Progetto integrativo sviluppato nell'ambito del corso di

Sistemi Intelligenti per Internet

(Anno Accademico 2014-2015)

Estrazione di riferimenti temporali ad eventi dal Web 2.0

Documentazione

docente di riferimento:

Fabio Gasparetti

a cura di

L.Martucci, C.Raponi, D.Santilli, L.Tomaselli

Indice

Capitolo 1

Introduzione

Il progetto[1] è relativo allo studio ed all'implementazione di un sistema di estrazione di riferimenti temporali ad eventi a partire da informazioni pubblicate su siti Web 2.0.
I riferimenti temporali trattati sono associati allo svolgimento di eventi inerenti al contesto musicale, quindi relativi a concerti e manifestazioni affini.
Lo sviluppo ha seguito un andamento sperimentale, in quanto nel corso dell'implementazione sono state effettuate diverse scelte realizzative, che ci hanno condotto all'utilizzo di alcuni strumenti e tecniche che sono state in parte mantenute e in parte scartate nella versione finale del progetto.

Le informazioni pubblicate sui siti analizzati sono state inizialmente elaborate nella loro interezza, infatti sia i titoli dei siti che i contenuti testuali degli stessi sono stati sottoposti ad un processamento che prevedeva l'applicazione di diversi tool in grado di far emergere i dati rilevanti per la nostra ricerca (protagonista, luogo e data dell'evento). Nello specifico abbiamo impiegato i seguenti *tool*:

- **Stanford Temporal Tagger, SUTime**[2]: libreria per il riconoscimento di espressioni temporali normalizzate, come ad esempio le date;
- **Stanford Named Entity Recognizer (NER)**[3]: libreria per l'etichettatura, all'interno di un testo, di sequenze di parole che identificano persone, organizzazioni e luoghi;
- **TagMe**[4]: libreria per l'identificazione di brevi frasi significative, all'interno di un testo non strutturato, in grado di fornire un collegamento ad una pagina di Wikipedia pertinente;

SUTime è stato utilizzato per far emergere dai testi tutte le date significative presenti, mentre NER e TagMe sono stati impiegati in modo combinato per evidenziare la presenza di termini associati ad informazioni che potessero identificare il protagonista/i e il luogo associato all'evento. I dati ricavati da questo tipo di approccio sono stati successivamente trattati per la formazione di un training set che fosse in grado di coprire il più possibile la molteplicità di tipologie di casi con cui tali informazioni sono presenti all'interno dei siti Web 2.0. Sono quindi state valutate diverse *feature* come ad esempio la presenza di questi

dati nel titolo della pagina, il numero di occorrenze e il numeroso di categorie di Wikipedia maggiormente rilevanti associate da TagMe ai diversi termini selezionati. Per applicare le tecniche di *Machine Learning* abbiamo fatto uso dei metodi di apprendimento supervisionato offerti da **SVM** (*Support Vector Machines*), il cui lavoro è stato sviluppato prelevamente attraverso l'uso della libreria **LIBLINEAR**[5], in grado di garantire la classificazione e l'analisi di testi brevi. In forma minore abbiamo sperimentato questo lavoro anche attraverso **Weka**[6].

Successivamente un'attenta analisi dei dati visualizzati, ci ha portato alla scelta di escludere i contenuti delle pagine dei siti selezionati, sia perchè la troppa entropia dei dati presenti conduceva verso valutazioni sbagliate, sia perchè nel corso dell'analisi è emersa una decisiva rilevanza dei contenuti dei titoli dei siti trattati. Analizzando una consistente mole di dati abbiamo quindi notato che le informazioni importanti per i nostri fini realizzativi sono, nella maggior parte dei casi, presenti nel titolo nel quale sono espressi in modo estremamente chiaro e sintetizzato. Tale scelta implementativa ci ha condotto verso l'utilizzo di ulteriori tecniche di elaborazione testuale maggiormente focalizzate sul *Part-of-Speech Tagging* (POST). In questo modo abbiamo creato un nostro sistema di tagging incentrato principalmente sui termini per noi rilevanti, in grado di fornire uno strumento capace di evidenziare, dato un sito Web 2.0, la presenza dei termini necessari per la descrizione di un evento musicale.

Capitolo 2

Struttura del progetto

2.1 Organizzazione

Il progetto è stato stato sviluppato in maniera sperimentale, utilizzando diversi approcci realizzativi supportati da differenti strumenti ed implementazioni. Questo raffinamento è stato supportato da diverse implementazioni che si sono susseguite nel corso dello sviluppo. Per lasciare traccia di questo percorso, all'interno del progetto sono stati volutamente lasciati tutti gli strumenti utilizzati a partire dalla prima all'ultima versione del progetto.
In particolare, la versione finale, che si focalizza sulle tecniche di *Part-of-Speech Tagging* (POST), fa riferimento alle sole classi dei package *Final_Version*.
L'organizzazione è la seguente:

- Final_Version.boilerpipe: classi per la rilevazione e la rimozione del surplus (rumore) all'interno del contenuto testuale principale di una pagina web;
- Final_Version.event: classi per l'interazione con Bing per l'estensione del training e del test set;
- Final_Version.lastFM: classi a supporto dell'interazione con **Last.fm**;
- Final_Version.main: classi principali per l'avvio di training e tagging insieme;
- Final_Version.parser: classi per il parsing dei dati;
- Final_Version.suTime: classi per l'interazione con **SUTime**;
- Final_Version.testing: classi per la parte di test;
- Final_Version.training: classi per la parte di training;

Altri file di supporto per la versione finale, come ad esempio quelli relativi ai training set piuttosto che ai risultati del testing, sono presenti in varie folder identificative.

