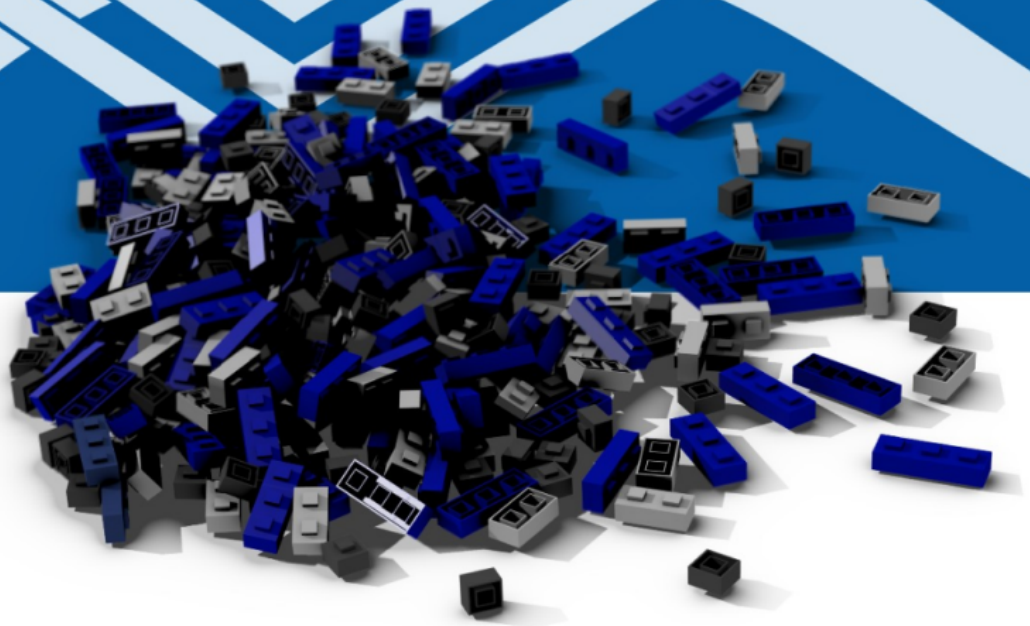
Your Data, Your Creativity!



