato [2],	accompagnata [6],	accreditati [2],	acquistano [1],
accentuarsi [1],	accompagnati [7],	accreditato [5],	acquistare [1],
acceso [1],	accompagnato [33],	accrescendo [1],	acquistata [6],
access [1],	accompagnava [2],	accrescere [3],	acquistati [1],
accessibili [1],	accompagnavano [2],	accrescersi [1],	acquistato [2],
accetta [2],	accompagnerà [1],	accresceva [1],	acquistava [1],
accettabili [1],	accompagnò [3],	accumulati [1],	acquisto [4],
accettando [2],	accomunava [1],	accusa [8],	acquistò [4],
accettare [2],	acconciatura [1],	accusandola [1],	acro [1],
accettarono [3],	acconsentì [4],	accusare [3],	acrobat [1],
accettata [1],	accontentare [1],	accusarono [3],	acronimo [3],
accettato [7],	accorciato [1],	accusata [3],	act [1],
accettò [7],	accordarono [1],	accusati [3],	action [2],
acciaio [2],	accordasse [1],	accusato [9],	acuff [17],
accidentali [1],	accordata [1],	accusatore [1],	acustica [12],
accidentalmente [2],	accordati [2],	accusatori [1],	acustiche [9],
accingeva [1],	accordi [7],	accusavano [2],	acustici [3],

Figura 23: elaborazione delle pagine dei cantanti presenti in Wikipedia versione italiana

Oltre a conteggiare le occorrenze totali delle parole appartenenti allo stesso argomento e della provenienza, il programma avrà in più il compito di segnalare il numero di documenti in cui tale termine appare. L'esempio di **Tabella 4** riporta il risultato ottenuto sui cantanti della versione italiana di Wikipedia.

Tabella 4: estratto per le pagine dei cantati presenti nella versione italiana di Wikipedia

Termini	Occorrenze totali	Documenti nei quali è presente
accompagnato	33	28
accusa	8	8
acustica	12	11
band	294	142
canzone	463	178

4.5 Aggregazione delle parole con la stessa radice

Per fare questa operazione ho utilizzato un programma che analizza le parole e le aggrega in base alle radici dei termini stessi come in **Figura 24**. Lo stemming **[15]** è il processo di riduzione di una parola alla sua forma radice, detta tema. Il tema non corrisponde necessariamente alla radice morfologica della parola: normalmente è sufficiente che le parole correlate siano mappate allo stesso tema anche se quest'ultimo non è una valida radice per la parola. La creazione di un algoritmo di

stemming è stato un grosso problema dell'informatica ed è spesso utilizzato nei motori di ricerca per l'espansione delle interrogazioni.

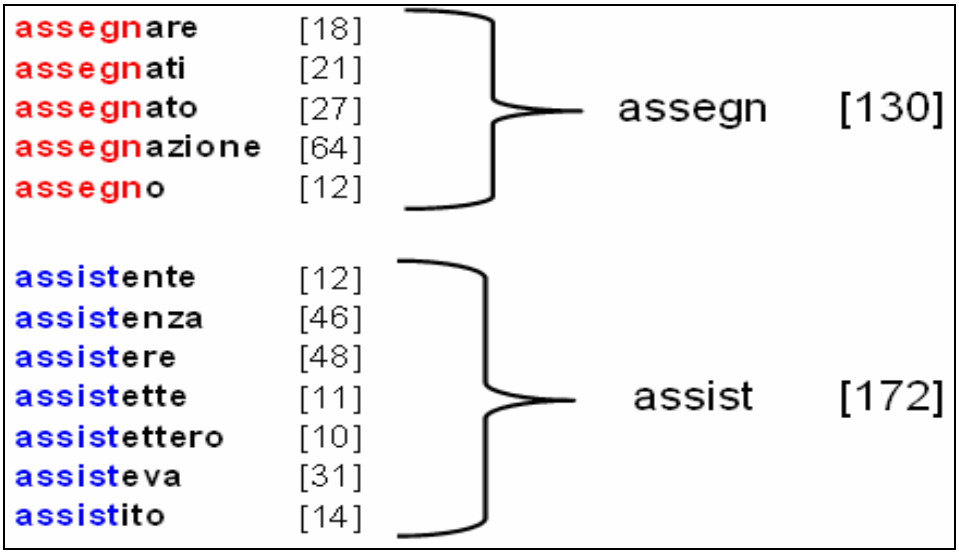


Figura 24: esempio di stemming sugli artisti di Wikipedia versione italiana

4.5.1 Utilizzo dello stemming per l'elaborazione dei dati ottenuti

Partendo dai file ottenuti nelle fasi precedenti di questo capitolo ho utilizzato uno stemmer per filtrare meglio i dati raccolti e così raggrupparli da singoli termini a gruppi di parole con la stessa radice.

Con questo programma le singole parole "rock", "rocker" e "rockeggiare" con i loro singoli contatori sarebbero state conglobate tutte nella radice "rock" con occorrenza pari alla somma delle singole occorrenze. Questo strumento risulta molto utile quando si devono elaborare enormi quantità di dati come in questo caso.

