

1st Plone for Research and University Day - Bologna, 20 Maggio 2010

COACH - Un workbench per l'analisi dei testi e l'estrazione di termini

Giovanni Toffoli (toffoli at uni.net) - LINK srl, Roma

Stefano Lariccia (stefano.lariccia at uniroma1.it) - Università di Roma "La Sapienza"

OBBIETTIVI

Disporre anche per l'italiano di una **suite di strumenti** e di **risorse linguistiche** che consentano di

- effettuare una migliore **indicizzazione full-text** dei documenti
- estrarre da essi **termini** significativi, candidati ad essere usati per il **tagging** dei contenuti stessi (nel contesto di blog, digital library, ecc.)..

Come abbiamo accennato in un altro contributo, l'estrazione di termini da un documento, o da un intero corpus, può costituire un passo verso

- la **classificazione semantica** del documento, se si dispone di una struttura di riferimento, come un *thesaurus*
- la costruzione di un *glossario* o di un'*ontologia di dominio*.

IL PROGETTO

Si tratta di *work-in-progress*, consistente in

- adattamento, estensione e
- integrazione in Plone

di strumenti per l'elaborazione di testi.

Focus sulla **lingua italiana**

- tool di questo tipo sono disponibili per la lingua inglese, sia sotto forma di *librerie/package* per Python o Plone, sia sotto forma di *servizi su web*
- ma sono praticamente inesistenti per l'italiano.

L'APPROCCIO

I principali componenti che intendiamo integrare in Plone sono:

NLTK (Natural Language ToolKit)

- una libreria di package Python
- un insieme di risorse linguistiche, soprattutto corpora

Risorse per l'italiano

- di libero dominio (o quasi)
- di ottima qualità

PERCHÉ PYTHON?

Crediamo che **Python** sia il linguaggio di programmazione *general-purpose* che meglio supporta la sperimentazione nel campo dell'analisi dei testi.

PERL

- è il linguaggio tradizionalmente più usato in questo campo
- ma è un linguaggio di scripting specialistico e dalla sintassi criptica

C++, Java

- sono ambienti troppo estesi
- il ciclo di sviluppo è pesante.

PERCHÉ PLONE?

Plone di per sé non aggiunge un contributo essenziale ad una base che includa

- Python, NLTK, risorse linguistiche adeguate.

Ma Plone può offrire sinergie non trascurabili:

- un sito Plone potrà beneficiare del package in corso di sviluppo ai fini di *indicizzazione e tagging*
- la piattaforma Zope/Plone è un'ottima base di partenza per un'applicazione che deve essere *accessibile su web*
- le funzionalità di *content management* e di *user management* di Plone facilitano la gestione di risorse linguistiche personalizzate (es: basi di documenti e file di parametri).

PROSPETTIVE DEL PROGETTO - 1

L'attività descritta intende essere un elemento di **aggregazione**

- da un punto di vista tecnico-realizzativo
- da un punto di vista della costituzione di una comunità di ricerca.

Il prodotto `link.nltk`

Quanto al primo punto di vista, per cui LINK srl è l'attuale riferimento

- stiamo sviluppando un package che provvisoriamente si chiama `link.nltk`, con richiamo alla società LINK srl e al package NLTK
- gradiremo ricevere contributi nella revisione delle specifiche e nella prosecuzione dello **sviluppo e collaudo del package**.

PROSPETTIVE DEL PROGETTO - 2

Il progetto COACH

Attualmente "**La Sapienza**" è il principale promotore della costituzione di una comunità di ricerca, che intende

- lanciare iniziative su cui richiedere contributi pubblici e privati al finanziamento, nell'ambito di **consorzi** e di **programmi di ricerca nazionali ed europei**
- definire e promuovere forme di collaborazione con enti che diffondono la **cultura italiana nel mondo**

Alcuni obiettivi:

- far tesoro delle risorse generate, nell'ambito di attività autonome, da ciascun membro della comunità
- un esempio: costituire corpora settoriali per l'esame di testi non contemporanei, introducendo variabili e modelli che consentano l'identificazione su una scala diacronica dei testi sottoposti a indagine.

CHE COSA E' NLTK - 1

Dal sito del progetto (<http://www.nltk.org>):

"Open source **Python modules, linguistic data and documentation** for research and development in natural language processing and text analytics, with distributions for Windows, Mac OSX and Linux".

Dall'articolo "Multidisciplinary Instruction with the Natural Language Toolkit"

(<http://www.aclweb.org/anthology/W/W08/W08-0208.pdf>):

Il Natural Language Toolkit, o NLTK, è stato sviluppato per dare accesso, ad una ampia gamma di studenti, a conoscenze e abilità di base nel campo del NLP ..

NLTK è una suite di moduli Python distributi sotto la licenza open source GPL.

CHE COSA E' NLTK - 2

(segue la citazione dall'articolo)

NLTK include una vasta raccolta di corpora, documentazione estesa e centinaia di esercizi, che lo rendono unico nel fornire un quadro esauriente per gli studenti che vogliono acquisire una comprensione del linguaggio di tipo computazionale.