Tutti gli altri package prensenti nel progetto, sono relativi a sperimentazioni passate, che sono state in parte mantenute e in parte abbandonate per la versione ultima del progetto.

2.2 Implementazione

Il progetto si compone principalmente di tre fasi:

- **Data Collection**: recupero delle informazioni attraverso l'analisi di siti Web 2.0;
- **Training**: definizione di regole di POST e creazione di un training set consistente;
- **Tagging**: classificazione di un testing set e valutazione dell'efficacia del sistema;

Di seguito sono approfondite le dinamiche implementative proprie di ciascuna fase.

2.2.1 Data Collection

L'oggetto principale dell'analisi sono stati i *titoli* dei siti di interesse, in quanto sufficientemente completi dal punto di vista informativo, per gli scopi realizzativi del progetto. La ricerca di quest'ultimi, necessaria per definire un training set quanto più possibile omogeneo, è partita da un insieme di informazioni certe. Infatti per raccogliere una quantità notevole di siti di interesse, è stato preso come riferimento uno dei principali siti musicali, **Last.fm**[7], e a partire dagli eventi archiviati nel sito, sono state raccolte una serie di informazioni relative ad eventi futuri. Interagendo con le API messe a disposizione, è stato possibile raccogliere per ogni istanza (evento musicale) i dati relativi a:

- evento (da intendere come protagonista/i dell'evento);
- data;
- luogo, comprensivo di città e sede;

Avendo a disposizione quindi delle informazioni di base certe, queste parole sono state utilizzate come *keywords* per estendere la raccolta dati con un motore di ricerca. Questo lavoro è stato sostenuto attraverso l'interfacciamento con **Bing** al quale sono state sottoposte delle query in grado di fornire una lista di URLs riferiti a siti di interesse. Nello specifico solo due terzi delle informazioni raccolte da Last.fm (evento e data), sono state utilizzate come chiavi di ricerca per ampliare la raccolta dei siti con il supporto di Bing. L'esito di queste operazioni ha fornito una lista di siti, dai quali è stato possibile ricavare i titoli, che sono stati successivamente processati. I risultati così come sono stati restituiti da Bing sono stati ulteriormente filtrati e analizzati dal tool **boilerpipe**[8], le cui librerie forniscono algoritmi per la rilevazione e la rimozione del surplus (rumore) all'interno del contenuto testuale principale di una pagina web.

La fase di collezione dei dati ha permesso quindi di raccogliere i titoli dei siti restituiti da Bing, in seguito ad una query avente come keywords informazioni certe associate ad eventi musicali futuri, recuperate attraverso l'interazione con Last.fm.

2.2.2 Training

Le pagine web raccolte nella fase precedente sono state processate al fine di poter costituire un training set consistente. Tale processamento ha previsto una serie di interventi a livello di elaborazione testuale, il primo dei quali è stato la rimozione all'interno del titolo dell'eventuale presenza del nome del sito. Frequentemente all'interno del titolo, è citato il nome del dominio di riferimento, informazione che essendo fortemente dipendente dal contesto è stata esclusa dall'analisi. Ulteriori interventi di questo genere, hanno previsto la riduzione degli spazi bianchi all'interno dei titoli e la separazione della punteggiatura rispetto ai termini presenti, in modo da facilitare le attività successive.
L'operazione principale della fase che precede l'effettivo training automatico è stata la definizione di regole di Part-of-Speech Tagging (POST), in grado di effettuare un processo di labelling automatico rispetto ai termini presenti nel titolo. Sono stati identificati quindi dei *tag* ricorrenti da associare alle parole che costituiscono il titolo, sulla base del significato dei termini e della loro occorrenza rispetto a determinate posizioni. I tag sono stati scelti partendo dall'osservazione di numerosi casi, prestando particolare attenzione alla presenza e alla collocazione dei termini all'interno del titolo.

Il training set di titoli taggati è stato successivamente utilizzato come input di un sistema di addestramento che ha permesso la generazione di un *corpus.model*, ovvero di un modello rappresentante delle conoscenze apprese. Intervenendo sulle regole di POST identificate, sono stati prodotti diversi modelli di apprendimento per valutare quale tipo di conoscenza fosse in grado di fornire le migliori prestazioni. La diversità dei modelli consiste essenzialmente nell'aggiunta, esclusione o variazione di alcuni tag.
L'ultimo modello prodotto, utilizza i seguenti tag per etichettare i titoli che costituiscono il training set:

Tag	*Associazione*
PPP	evento
DDD	data dell'evento
CCC	città dell'evento
SSS	sede dell'evento
ET	termini di congiunzione (es. *and*)
ART	articoli (es. *the*)
AAA	termini che precedono il luogo dell'evento (es. *at*)
PRED	termini che precedono una data (es. *on*)
POSTP	termini che seguono l'evento (es. *tickets & tour dates*)
SOCIAL	termini legati ai social network (es. *on twitter*)
SELL	termini legati alla vendita dei biglietti per l'evento (es. *tickets*)
CONCERT	termini legati all'evento come manifestazione (es. *tour*)
MUSIC	termini legati in generale al mondo della musica (es. *lyrics*)
ALTRO	termini restanti

