

Firenze, 25 Settembre

I Big Data e la Linguistica Computazionale Il progetto SEMPLICE

Alessandro Lenci CoLing Lab, Università di Pisa



Dare un "senso" ai Big Data

ATION TO ATI

Noam Chomsky

Big Data is a step forward. But our problems are not lack of access to data, but understanding them. [Big Data] is very useful if I want to find out something without going to the library, but I have to understand it, and that's the problem.



Big Text Data

























Big Text Data = Big² Data



I dati testuali aumentano esponenzialmente la complessità del problema Big Data

- i testi sono essi stessi dati
- ciascun testo contiene a sua volta grandi quantità di dati
 - riferimenti ad altre entità
 - eventi
 - relazioni tra dati intra-testuali
 - relazioni con dati extra-testuali
- i dati contenuti nei testi non sono strutturati

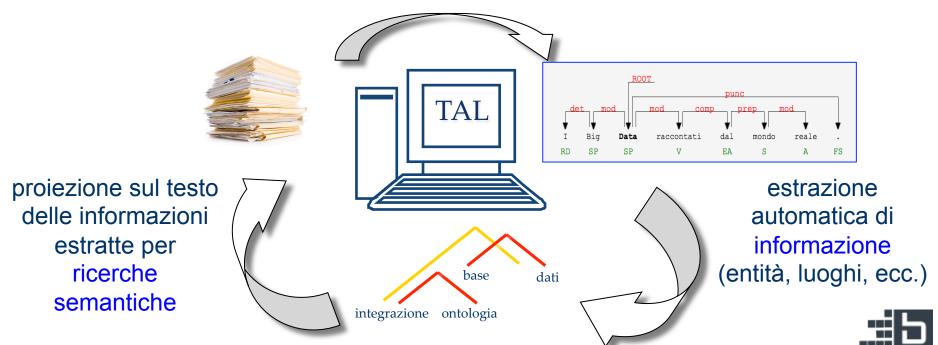


Dal testo alla conoscenza

ATIS Z. 1343

Il Trattamento Automatico della Lingua (TAL)

analisi automatica della struttura linguistica



SEMPLICE











SEMantic Instruments for PubLic Administrators and CitizEns

















SEMPLICE



SEMantic Instruments for PubLic Administrators and CitizEns

Mission

sviluppo di una piattaforma di servizi in ambiente cloud per il supporto alle attività della P.A. nella programmazione e gestione documentale

Stakeholders

Comune di Sansepolcro
Comune di Sesto Fiorentino
Comune di Città di Castello
Unione dei Comuni della
Valtiberina Toscana

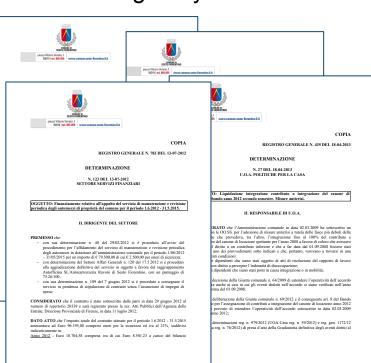


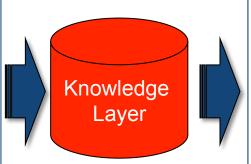




Analisi semantica in SEMPLICE

II Knowledge Layer









aggiudicato provvisoriamente il servizio di cui all'oggetto alla Cooperativa Sociale Il Poliedro,