La base di codice di NLTK, oltre 100.000 linee di codice Python, include il supporto per l'accesso ai corpora, tokenizzazione, stemming, tagging, chunking, parsing, clustering, classificazione, modellazione del linguaggio, interpretazione semantica .. e molto altro ancora.

Per dare un'idea del suo impatto: NLTK è stato usato in più di 60 corsi universitari in 20 paesi ..

LE RISORSE LINGUISTICHE ITALIANE - 1

In un rapporto del 2002, **Nicoletta Calzolari**, esponente del venerando **CNR-ILC**, scriveva:

"Il ruolo infrastrutturale delle Risorse Linguistiche nell'ambito del TAL [trattamento automatico della lingua] richiede che esse vengano:

- disegnate, costruite, validate in cooperazione con i potenziali utilizzatori ..
- **costruite riutilizzando risorse parziali disponibili**
- armonizzate con le risorse di altre lingue europee e non ..
- valutate con metodologie riconosciute a livello internazionale
- **messe a disposizione della intera comunità nazionale**
- mantenute e aggiornate ..

Rif.: www.isticom.it/documenti/rivista/2002_065.pdf

LE RISORSE LINGUISTICHE ITALIANE - 2

Ci aspettiamo che nel medio termine la comunità di ricerca proposta divenga una fonte di risorse linguistiche di qualità: corpora, lessici, metodi di analisi, algoritmi, ecc.

Quanto al breve termine, LINK ha effettuato una ricerca rapidissima, informale - e certamente incompleta - sul panorama italiano dei gruppi di ricerca di linguistica computazionale (LC); come noto, sono presenti numerosi centri di eccellenza, ma poco emerge del contributo che essi potrebbero dare a iniziative quale la nostra.

Siamo arrivati alla conclusione preliminare che, per concretezza e apertura, si distinguono i gruppi di

- Università di Trento a Rovereto (rif. Marco Baroni)
- Università di Bologna a Forlì (rif. Eros Zanchetta)

LE RISORSE LINGUISTICHE ITALIANE - 3

Per esempio

- sul web si trovano facilmente materiali didattici del prof. Baroni sui passi di elaborazione propedeutici alla "estrazione di termini": *tokenizzazione*, *POS-tagging*, *chunking*, ecc.
- in particolare si può scaricare liberamente un ampio *lessico morfologico*, **morph-it**, compilato qualche anno fa dal gruppo del prof. Baroni come sottoprodotto della costruzione di un ampio corpus di articoli dal quotidiano "**La Repubblica**"
- sul web è illustrata un'iniziativa internazionale, che ha i gruppi summenzionati tra i promotori, e che ha già prodotto per 3-4 lingue, tra cui l'Italiano, vasti corpora *bilanciati* e *annotati*, usando una metodologia ben documentata di estrazione dei documenti dal web:
<http://wacky.sslmit.unibo.it/doku.php>

CHE COSA C'ENTRA PLONE - 1

Plone parte ben piazzato:

- rappresenta il testo nel formato universale *Unicode*
- supporta l'interazione con l'utente (tramite il browser) e il sistema operativo nelle codifiche più moderne, come *utf-8*

Il core di Plone include, tra l'altro (così, alla rinfusa)

- estrazione del *plain-text* da contenuti nei formati HTML, PDF, Microsoft-Office, OpenOffice, ecc., secondo un'architettura estensibile a *plug-in*
- ricerca full-text su tutti i contenuti testuali, in diverse modalità: è di effetto la ricerca istantanea, *LiveSearch*
- **indici full-text** nativi, del tipo *ZCTextIndex*, costruiti e utilizzati nel contesto di un'**architettura modulare**: sia il testo da indicizzare, sia la query, sono sottoposte ad un *pipeline* di passi elaborativi.

CHE COSA C'ENTRA PLONE - 2

Plone offre il supporto multi-lingua

- nella gestione dell'interfaccia utente (UI)
- nella gestione di versioni in lingua dei contenuti.

A livello di base, Plone

- implementa un sofisticato meccanismo di "negoziamento" per decidere di volta in volta quale lingua usare, quando le impostazioni del sito e quelle del browser prevedono diverse opzioni
- consente di annotare i contenuti con metadati che specificano se contengono testo in una lingua specifica.

Tuttavia in Plone il supporto multi-lingua non si estende al trattamento del testo e all'indicizzazione dei contenuti.

ALCUNE ESTENSIONI DI PLONE - 1

Esaminiamo rapidamente alcuni "prodotti" che hanno qualche connessione con il nostro progetto:

- LinguaPlone
<http://pypi.python.org/pypi/Products.LinguaPlone>
- TextIndexNG
<http://pypi.python.org/pypi/Products.TextIndexNG3>
- SemanticIndex
<http://www.linkback.net/products/semanticindex>
- topia.termextract
<http://pypi.python.org/pypi/topia.termextract/>
- collective.classification
<http://pypi.python.org/pypi/collective.classification/0.1b1>

ALCUNE ESTENSIONI DI PLONE - 2

LinguaPlone

LinguaPlone non appartiene al nucleo Plone, ma è una delle sue estensioni di più "venerande":

- supporta in particolare il site administrator e i redattori nello strutturare un sito multi-lingua e nel disciplinare le attività di traduzione dei contenuti
- si integra con il meccanismo di negoziazione della lingua e consente di specificare la lingua di un contenuto con una granularità più fine.