Una variante rispetto ai tag precedentemente illustrati è utilizzata nel primo modello prodotto (*base*) in cui è stato inserito il tag **SEPA** per identificare la presenza dei termini di punteggiatura più ricorrenti. Tale approccio è stato suc-

cessivamente abbandonato in seguito a diverse scelte implementative intraprese (vedi paragrafo 3.1).
L'associazione di ciascun tag ai termini nel titolo è stata effettuata in modi differenti. Per quanto riguarda l'assegnazione del tag DDD, riferito alla presenza della data, il lavoro è stato supportato dal tool SUTime, mentre per quanto riguarda l'assegnazione dei tag PPP, CCC e SSS si sfruttano le conoscenze dei valori noti per quei determinati campi, che si ricavano inizialmente con l'interazione con Last.fm. Per le restanti assegnazioni di tag si procede semplicemente effettuando delle operazioni di matching tra le parole del titolo e i termini identificati rilevanti per ciascuna categoria riportata in tabella. Tutte le parole restanti, escluse da queste considerazioni, vengono taggate con ALTRO.
L'ultima operazione necessaria per la definizione del training set prevede un *parsing* delle istanze finora considerate nel formato .**brown**[9]. Tale lavoro è affidato ad un parser capace di convertire i titoli, taggati secondo le regole di POST definite precedentemente, in un file in formato .brown in grado di essere interpretato correttamente in fase di addestramento. Di seguito è mostrato un esempio di istanza trattata:

Titolo originale: *Placebo - Liverpool - Tue, March 10, 2015*

Titolo taggato e convertito in formato brown: Placebo/PPP -/SEPA Liverpool/CCC -/SEPA Tue/DDD ,/SEPA March/DDD 10/DDD ,/SEPA 2015/DDD

La costituzione del training set prevede un'ulteriore fase di filtraggio delle istanze considerate. In modo automatico, sono state selezionate per formare il training set solo quei titoli taggati in cui fossero presenti i tag associati alle informazioni principali. Sono quindi stati esclusi tutti quei titoli che non menzionavano tra i loro tag quelli riferiti a PPP, DDD, CCC e SSS, associati quindi all'evento, alla sua data e al suo luogo.

Il sistema utilizzato per il supporto nelle fasi di training e di tagging è **Jitar**[10], un semplice Part-of-speech tagger, basato sul trigramma *Hidden Markov Model* (HMM). Un HMM è un modello statistico in cui il sistema da modellare viene assunto essere un processo di Markov con parametri sconosciuti.

2.2.3 Tagging

Il *corpus.model*, creato come risultato della fase di training in risposta ai dati impostati per l'addestramento, è utilizzato a sua volta come input per la fase di tagging. Le conoscenze apprese sono quindi sfruttate dal sistema per taggare nuovi titoli in modo da evidenziare o meno la presenza all'interno di questi, di informazioni consistenti per la descrizione di un evento musicale.
Il procedimento di raccolta dati è analogo a quello effettuato per la costituzione del training set: partendo da informazioni certe recuperate grazie all'interazione con Last.fm, si è estesa la ricerca attravero Bing per ottenere una quantità notevole di siti da sottoporre al sistema. Tali titoli sono stati trattati analogamente alla fase di training anche per quanto riguarda le operazioni di elaborazione testuale, come la riduzione degli spazi e l'esclusione dell'eventuale presenza dei domini. Il successivo tagging è stato affidato al sistema sulla base delle conoscenze descritte nel modello prodotto ed utilizzato come elemento di input.

Il risultato di queste operazioni può essere riassunto nell'esempio mostrato di seguito:

Titolo test: *Hevia at Teatro Brancaccio (Roma) on 17 Mar 2015 – Last.fm*

Titolo taggato: PPP AAA SSS SSS (CCC) PRED DDD DDD DDD

Attraverso un'analisi dei tag restituiti dal sistema è possibile osservare la presenza o meno dei termini necessari per la descrizione di un evento musicale. La presenza all'interno dei risultati dei tag PPP, DDD, CCC e SSS associati rispettivamente all'evento, alla data, alla città e alla sede in cui questo si verificherà, è indice della validità del titolo e quindi del sito corrispondente. La validità consiste nella capacità della pagina web di riferirsi ad un evento musicale che avrà luogo in futuro.
Le deduzioni relative all'esempio precedente sono:

PERSONA: *Hevia*
CITTA: *Roma*
SEDE: *Teatro Brancaccio*
DATA: *17 Mar 2015*

La completezza o meno di queste informazioni, anche rispetto ai vari tipi di modelli realizzati è stata valutata nella sezione successiva, nella quale si stima l'accuratezza generale del sistema implementato.