4.6 Confronto tra diverse sorgenti

In questa ultima parte relativa alla elaborazione dei dati è stato fatto un confronto tra le frequenze di utilizzo dei termini appartenenti a diverse sorgenti ma della stessa lingua. La frequenza è stata calcolata dividendo il numero di documenti nei quali il termine è presente con il numero di occorrenze totali del termine. Il numero derivante da questo calcolo sarà compreso tra 0 e 1. Se il termine è vicino ad 1 significa che il numero di documenti è vicino al numero di occorrenze e quindi che i termini sono presenti poche volte e solo in quelle determinate pagine. Se il numero si avvicina a 0 vuol dire che in pochi documenti è presenti molte volte lo stesso

termine, ciò significa che la ricerca di quel termine in quei documenti sarà inefficace come parola chiave. Quindi più la frequenza di un termine è avvicina ad 1 e più tale termine risulterà parola chiave di quel testo. Indicate in rosso nell'ultima colonna della

Tabella 5. le frequenze che si avvicinano maggiormente ad 1 tra le accoppiate termine, versione e argomento.

Tabella 5: esempio di confronto tra varie sorgenti

termine	versione	argomento	sorgente	numero di documento	occorrenze totali	frequenza
concert	italiana	artisti	Wikipedia	**147**	**324**	**0.45**
			Last.fm	**412**	**711**	**0.58**
sing	inglese	artisti	Wikipedia	**230**	**312**	**0.74**
			Last.fm	**476**	**782**	**0.61**
assol	italiana	artisti	Wikipedia	**140**	**190**	**0.74**
			Last.fm	**375**	**700**	**0.54**
live	inglese	artisti	Wikipedia	**94**	**118**	**0.80**
			Last.fm	**284**	**471**	**0.60**

Capitolo 5

Conclusioni

L’obiettivo di questo elaborato è stato quello di costruire un componente di un sistema automatico di reperimento delle informazioni musicali che raccogliesse da vari siti Web di informazione musicale le notizie in maniera automatica. Tale componente permette, dopo aver elaborato tali informazioni, di trovare relazioni o aspetti singolari in grado di rendere un determinato brano musicale o un artista diverso ed identificabile univocamente rispetto ad altri. Le fasi di questo elaborato sono state essenzialmente tre: l’analisi dei requisiti, l’implementazione e l’elaborazione dei dati.

La fase dell’analisi dei requisiti ha riguardato lo studio dei modelli dell’Information Retrieval e le tecniche di Music Information. Questo primo periodo è stato necessario per definire meglio quale era il problema da affrontare e per trovare il modo di poterlo risolvere nella maniera migliore. All’interno di questa fase ho analizzato i più conosciuti siti web per la ricerca delle informazioni musicali e del software che poteva essere utilizzato nella fase di implementazione.

La fase dell’implementazione è risultata il momento fondamentale dell’intero elaborato in quanto proprio in questa fase si è dovuto mettere in pratica quello che si era visto teoricamente nella prima parte. Durante tale periodo sono stati affrontati i problemi e le difficoltà che si incontrano nel lavorare con il web. Il più gravoso è stato l’implementazione dello Stax parser di Last.fm che diversamente da quello di

Wikipedia aveva delle difficoltà nella lettura del codice sorgente. Per questo motivo ho dovuto risolvere il problema utilizzando un programma che leggesse solamente la parte di codice di nostro interesse.

La fase dell'elaborazione dei dati è stata la conclusione del lavoro svolto nei mesi precedenti ed è servita per verificare quello che era stato solo supposto nelle fasi precedenti. In questo elaborato sono stati utilizzati i dati provenienti dai siti internet di Wikipedia e Last.fm che rispetto ad un artista o ad una canzone hanno comportamenti diversi.

- Wikipedia adotta un comportamento neutro in quanto descrive solamente i fatti o i personaggi legati ad una canzone o ad un artista senza esprimerne giudizi.
- Last.fm esprime giudizi la maggior parte delle volte positivo riguardo al brano o all'artista, in quanto viene aggiornato periodicamente dai fan o direttamente dalle etichette discografiche vicine agli artisti.

Un buon sviluppo futuro potrebbe essere quello di utilizzare come sito di riferimento Amazon.com. In questo sito vengono venduti e scambiati CD musicali e libri che riguardano il mondo della musica e per ogni oggetto di questo tipo l'utente che ha acquistato o venduto può esprimere i suoi giudizi su un album musicale. In questo modo le informazioni che si possono reperire non saranno più esclusivamente neutre o tendenti al positivo ma potrebbero essere negative o pessime. In questo sito, inoltre è possibile esprimere il proprio favore o il proprio dissenso esprimendo il proprio voto, sotto forma di stelline, riguardo quello che si cerca. Analizzando Amazon.com si potrebbero reperire molte altre informazioni riguardo a brani e a cantanti completando e arricchendo questo progetto.

Bibliografia

[1] R. BAEZA-YATES, B. RIBEIRO-NETO
"Modern Information Retrieval"
Addison-Wesley, 1999

[2] URL: http://www.cbi.umn.edu/collections/inv/cbi00081.html

[3] URL: http://www.music-ir.org/

[4] R MIOTTO, N. ORIO
"A Probabilistic Approach to Merge Context and Content Information for Music Retrieval" Proceedings of the The International Society for Music Information Retrieval (ISMIR), pp 15-20
Utrecht (The Netherlands), 2010

[5] URL: http://it.wikipedia.org/wiki/Wikimedia_Foundation

[6] URL: http://it.wikipedia.org/wiki/

[7] URL: http://it.wikipedia.org/wiki/Wikimedia_Commons

[8] URL: http://www.lastfm.it/

[9] URL: http://www.amazon.com/

[10] URL: http://www.youtube.com/

[11] URL: http://java.sun.com/developer/technicalArticles/ThirdParty/WebCrawler/

[12] URL: http://java-source.net/open-source/parser-generators

[13] URL: http://download.oracle.com/docs/webservices/Webservices/docs/StaxParser.html

[14] URL: http://www.rollingstone.com/

[15] URL: http://www.comp.lancs.ac.uk/computing/research/stemming/general/