a soddisfare

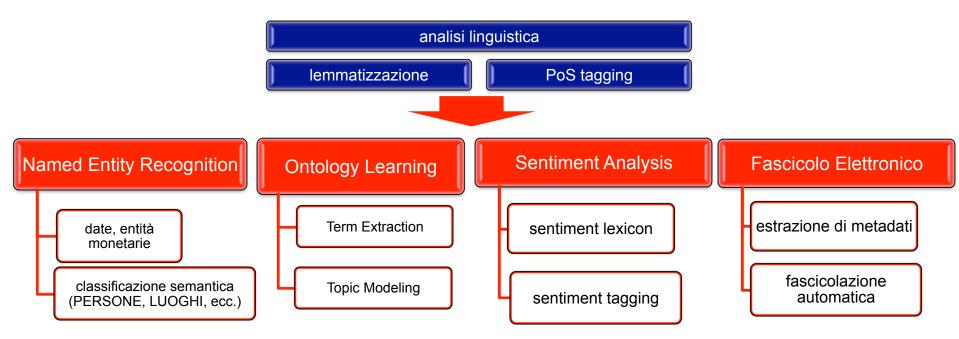
sotto riserva di legge, stante il dettato dell'art. 11 comma 9 ultimo periodo del D. Lgs. 163/2006 e s.m.i., che consente l'esecuzione d'urgenza nei carai in cui la mancata esecuzione immediata della prestazione dedotta nella gara, determinerebbe un grave danno all'interesse pubblico che è destinato



Analisi semantica in SEMPLICE



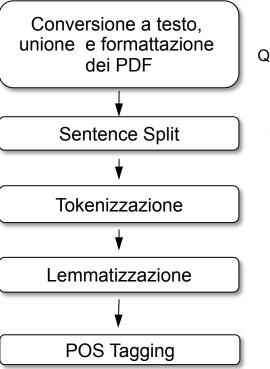
II Knowledge Layer





Pipeline di analisi linguistica





NER

Questo è un esempio di analisi. Il testo poi continua con altre frasi...

Frase 1:Questo è un esempio di analisi.

Frase 2:II testo poi continua con altre frasi...

id	for			
1	Out	id	form	lemma
2	è	1	Questo	questo
3	un	2	è	essere

id	form	lemma	cpos	pos	morph
1	Questo	questo	P	PD	num=s gen=m
2	è	essere	V	V	num=s per=3 mod=i ten=p
3	un	un	R	RI	num=s gen=m
4	esempio	esempio	S	S	num=s gen=m
5	di	di	E	E	
6	analisi	analisi	S	S	num=n gen=f
7			F	FS	

Named Entity Recognition



1	Comunità	Comunità	S	SP	_	B-GPE
2	Montana	Montana	S	SP	_	I-GPE
3	"	F	FB	_	0	I-GPE
4	Valtiberina	Valtiberina	S	SP	num=s gen=f	I-GPE
5	"	F	FB	_	0	I-GPE
1	Via	via	S	S	num=s gen=f	B-LOC
2	s.	S.	S	SP	_	I-LOC
3	Giuseppe	Giuseppe	S	SP	_	I-LOC
4	,	,	F	FF	_	I-LOC
5	32	32	Ν	Ν	_	I-LOC
6	-	-	F	FC	_	0
7	Sansepolcro	Sansepolcro	S	SP	_	B-GPE
8	-	-	F	FC	_	0
1	Determinazione	determinazione	S	S	num=s gen=f	0
2	Dirigenziale	Dirigenziale	S	SP	_	0
3	Numero	Numero	S	SP	_	0
4	1.018	1.018	Ν	Ν	_	0
5	del	di	Е	EΑ	num=s gen=m	0
6	05/11/2010	05/11/2010	N	N		0

Stanford CoreNLP NER (Finker et al. 2005) addestrato sul corpus I-CAB (Magnini et al. 2006)

Persone Luoghi Entità geopolitiche Organizzazioni

Comunità Montana " Valtiberina "

Via s. Giuseppe, 32- Sansepolcro-

Determinazione Dirigenziale Numero 1.018 del 05/11/2010

Direzione Attività Produttive.

Oggetto:

P.s.r. 2007/2013 Fase 3- Bando Misura 112 " insediamento di Giovani Agricoltrori "- Approvazione Istruttorie di Ammissibilita' al Finanziamento e Autorizzazione All' emissione Dei Relativi Atti di Assegnazione.