Capitolo 3

Risultati

3.1 Modelli analizzati

Il sistema realizzato è stato testato su cinque diverse tipologie di *corpus.model*, generati a partire da cinque diversi training set. Ognuno di questi rappresenta un diverso caso analizzato, per testare i miglioramenti o i peggioramenti dipendenti dalle differenti conoscenze apprese, rispetto al training precedenti. Ogni modello rispecchia un caso specifico:

- *base*: training costituito con le regole di POST definite nella tabella precedente (con l'aggiunta del tag **SEPA**);
- *senza punteggiatura*: training in cui le istanze sono prive di punteggiatura;
- *gestione punteggiatura*: training in cui la punteggiatura presente nelle istanze è taggata con se stessa (es. il simbolo | ha come tag |);
- *gestione parentesi*: training in cui eventuali parentesi presenti nelle istanze sono gestite in modo specifico;
- *gestione orario evento*: training in cui l'eventuale presenza dell'orario all'interno delle istanze è gestita in modo specifico;

La definizione di questi modelli è stata sviluppata in modo incrementale, specializzandosi sempre di più nella ricerca e nella gestione di elementi mirati.
Questi cinque modelli sono stati tutti valutati secondo un duplice punto di vista. Ciascun caso analizzato è stato prima testato attraverso un training di istanze selezionate come più rappresentative e successivamente su un numero di istanze maggiori. Ognuno dei modelli è stato quindi valutato due volte, in base alla diversa quantità di istanze presenti nel training set. Nello specifico, i training set che sono stati valutati per ogni caso sono:

caso	training selezionato	training con più istanze
base	*trainingSepaSel.brown*	*trainingSepa.brown*
senza punteggiatura	*trainingSenzaPuntSel.brown*	*trainingSenzaPunt.brown*
gestione punteggiatura	*trainingTaggaPuntSel.brown*	*trainingTaggaPunt.brown*
gestione parentesi	*trainingParentesiSel.brown*	*trainingParentesi.brown*
gestione orario evento	*trainingSenzaTimeSel.brown*	*trainingSenzaTime.brown*

3.2 Testing e statistiche

Ogni modello è stato valutato rispetto ad un testing set comune che si è formato a partire dalla raccolta automatica dei dati, che parte dalle informazioni di Last.fm e che si estende attraverso Bing. In particolare la costituzione del test set è stata realizzata imponendo, nelle operazioni di interazione con Last.fm, dei luoghi specifici da osservare. Sono state scelte 20 città tra le più rappresentative per ricavare una serie di informazioni certe dalle quali far ampliare la ricerca con il supporto di Bing. Le città selezionate sono: *Roma*, *Londra*, *New York*, *Los Angeles*, *Stoccolma*, *Parigi*, *Helsinki*, *Canberra*, *Chicago*, *Austin*, *Amsterdam*, *Liverpool*, *Boston*, *Detroit*, *Dublino*, *Houston*, *Phoenix*, *Dallas*, *Denver* e *Manchester*.
La ricerca delle istanze da testare è stata estesa nel seguente modo:

- per ciascuna città sono stati ricavati da Last.fm i 10 principali eventi associati;
- da ogni evento sono state estratti 2/3 di informazioni principali (evento e data);
- i 2/3 di informazioni principali sono stati utilizzati come keywords per impostare la ricerca con Bing;
- sono stati selezionati sono i primi 10 risultati restituiti da Bing;
- a partire dai risultati selezionati si sono ricavati i titoli per la formazione del test set;

Avendo quindi impostato 20 città da osservare, 1 evento associato e avendo filtrato i risultati di Bing ai primi 10 elementi, la dimesione del test set analizzato conta potenzialmente 2000 istanze. Nel corso dell'estrazione però è possibile perdere e scartare alcune istanze che non sono idonee per l'elaborazione che il sistema deve effettuare. Per questo motivo delle 2000 istanze potenziali, solo circa 500 sono state scelte per la definizione del test set.

Per ogni modello analizzato sono state prodotte delle statistiche in grado di valutare analiticamente l'efficacia di ciascuna applicazione. Per quanto riguarda le statistiche rispetto alla fase di training, il lavoro è stato supportato da alcune componenti di Jitar già predefinite in grado di fornire un valore di accuratezza rispetto all'andamento dell'addestramento. I risultati forniti hanno permesso, sia di tener traccia della valutazione di ciascun *fold* della Cross Validation, sia di osservare una stima di precisione complessiva di tutta la fase di training (*overall accuracy*).
Per la valutazione dell'accuratezza del test set sono stati invece definiti dei parametri specifici. Tale definizione ha avuto due livelli di osservazione:

- analisi rispetto al dominio delle istanze osservate;
- analisi generale rispetto all'intero test test;

L'osservazione a livello di dominio è relativa a delle statistiche associate a sottoinisiemi di istanze del test set, raggruppate sulla base del dominio di appartenenza. Questa tipologia di statistiche sono ad esempio relative a tutte le istanze

del test set, ovvero a tutti i titoli di siti, estratti a partire da *songkick.com*, una delle principali pagine web che fornisce informazioni in merito ad eventi musicali. Tali stime sono state fornite in modo da permettere l'osservazione dell'andamento dell'accuratezza rispetto alla diversa tipologie di titoli, e quindi domini, che sono stati studiati. Le valutazioini fornite in merito a questo approccio, per ogni dominio, sono relative al numero di informazioni principali presenti. Per informazioni principali si intendono la presenza dell'evento, della data e del luogo. Per ciascun dominio sono quindi riportate:

- % di volte che sono taggate 3/3 informazioni principali;
- % di volte che sono taggate 2/3 informazioni principali;
- % di volte che sono taggate 1/3 informazioni principali;
- % di volte che sono taggate 0/3 informazioni principali;