BIGDATA TECH 2017

Big Text Data

Come estrarre informazione dai dati testuali

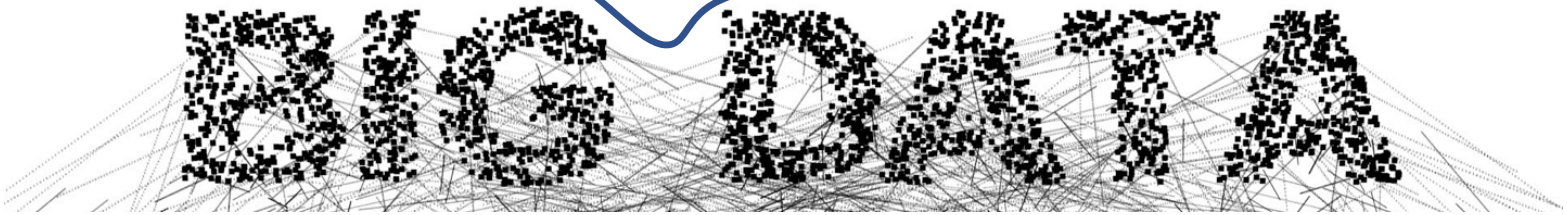
Alessandro Lenci, Lucia Passaro

Computational Linguistics (CoLing) Lab

Università di Pisa

CENTRO SVIZZERO - MILANO - 12 OTTOBRE 2017

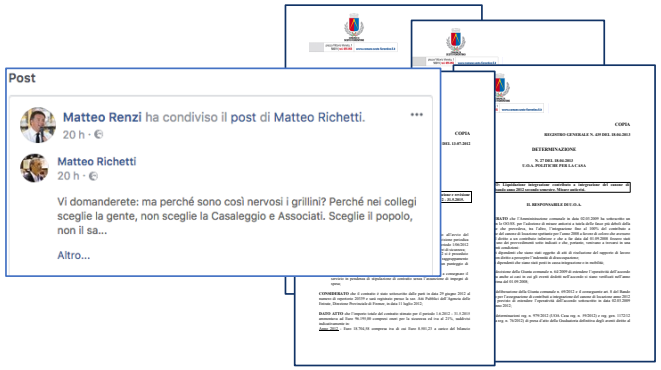
TEXT

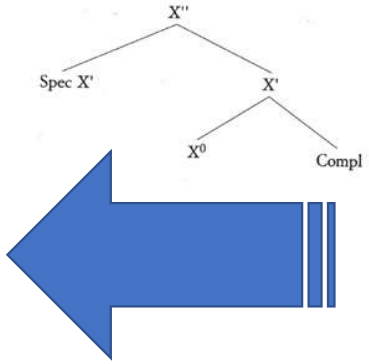

BIG DATA


la Repubblica.it
il mondo in diretta
24 ore su 24
WIKIPEDIA

Le sfide (e opportunità) dei Big Text Data

- Ogni singolo testo è esso stesso una fonte di big data
  - persone, luoghi, date, eventi, ecc.
  - relazioni tra dati intra-testuali
  - relazioni con dati extra-testuali
- Dati non strutturati ad alto tasso di variabilità
- I contenuti informativi sono impliciti e la loro estrazione richiede la "comprensione linguistica" del testo