Dirigente Responsabile: Arch. Siro Veri

Uffici destinatari: SVILUPPO Economico Segreteria

Allegati: il Dirigente

Richiamato il Programma di Sviluppo Rurale della *Toscana* 2007/2013, approvato dalla *Commissione delle Comunità Europee* con Decisione del 16/10/2007 n. C (2007) 4664 e le sue successive modifiche ed integrazioni;

Richiamato il Documento Attuativo Regionale del P.s.r. 2007/2013, approvato con delibera G.r. n. 1319/2009, e le sue successive modifiche ed integrazioni; Richiamato il Decreto n. 515 del 11/02/2010, del Settore Valorizzazione dell' imprenditoria agricola della Regione Toscana, con il quale è stato approvato il bando contenente le disposizioni tecniche e procedurali per la per la concessione del premio previsto dalla misura 112 " Insediamento di giovani agricoltori, relativo alla Fase lii - fondi 2010:

Richiamato il Decreto n. 2079 del 3/05/2010, del Settore Valorizzazione dell' imprenditoria agricola della Regione Toscana, il quale approva modifiche ed integrazioni al bando della misura 112 approvato con decreto 515/2010;

Richiamata la propria determinazione n. 616 del 02/07/2010, con la quale si prende atto della graduatoria preliminare relativa alla misura in oggetto, per la fase 3- fondi 2010 (elenco n. 94/2010) , elaborata dal sistema A.r.t.e.a.;

Ricordato che, sulla base delle risorse disponibili per la misura in oggetto nell' annualità 2010, in tale graduatoria sono state individuate n. 3 (tre) domande finanziabili (Brogialdi Elisa, Pellegrini Marco, Giovagnini Matteo);



Ontology Learning in SEMPLICE

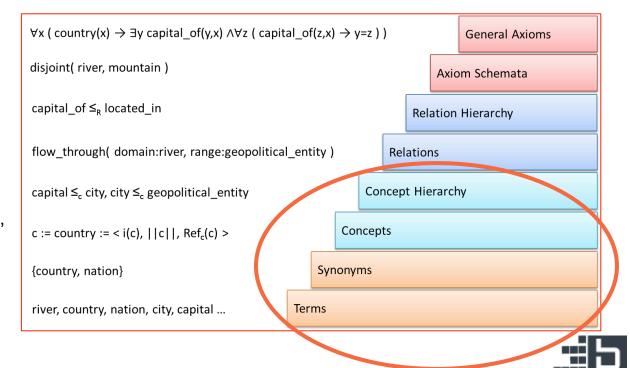


Term extraction

 estrazione automatica di termini semplici e complessi

Topic modeling

 creazione di clusters di termini (e documenti) semanticamente simili, sulla base della loro distribuzione statistica nella collezione



Term Extraction

EXTra (Lenci & Passaro 2014)

Termine	Lmi	Termine normaliz
attestare-v#il-rd#copertura-s	34.550,2314	attestare_il_coper
apporre-v#il-rd#visto-s	22.870,6002	apporre_il_visto
uso-s#amministrativo-a	22.690,1821	uso_amministrativ
imputare-v#il-rd#spesa-s	21.870,4133	imputare_il_spesa
richiamare-v#il-rd#deliberazione-s	18.974,5263	richiamare_il_delit
richiamare-v#il-rd#delibera-s	18.151,1021	richiamare_il_delit
impegnare-v#il-rd#somma-s	14.772,2439	impegnare_il_som
consentire-v#il-rd#copertura-s	13.724,9810	consentire_il_cope
tempo-s#determinato-a	12.343,6810	tempo_determinat
gestione-s#ordinario-a	11.865,9074	gestione_ordinario
funzione-s#di-e#direzione-s	11.547,0920	funzione_di_direzi
fase-s#transitorio-a	10.648,3771	fase_transitorio
verificare-v#il-rd#disponibilità-s	9.140,7242	verificare_il_dispo
dotazione-s#di-e#personale-s	8.777,6550	dotazione_di_pers
votazione-s#unanime-a	8.717,1744	votazione_unanim
risorsa-s#finanziario-a	8.694,6797	risorsa_finanziario
esecuzione-s#di-ea#contratto-s#di-e#servizio-s	8.514,5249	esecuzione_di_co
numero-s#di-ea#intervenuto-s	8.491,1245	numero_di_interve
dotazione-s#strumentale-a	8.044,0166	dotazione_strume
richiamare-v#il-rd#determinazione-s	7.875,9360	richiamare_il_dete
manutenzione-s#ordinario-a	7.702,9214	manutenzione_ord
intervento-s#di-e#assistenza-s#sociale-a	7.685,5150	intervento_di_assi
determinazione-s#dirigenziale-a	7.649,6235	determinazione_di