L'osservazione generale rispetto all'intero test set ha permesso di fornire statistiche meno focalizzate ma di carattere più complessivo. Per l'intero insieme di istanze sono state fatte valutazioni di diversa natura. Inizialmente per l'intero test set è stato definito un valore percentuale di precisione definito nel seguente modo:

$$\textbf{score}(\%) = \frac{puntiTestSet * 100}{max\_puntiTestSet}$$

dove *puntiTestSet* è ottenuto analizzando l'intero test set e sommando, per ogni istanza, il valore 3, 2, 1 o 0 in base al numero di informazioni principali taggate. Un titolo in cui sono presenti tutte e 3 le informazioni darà contributo 3, mentre ad esempio uno privo della sola data darà contributo 2. Il valore *max_puntiTestSet* è invece calcolato considerando il massimo punteggio che un test set può raggiungere, considerando che in ogni sua istanza possano essere presenti tutte e 3 le informazioni principali; considerando quindi un contributo di 3 da parte di tutte le istanze del test set. Ad esempio su un test set di 60 titoli, il *max_puntiTestSet* è di 60*3 = 180 punti, con un conseguente **score** pari al 100%.
Un' ulteriore tipo di valutazione effettuata è relativa alla misurazione della quantità di informazioni rilevanti rispetto a quelle principali. Per l'intero test set sono quindi riportate:

- % di volte che sono taggate 3/3 informazioni principali;
- % di volte che sono taggate 2/3 informazioni principali;
- % di volte che sono taggate 1/3 informazioni principali;
- % di volte che sono taggate 0/3 informazioni principali;

Una successiva analisi statistica ha permesso di calcolare invece, rispetto all'interno test set, la percentuale di volte in cui è stato possibile rilevare l'evento, il luogo e la data nello specifico. Per l'intero test set sono quindi riportate:

- % di volte in cui è stato taggato l'evento;

- % di volte in cui è stato taggato il luogo;
- % di volte in cui è stata taggata la data;

Per ogni modello descritto in precedenza sono quindi state fornite le seguenti valutazioni statistiche, in grado di stimare l'accuratezza del sistema rispetto ai vari training set e quindi sulla base di diversi *corpus.model*.

3.3 Confronto tra modelli

3.3.1 Valutazioni rispetto all'intero test set

Il primo *corpus.model* che è stato realizzato e analizzato è quello base, costituito a partire dal training costituito con le regole di POST definite inizialmente (con l'aggiunta del tag **SEPA**). Il confronto, tra la versione generata a partire dal training selezionato e quello più esteso, ha prodotto i seguenti risultati per quanto riguardo la fase di training:

	trainingSepaSel.brown	*trainingSepa.brown*
overall accuracy (cross validation)	91,25%	88,80%

Mentre per quanto riguarda le statistiche relative all'intero test set:

	trainingSepaSel.brown	*trainingSepa.brown*
Score	79,9%	81,1%
3/3 info principali	54,1%	53,4%
2/3 info principali	31,8%	36,4%
1/3 info principali	13,1%	10,2%
0/3 info principali	0.963%	0%
evento	66,9%	63,2%
luogo	73,0%	74,8%
data	81,9%	78,2%

Successivamente si è pensato di osservare come potesse variare l'accuratezza escludendo completamente la punteggiatura dai titoli dei siti analizzati. In questo modo è stato realizzato il *corpus.model* a partire da un training set che non contemplasse la presenza dei segni di interpunzione al proprio interno. I risultati prodotti in fase di training sono stati:

	trainingSenzaPuntSel.brown	*trainingSenzaPunt.brown*
overall accuracy (cross validation)	79,99%	85,78%

Mentre per quanto riguarda le statistiche relative all'intero test set:

	trainingSenzaPuntSel.brown	*trainingSenzaPunt.brown*
Score	80,8%	82,7%
3/3 info principali	52%	56,8%
2/3 info principali	38,3%	34,3%
1/3 info principali	9,63%	8,86%
0/3 info principali	0%	0%
evento	58,2%	63,8%
luogo	80,2%	77,8%
data	73,8%	78,8%

L'importanza della punteggiatura all'interno dei titoli è però una caratteristica testuale che deve essere considerata soprattutto per la sua capacità di separare e definire la posizione delle informazioni. Per questo è stato definito un *corpus.model* a partire da un training in cui ogni elemento di interpunzione fosse taggato con se stesso. I risultati prodotti in fase di training sono stati:

	trainingTaggaPunSel.brown	*trainingTaggaPun.brown*
overall accuracy (cross validation)	84,43%	88,93%

Mentre per quanto riguarda le statistiche relative all'intero test set:

	trainingTaggaPunSel.brown	*trainingTaggaPun.brown*
Score	83,0%	78,8%
3/3 info principali	56,5%	50,5%
2/3 info principali	36,4%	36,0%
1/3 info principali	6,74%	12,5%
0/3 info principali	0,385%	0%
evento	62,0%	63,6%
luogo	86,5%	70,1%
data	80,2%	80,2%

Le valutazioni sui modelli successivi si sono invece focalizzate su dettagli testuali specifici come ad esempio la presenza e la gestione delle parentesi all'interno dei titoli. Questo ha permesso di creare un *corpus.model* a partire da un training set specifico che fosse in grado di trattare con cura questa particolarità testuale. I risultati prodotti in fase di training sono stati:

	trainingParentesiSel.brown	*trainingParentesi.brown*
overall accuracy (cross validation)	84,53%	88,93%

Mentre per quanto riguarda le statistiche relative all'intero test set:

	trainingParentesiSel.brown	*trainingParentesi.brown*
Score	82,3%	83,6%
3/3 info principali	56,1%	58,6%
2/3 info principali	35,5%	33,5%
1/3 info principali	7,9%	7,9%
0/3 info principali	0,578%	0%
evento	62,6%	64,5%
luogo	82,3%	79,0%
data	81,3%	77,3%