A noi LinguaPlone non piace molto:

- perché è un'estensione complessa e "intrusiva"
- ma forse non l'apprezziamo non avendo molta esperienza di gestione di siti multi-lingua di grandi dimensioni.

ALCUNE ESTENSIONI DI PLONE - 3

TextIndexNG

TextIndexNG è un'altra estensione "veneranda", più correlata di LinguaPlone con il nostro progetto. Intende rimpiazzare in modo indolore il modulo ZCTextIndex, fornendo un supporto avanzato a indicizzazione e ricerca full-text; principali caratteristiche:

- supporto multi-lingua
- ampia scelta di plug-in per sfruttare l'architettura a pipeline già abbozzata in ZCTextIndex

I plug-in includono, tra l'altro, per diverse lingue

- risorse quali lessici, elenchi di "stopword", ecc.
- moduli di elaborazione quali splitters, stemmers, normalizers, ecc.

Documentazione un po' datata ma molto interessante in:

http://www.linkback.net/doc/searching_plone.pdf/view

ALCUNE ESTENSIONI DI PLONE - 4

SemanticIndex (prodotto non pubblicato)

SemanticIndex, sviluppato da LINK originariamente per l'applicazione **Interop-KMap**, poggia su PloneSaurus e supporta la classificazione semantica, manuale e automatica, dei contenuti di un sito Plone, con riferimento ad una o più tassonomie di termini-concetti

In modalità automatica o semi-automatica (supervised), SemanticIndex estrae i termini mediante confronto del testo con i label lessicali della tassonomia; nel fare ciò usa

- **xpdf** per la conversione in plain text del PDF
- **TreeTragger**, free ma non open source, per estrarre frammenti di testo da cui estrarre eventuali candidati www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger
- algoritmi elaborati dal gruppo della prof. **Velardi** (La Sapienza - Dip. di Informatica) per valutarne la rilevanza.

ALCUNE ESTENSIONI DI PLONE - 5

topia.termextract

Si tratta di un prodotto molto recente, sperimentale, che ha l'obiettivo specifico di assistere nell'estrazione di "termini" rappresentativi di un contenuto.

Dalla homepage del package:

- questo package identifica i termini "importanti" all'interno di un contenuto testuale
- esso implementa l'estrazione dei termini mediante semplici algoritmi di POS-tagging (POS = part-of-speech)
- esegue una semplice analisi statistica per individuare i termini e assegnare ad essi un "peso" (rilevanza).
- **topia.termextract** produce risultati di qualità intermedia tra un POS-tagger come TreeTagger ed il servizio **Yahoo Keyword Extraction**.

ALCUNE ESTENSIONI DI PLONE - 6

collective.classification

Si tratta di un prodotto ancora più recente, che abbiamo scoperto da pochi giorni, e che condivide con il nostro progetto l'interesse per il **package NLTK**.

collective.classification

- intende fornire un insieme di strumenti per la **classificazione automatica dei documenti**
- è un prodotto in via di sviluppo, con obiettivi di sperimentazione
- attualmente supporta solo **inglese e olandese**
- è "fortemente influenzato da topia.termextract"
- fa uso del package NLTK.

QUALCHE DETTAGLIO SU COACH/LINK.NLTK - 1

Stiamo sviluppando un "prodotto" per Plone.

Siamo già un po' avanti, ma ci fermeremo un po' a

- studiare meglio TextIndexNG
- studiare/testare `topia.termextract` e `collective.classification`, anche per decidere se cercare di unire gli sforzi.

`link.nltk` estende alcune API di NLTK per

- facilitarne l'uso in Plone
- integrare alcune risorse linguistiche italiane (lessici, corpora, ecc.), in particolare quelle sviluppate dai gruppi che fanno riferimento ai proff. Baroni e Zanchetta.

QUALCHE DETTAGLIO SU COACH/LINK.NLTK - 2

link.nltk include già alcuni moduli elaborativi dedicati a

- lessici: creazione di indici, browsing, ecc.
- corpora: importazione, uso per estrazione di lessici, addestramento di algoritmi di learning, ecc.
- estrazione di termini, mediante: tokenizzazione, POS-tagging, riassiemamento dei token in frammenti di testo (chunk), eventuale confronto con terminologie di dominio.

link.nltk include un prototipo di UI per la manipolazione e il browse di lessici e corpora, e per l'estrazione di termini; in linea di massima, per ogni tipo di funzione, si dovrebbe

- poter scegliere tra diverse opzioni: lingua, formato del lessico e/o del corpus, tipo di algoritmo, ecc.
- leggere/scrivere risorse linguistiche, testi e parametri da locale (upload/download tramite browser, campi di testo dei form) e da remoto: filesystem del server, file di Plone.