Registro	Generale	n.	961	Del	26-09-2012	
Copia						

Determinazione n. 140 Del 26-09-2012

Settore Servizi Finanziari

Oggetto: Affidamento del servizio di manutenzione presidi antincendio posti negli edifici di proprietà o ad uso dell'ente; integrazione finanziamento anno 2012. il Dirigente Del Settore Richiamata la deliberazione Consiglio Comunale n. 22 del 14 marzo 2012 di approvazione del bilancio di previsione per l' anno 2012 e pluriennale per il triennio

2012/2014; Richiamata la deliberazione di Giunta Municipale n. 74 del 2 maggio 2012 di approvazione del Piano Esecutivo di Gestione per l' esercizio 2012;

Premesso che: con propria determinazione n. 13 del 25 gennaio 2012 si è proceduto all' affidamento

- del servizio di manutenzione presidi antincendio posti negli edifici di proprietà o ad uso dell' ente per l' anno 2012 alla ditta rf Antincendio di Empoli; con propria determinazione n. 67 del 17 aprile 2012 si è provveduto a finanziare la spesa relativa all' affidamento di cui sopra;

Dato Atto che nel corso dell' anno è stato necessario effettuare alcuni interventi di manutenzione che non rientravano nelle attività ordinarie per cui le risorse impegnate ad oggi non sono sufficienti per pagare i canoni manutentivi del secondo semestre del 2012;

Preso Atto che il servizio di cui sopra è improrogabile ed indifferibile per motivi di sicurezza pubblica e la sua mancata effettuazione potrebbe causare un danno al patrimonio immobiliare e mobiliare dell' ente;

Dato Atto che il codice identificativo di gara Zb5035a530 è stato acquisito con la determinazione di affidamento del servizio; Dato Atto del rispetto di quanto previsto dal D.l. 168 del 12/07/2004 convertito in Legge

191/04 e confermato dal disposto dell' art. 2, comma 573, della l. 244 del 24 dicembre 2007

(Finanziaria per il 2008);

Dato Atto altresì del rispetto di quanto previsto dalla normativa di cui al D.l. n. 95 del 6 luglio 2012, " Disposizioni urgenti per la revisione della spesa pubblica articolo 1) commi 1, 3 e 13, convertito in Legge 7 agosto 2012, n. 135 ";

Richiamato il Regolamento sull' ordinamento degli uffici e dei servizi- parte i-

Regolamento di Organizzazione, approvato con deliberazione della Giunta Municipale n. 109 del 9 giugno 2008, integrato con deliberazione della stessa Giunta Comunale n. 97 del 18 maggio 2009;

Visto il Decreto del Sindaco n. 4 del 30 giugno 2011 di *proroga dell' incarico dirigenziale* di responsabilità del Settore Servizi Finanziari ai sensi dell' art. 110, comma 1), del D.Lgs.

267/2000: Visto l' art. 151, comma 4, del D.lgs. n. 267/2000;

Visto l' art. 4 del vigente regolamento di contabilità;

1) Di integrare il finanziamento di spesa del servizio di manutenzione presidi antincendio posti negli edifici di proprietà o ad uso dell' ente relativamente all' anno in corso;

2) Di dare atto che la somma necessaria per pagare alla ditta rf Antincendio di Empoli i canoni manutentivi del secondo semestre del 2012 ammonta ad Euro 4.885,48 compresa iva;

3) Di impegnare la spesa sul bilancio del corrente esercizio finanziario come segue:

Cap. 36230 Euro 1.000,00

Cap. 9230 Euro 1.000,00

Cap. 16230 Euro 383,34

Cap. 17230 Euro 324,22 Cap. 18230 Euro 1.177,92

Cap. 23230 Euro 1.000,00

Topic Modeling

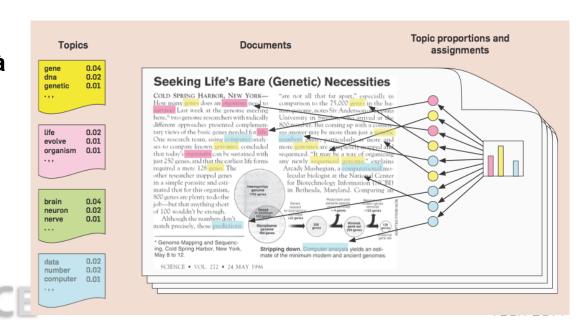


Modelli generativi bayesiani che individuano gli argomenti (topic) di un testo (Blei et al. 2003, Blei 2012)

• un topic è un argomento latente rappresentato da una distribuzione di probabilità delle parole

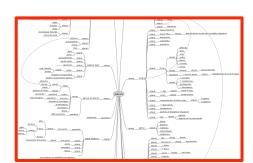
Possibilità di calcolare similarità semantica tra termini e tra documenti

termini (documenti) simili hanno simili distribuzioni nei topics



Topic Modeling in SEMPLICE

ontologia di **SEMPLICE** 01s





topic

alloggio nucleo_familiare

iniziativa manfestazione

bonifica rifiuto

alloggio nucleo_familiare

iniziativa manfestazione

bonifica rifiuto

EDILIZIA

CULTURA TURISMO E **SPORT**

AMBIENTE

topic labeling





Mallet



testi

annotati

con EXTra



Topic Modeling in SEMPLICE



Demografico

Minore Loculo Anno_in_corso C Disponibilità_finanziario_in_cor

Edilizia

Alloggio Nucleo Gestore Auto G

Sicurezza Ribasso Incentivo Peri Netto_di_ribasso Immobile Impre



Alloggio Nucleo Gestore Auto Grad Alloggio_di_edilizia_residei

Sorgente

Sesto Fiorentino

Sesto Fiorentino

Sesto Fiorentino

Sesto Fiorentino

Copia

Registro Generale n. 1308 Del 13-12-2011

Determinazione n. 135 Del 13-12-2011

U.o.a.

Politiche Per la Casa

Oggetto: Assegnazione di alloggio di edilizia residenziale pubblica al nucleo familiare di K.s. posizione n. 110 nella vigente graduatoria. il

Responsabile di U.o.a.

Premesso che da parte del Servizio Politiche della casa è stato avviato il procedimento per la verifica della sussistenza dei requisiti richiesti dalla Legge regionale Toscana n. 96/1996, nei confronti dei concorrenti collocati nella vigente graduatoria di Erp per l'assegnazione dei nuovi alloggi realizzati nella Umi 9 del PI1:

Richiamata la propria determinazione reg. gen. n. 1100/2011 (reg. Uoa Politiche per la casa n. 105/2011) del 08.11.2011 con la quale si stabiliva che 1 dei 16 alloggi dimensionati per nuclei composti da quattro persone veniva riservato alla mobilità e che la scelta degli alloggi sarebbe stata effettuata prioritariamente dal nucleo interessato al cambio di alloggio e successivamente dagli altri nuclei secondo la posizione ricoperta nella vigente graduatoria di

Erp;

Considerato che soddisfatta la riserva di cui sopra è stato individuato quale dodicesimo nucleo familiare avente diritto all' assegnazione di uno degli alloggi di metratura idonea per un nucleo composto da quattro persone quello di K.s. collocato in posizione n. 110 della graduatoria sopracitata approvata con determinazione dirigenziale n. 832/2011;

Dato Atto che con dichiarazione del 06 dicembre 2011 prot. n. 58623, K.s. ha scelto e accettato, tra quelli disponibili, l' alloggio identificato con il numero di interno a2 posto al piano terra, scala a, del fabbricato sito in Largo ix novembre n. 2 di cui sono pertinenze il posto auto e la cantina identificati con il numero a2 e posti al piano interrato del fabbricato;

Rilevato Altresì che all' atto della scelta e formale accettazione dell' alloggio, K.s. ha dichiarato di essere a conoscenza di quanto indicato ai punti 2) e 3) del dispositivo della presente determinazione;