L'ultimo modello prodotto si è interessato della gestione, all'interno delle istanze, delle espressioni legate agli orari degli eventi. Tale informazione, in grado di condizionare l'accuratezza delle elaborazioni, è stata esclusa dai tag identificativi della data. Si è quindi prodotto un *corpus.model* da un training in cui tali espressioni fossero trattate secondo quest'approccio. I risultati prodotti in fase di training sono stati:

	trainingSenzaTimeSel.brown	*trainingSenzaTime.brown*
overall accuracy (cross validation)	84,29%	89,20%

Mentre per quanto riguarda le statistiche relative all'intero test set:

	trainingSenzaTimeSel.brown	*trainingSenzaTime.brown*
Score	82,1%	83,9%
3/3 info principali	55,7%	59%
2/3 info principali	35,6%	33,7%
1/3 info principali	8,09%	7,32%
0/3 info principali	0,578%	0%
evento	62,6%	64,7%
luogo	82,5%	79,0%
data	81,1%	78,0%

I risultati associati ad ogni modello mostrano che, per quanto riguarda il numero di informazioni principali taggate correttamente, in tutti i test le percentuali più alte sono associate al caso in cui sono rilevate 3/3 informazioni. Seguono subito dopo i casi in cui le informazioni rilevanti matchate sono 2/3, mentre i casi in cui nessuna delle informazioni principali è identificata non superano mai valori del 1%. Analizzando invece, per l'intero test set, le volte in cui è possibile taggare in modo specifico l'evento, la data e il luogo, è emerso che le due informazioni che più frequentemente risultano rilevate in modo corretto sono il luogo e la data con range di valori che oscillano tra il 70% e l'80%. Mentre la rilevazione dell'evento si attesta mediamente intorno al 65%. Tali risultati sono dovuti al modo in cui sono espressi tali tipi di informazioni all'interno dei titoli dei siti. I dati associati all'evento, ovvero il nome del cantante o della manifestazione musicale, sono più facilmente soggetti a fraintedimenti e ambiguità di natura

linguistica rispetto alle altre tipologie di informazioni.

Paragonando tutti i modelli testati, il confronto mostra che con il raffinamento del modello, la prestazioni del sistema tendono a migliorare. In particolare per quanto riguarda l'*overall accuracy*, il valore indice della fase di addestramento, è possibile notare come la gestione di un numero più consistente di istanze, fa crescere la precisione. Questo a dimostrazione del fatto che un maggior numero di esempi di training, fornisce al sistema una base conoscitiva più ampia dalla quale apprendere informazioni importanti per la definizione del modello. Con un andamento meno incisivo, ma pur sempre incrementale, con l'aumentare delle istanze aumenta anche il valore di score per ogni modello. Tale risultato testimonia che con un addestramento più accurato, anche le prestazioni in fase di test migliorano, portando alla rilevazione corretta di più informazioni. La tabella che segue mostra il confronto tra i vari modelli, per quanto riguarda i soli valori complessivi di accuratezza (in fase di training) e score (in fase di testing).

	overall accuracy	**score**
trainingSepaSel.brown	91,25%	79,9%
trainingSepa.brown	88,80%	81,1%
trainingSenzaPuntSel.brown	79,99%	80,8%
trainingSenzaPunt.brown	85,78%	82,7%
trainingTaggaPuntSel.brown	84,43%	83,0%
trainingTaggaPunt.brown	88,93%	78,8%
trainingParentesiSel.brown	84,53%	82,3%
trainingParentesi.brown	88,93%	83,6%
trainingSenzaTimeSel.brown	84,29%	82,1%
trainingSenzaTime.brown	89,20%	83,9%

3.3.2 Valutazioni rispetto ai domini

Un'analisi più dettagliata rispetto ai risultati ha permesso anche di evidenziare le prestazioni rispetto ai domini più frequentemente riscontrati all'interno del test set. Di seguito sono mostrate le stime associate all'analisi dei principali domini rintracciati e le relative valutazioni prestazionali dipendenti dai vari modelli utilizzati. Per questione di visibilità le prestazioni dei vari modelli sono separti in due tabelle: nella prima sono presenti i risultati associati ai training selezionati, mentre nella seconda quelli associali ai training con un numero più ampio di istanze. Inoltre per rendere più leggibili i risultati, all'interno delle tabelle, sono stati utilizzati al posto dei nomi dei modelli, delle lettere identificative spiegate nella leggenda sottostante.

Leggenda:

A=*trainingSepaSel.brown*	B=*trainingSepa.brown*
C=*trainingSenzaPuntSel.brown*	D=*trainingSenzaPunt.brown*
E=*trainingTaggaPuntSel.brown*	F=*trainingTaggaPunt.brown*
G=*trainingParentesiSel.brown*	H=*trainingParentesi.brown*
I =*trainingSenzaTimeSel.brown*	L=*trainingSenzaTime.brown*

Di seguito è mostrato il confronto tra l'applicazione dei vari modelli, focalizzando l'attenzione sui domini più frequentemente riscontrati. I modelli considerati in questo caso sono quelli generati a partire dai training selezionati.

dominio	#	info	A	C	E	G	I
ticketfly	22	(3/3)	59,1%	45,5%	40,9%	45,5%	40,9%
		(2/3)	31,8%	50%	50%	54,5%	59,1%
		(1/3)	9,09%	4,55%	9,09%	0%	0%
		(0/3)	0%	0%	0%	0%	0%
last.fm	55	(3/3)	43,6%	70,9%	70,9%	67,3%	67,3%
		(2/3)	52,7%	29,1%	27,3%	27,3%	27,3%
		(1/3)	3,64%	0%	1,82%	5,45%	5,45%
		(0/3)	0%	0%	0%	0%	0%
songkick	49	(3/3)	71,4%	38,8%	61,2%	61,2%	61,2%
		(2/3)	22,4%	42,9%	38,8%	38,8%	38,8%
		(1/3)	6,12%	18,4%	0%	0%	0%
		(0/3)	0%	0%	0%	0%	0%
concertful	19	(3/3)	100%	63,6%	78,9%	78,9%	78,9%
		(2/3)	0%	36,8%	21,1%	21,1%	21,1%
		(1/3)	0%	0%	0%	0%	0%
		(0/3)	0%	0%	0%	0%	0%
bandsintown	55	(3/3)	38,2%	63,2%	63,6%	63,6%	63,6%
		(2/3)	27,3%	32,7%	32,7%	32,7%	32,7%
		(1/3)	9,09%	3,64%	3,64%	3,64%	3,64%
		(0/3)	0%	0%	0%	0%	0%
ticketnetwork	22	(3/3)	50%	59,1%	45,5%	45,5%	45,5%
		(2/3)	27,3%	36,4%	50%	50%	50%
		(1/3)	22,7%	4,55%	4,55%	4,55%	4,55%
		(0/3)	0%	0%	0%	0%	0%

Questi confronti tra i domini permettono di evidenziare quale tipologia di titoli fornisce le prestazioni migliori se sottoposta al sistema. Emerge ad esempio che per i titoli appartenenti al dominio *bandsintown* il passaggio dal primo al secondo modello selezionato, ha portato un miglioramento consistente che si è andato a stabilizzare con le modifiche successive. In altri casi come ad esempio per i domini *ticketfly* e *ticketnetwork* l'andamento è analogamente opposto. Inoltre, come già è stato osservato nei test precedenti, i valori percentuali più bassi si riscontrano nei casi di identificazione di una o nessuna informazione rilevante.

Di seguito, invece, è mostrato il confronto tra l'applicazione dei vari modelli, focalizzando l'attenzione sui domini più frequentemente riscontrati. I modelli considerati in questo caso sono quelli generati a partire dai training aventi un numero maggiore di istanze.

dominio	#	info	B	D	F	H	L
ticketfly	22	(3/3)	40,9%	50%	36,4%	50%	54,5%
		(2/3)	50%	40,9%	40,9%	45,5%	40,9%
		(1/3)	9,09%	9,09%	18,2%	4,55%	4,55%
		(0/3)	0%	0%	4,55%	0%	0%
last.fm	55	(3/3)	40%	72,7%	16,4%	72,7%	72,7%
		(2/3)	50,9%	27,3%	69,1%	27,3%	27,3%
		(1/3)	9,09%	0%	14,5%	0%	0%
		(0/3)	0%	0%	0%	0%	0%
songkick	49	(3/3)	65,3%	73,5%	57,1%	65,3%	67,3%
		(2/3)	32,7%	22,4%	36,7%	32,7%	30,6%
		(1/3)	2,04%	4,08%	6,12%	2,04%	2,04%
		(0/3)	0%	0%	0%	0%	0%
concertful	19	(3/3)	100%	89,5%	100%	100%	100%
		(2/3)	0%	10,5%	0%	0%	0%
		(1/3)	0%	0%	0%	0%	0%
		(0/3)	0%	0%	0%	0%	0%
bandsintown	55	(3/3)	61,8%	63,6%	69,1%	67,3%	67,3%
		(2/3)	36,4%	32,7%	29,1%	30,9%	30,9%
		(1/3)	1,82%	3,64%	1,82%	1,82%	1,82%
		(0/3)	0%	0%	0%	0%	0%
ticketnetwork	22	(3/3)	50%	50%	50%	50%	45,5%
		(2/3)	27,3%	36,4%	31,8%	36,4%	40,9%
		(1/3)	22,7%	13,6%	18,2%	13,6%	13,6%
		(0/3)	0%	0%	0%	0%	0%

Da queste valutazioni emerge come la considerazione di più istanze in fase di addestramento permette di raffinare la precisione dei risultati associaiti ai domini più frequenti. In particolare ciò è evidente per il dominio *concertful* che in tutti i modelli considerati, ad eccezione di uno, vanta un'accurattezza del 100% di titoli perfettamente taggati, con 3 informazioni su 3 rilevate. I risultati permettono inoltre di sottolineare come il raffinamento del modello permetta di migliorare le prestazioni. Questo è visibile ad esempio per il dominio *last.fm* per il quale, considerando il passaggio dal modello B al modello L, si passa dal 40% al 72% nel caso 3/3, dal 50,9% al 27,3% nel caso 2/3 e dal 9,09% allo 0% nel caso 1/3, evidenziando un incremento delle prestazioni che permette di aumentare le volte in cui si rilevano tutte le informazioni correttamente.