Dato Atto che

- in data 20.06.2011 è stata acquisita agli atti del Comune la relazione tecnica Verifica superamento rischio idraulico a seguito opere realizzate, redatta da S.in.ter. srl e attestata dal dott. ing. Federico Domenichini, estensore dello studio per la messa in sicurezza idraulica del 2004, a corredo dei piani di lottizzazione denominati Pl1- f. 31 Via Petrosa e

Pl13- F.25 via Lazzerini, e della relazione idraulica del 2006 a corredo del progetto esecutivo delle opere di urbanizzazione;

- la suddetta Verifica superamento rischio idraulico a seguito opere realizzate per eventi con tempo di ritorno due centennali, dà conto delle condizioni di sicurezza conseguite, alla data di attestazione del tecnico, rispetto agli interventi previsti e realizzati e quelle che lo saranno al completamento;
 in data 30.11.2011, con lettera, la società S.in.ter srl ha aggiornato lo stato dei lavori effettuati rispetto alla documentazione di cui al sopracitato punto evidenziando in particolare che: a) è stata completata, a eccezione della bocca di sfioro, la vasca di laminazione sul Canale di Cinta orientale; b) è stato effettuato il rialzamento degli argini (sponde) del canale;
- in data 01.12.2011 la società S.in.ter srl ha integrato la comunicazione del 30.11.2011 riepilogando che lo stato di attuazione dei lavori relativi alle opere di urbanizzazione del

Sesto Fiorentino	1324381970871.pdf	EXTra	Entities	0,77
Sesto Fiorentino	1310649324872.pdf	EXTra	Entities	0,77





Fascicolo elettronico

In progress

Estrazione di metadati

Atto, riferimenti ad altri documenti, Oggetto, ecc.

Associazione di documenti amministrativi riferiti a uno stesso procedimento

es. Procedimento di Acquisto

Caso d'uso

Composizione automatica di fascicoli informatici





- con nomina dei componenti:
- con determinazione dirigenziale n.1795 del 31.12.13 si sono approvati i verbali di gara e si è aggiudicato provvisoriamente il servizio di cui all'oggetto alla Cooperativa Sociale Il Poliedro, sotto riserva di legge, stante il dettato dell'art. 11 comma 9 ultimo periodo del D. Lgs. 163/2006 e s.m.i., che consente l'esecuzione d'urgenza nei casi in cui la mancata esecuzione immediata della prestazione dedotta nella gara ,determinerebbe un grave danno all'interesse pubblico che è destinato a soddisfare:





Sentiment analysis

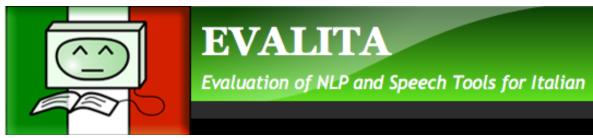


In progress

Identificazione della polarità (positiva, negativa o neutra) di tweets, forum, rassegna stampa della PA

CoLing Lab Tweet Analyzer

Terzo classificato (su 11 sistemi) al Task SENTIPOLC di Evalita 2014





Conclusioni



SEMPLICE mette le potenzialità della linguistica computazionale al servizio della PA

- · fase di sviluppo quasi completata
- trasformazione dei dati estratti in Linked Open Data (Seacom)
- integrazione tra dati estratti dai testi e dati strutturati (Bnova)
- prossimo inizio della sperimentazione con le PA stakeholder

La linguistica computazionale rappresenta un elemento chiave per affrontare la sfida dei Big (Text) Data

- organizzare e gestire i dati testuali
- trasformare i dati non strutturati nei testi in dati espliciti e strutturati

Non solo in macro-contesti, ma anche nelle piccole e medie PA







Pisa, 9-10 dicembre 2014

La prima Conferenza di Linguistica Computazionale Italiana

HOME

CALL FOR PAPERS

SOTTOMISSIONI

ISCRIZIONE PRO

PROGRAMMA

SEDE E ALBERGHI

ORGANIZZAZIONE





CoLing Lab

http://colinglab.humnet.unipi.it









Alessandro Lenci



Gianluca E. Lebani



Lucia Passaro





Grazie