3.4 Test

Come ultima attività è stato prodotto un ulteriore caso di studio che è stato addestrato e verificato per fornire ulteriori dati di test.
Per quanto riguarda la formazione del training set sono state selezionate 20 città per l'interazione con Last.fm da cui sono state ricavate informazioni certe per estendere la ricerca. Queste città sono: *Amsterdam*, *London*, *New York*,*Los Angeles*, *Stoccolma*, *Paris*, *Helsinki*, *Canberra*, *Chicago*, *Austin*,*Buffalo*, *Newport Beach*, *Olympia*, *Springfield*, *Sacramento*, *Miami*, *Las Vegas*, *Atlanta*, *Philadelphia* e *Salt Lake City*. A partire dai primi 10 eventi musicali certi, programmati per ciascuna di queste città, si è estesa la ricerca su Bing selezionando i primi 10

risultati offerti. Delle potenziali 2000 istanze le operazioni di filtraggio hanno ridotto il numero a 529. Questo training set è stato sottoposto alle regole di POST e convertito con il parser nel formato .brown. Il modello scelto è stato quello che in fase di sperimentazioni aveva mostrato i migliori valori prestazionali, ovvero quello in grado di gestire la presenza di eventuali parentesi e di espressioni legate all'orario. Il training automatico supportato da Jitar ha permesso la realizzazione di un corpus.model rappresentativo delle conoscenze apprese. Il risultato prodotto in fase di training è stato:

$$overall\ accuracy = 96{,}61\%$$

Nello specificato per i valori per ogni fold (10 considerati) della Cross Validation sono:

Fold 0 accuracy: 98.03
Fold 1 accuracy: 97.89
Fold 2 accuracy: 96.67
Fold 3 accuracy: 96.06
Fold 4 accuracy: 96.86
Fold 5 accuracy: 96.60
Fold 6 accuracy: 95.48
Fold 7 accuracy: 94.39
Fold 8 accuracy: 96.65
Fold 9 accuracy: 97.45

Tale valore è risultato superiore rispetto ai casi precedentemente analizzati.

Il test set è stato formato in modo analogo al training set, selezionato ulteriori altre 20 città. In questo caso sono state scelte: *Roma*, *Liverpool*, *Boston*, *Detroit*, *Dublino*, *Houston*, *Phoenix*, *Dallas*, *Denver*, *Manchester*,*Birmingham*, *Leeds*, *Leicester*, *Newcastle*, *Portsmouth*, *Nottingham*,*Southampton*, *Cardiff*, *Glasgow* e *Edimburgo*. Questa volta, delle potenziali 2000 istanze per il test set, ne sono state filtrate solo 481. Il risultato generale prodotto nella fase di testing è stato:

$$score = 84.6\%$$

Mentre le valutazioni, sempre relative all'intero test set, di carattere più specifico hanno evidenziato le seguenti prestazioni:

3/3 info principali	58,2%
2/3 info principali	37,8%
1/3 info principali	4,37%
0/3 info principali	0%

evento	59,5%
luogo	73,8%
data	70,3%

Coerentemente con i risultati analizzati in precedenza, le migliori percentuali si riscontrano nei casi di corretto matching di tutte e tre le informazioni principali,

mentre i casi di mancata rilevazione sono dello 0%. Anche per quanto riguarda il tagging dei dati specifici,come in precedenza, i risultati più alti sono correlati alla corretta identificazione di luogo e data, mentre leggermente meno precisi risultano essere quelli legati all'evento.

Per quanto riguarda invece i risultati, valutati a livello di dominio delle istanze osservate, sono emerse le seguenti considerazioni rispetto ai siti più frequentemente riscontrati:

dominio	#	info	*corpus.model*
ticketfly	13	(3/3)	69,20%
		(2/3)	23,1%
		(1/3)	7,69%
		(0/3)	0%
last.fm	49	(3/3)	59,2%
		(2/3)	40,8%
		(1/3)	0%
		(0/3)	0%
songkick	47	(3/3)	59,6%
		(2/3)	38,3%
		(1/3)	2,13%
		(0/3)	0%
stereoboard	41	(3/3)	68,3%
		(2/3)	31,7%
		(1/3)	0%
		(0/3)	0%
bandsintown	56	(3/3)	67,9%
		(2/3)	26,8%
		(1/3)	5,36%
		(0/3)	0%
ticketnetwork	13	(3/3)	53,8%
		(2/3)	46,2%
		(1/3)	0%
		(0/3)	0%

Anche in questo caso si nota come, la combinazione dei casi in cui sono rilevate 3 e 2 informazioni rispetto alle 3 complessive, porti a risultati considerevoli, che in presenza di alcuni domini arriva all'accurezza massima ottenibile.

Riferimenti

[1] L'intero progetto è disponibile presso la repository GitHub: https://github.com/LM7/SistemiIntelligentiPerInternet

[2] **Stanford Temporal Tagger, SUTime**: http://nlp.stanford.edu/software/sutime.shtml

[3] **Stanford Named Entity Recognizer (NER)**: http://nlp.stanford.edu/software/CRF-NER.shtml

[4] **TagMe**: http://tagme.di.unipi.it/

[5] **LIBLINEAR**: http://www.csie.ntu.edu.tw/ cjlin/liblinear/

[6] **Weka**: http://www.cs.waikato.ac.nz/ml/weka/

[7] **Last.fm**: http://www.lastfm.it/

[8] **boilerpipe**: http://boilerpipe-web.appspot.com/

[9] **brown**: http://en.wikipedia.org/wiki/Brown_Corpus

[10] **Jitar**: https://github.com/danieldk/jitar;