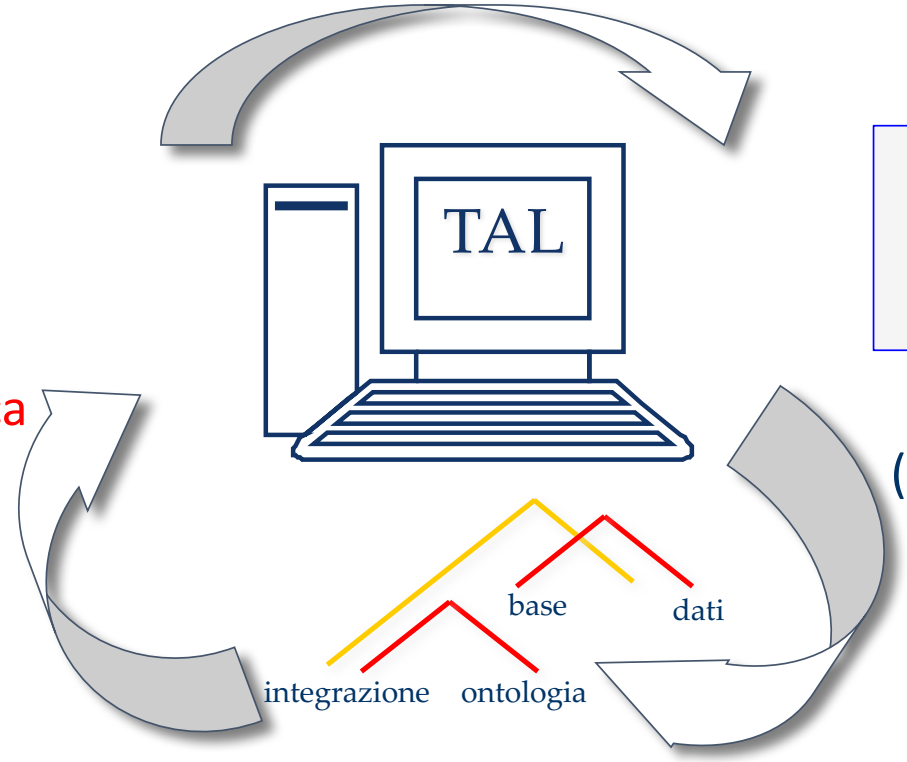
LANGUAGE

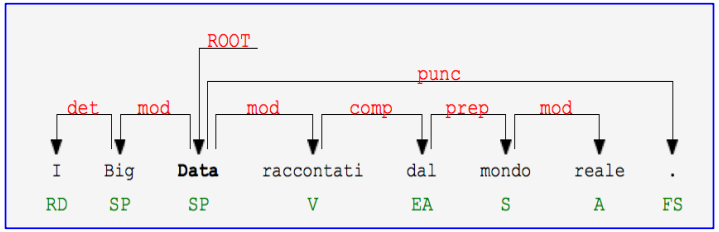
MEANING

Il Trattamento Automatico della Lingua (TAL)

analisi automatica della struttura linguistica

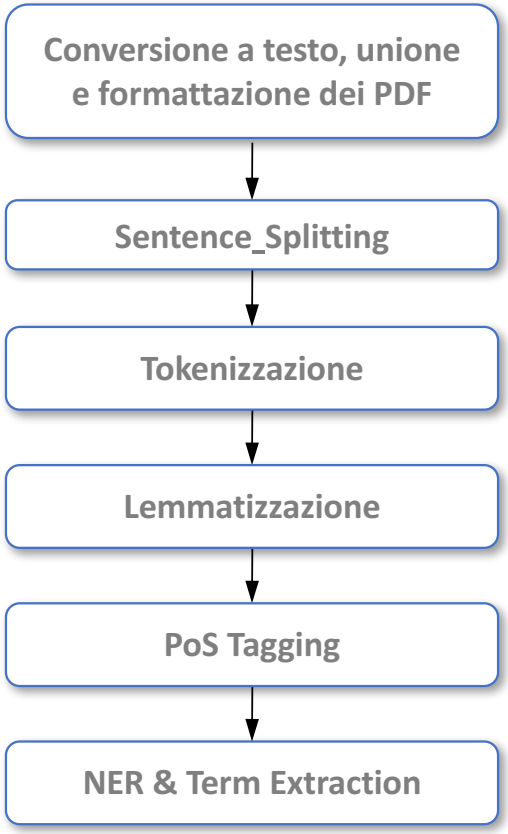
indicizzazione semantica
(smart searches)





analisi semantica
(chi, cosa, dove, sentiment, ecc.)

Annotazione linguistica



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

Frase 1:Questo è un esempio di analisi.

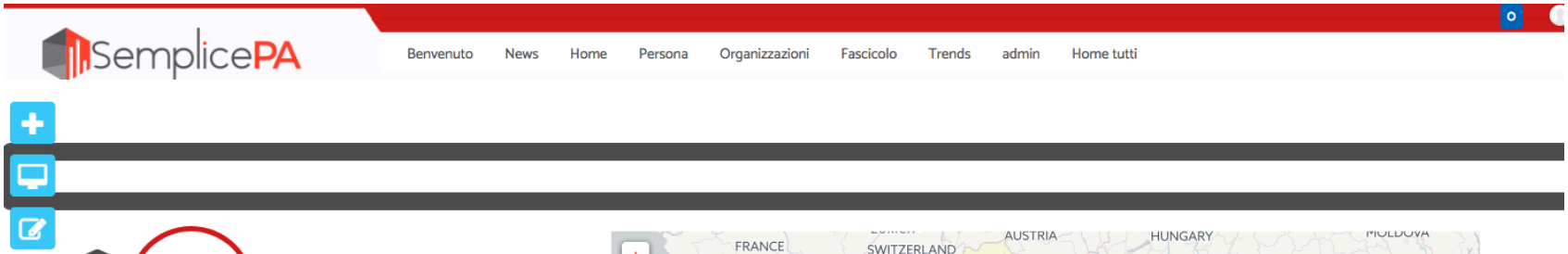
Frase 2:Il testo poi continua con altre frasi…

id	form
1	Questo
2	è
3	un

id	form	lemma
1	Questo	questo
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	_

TAL, Big Text Data e PA



994

COMUNI

928742

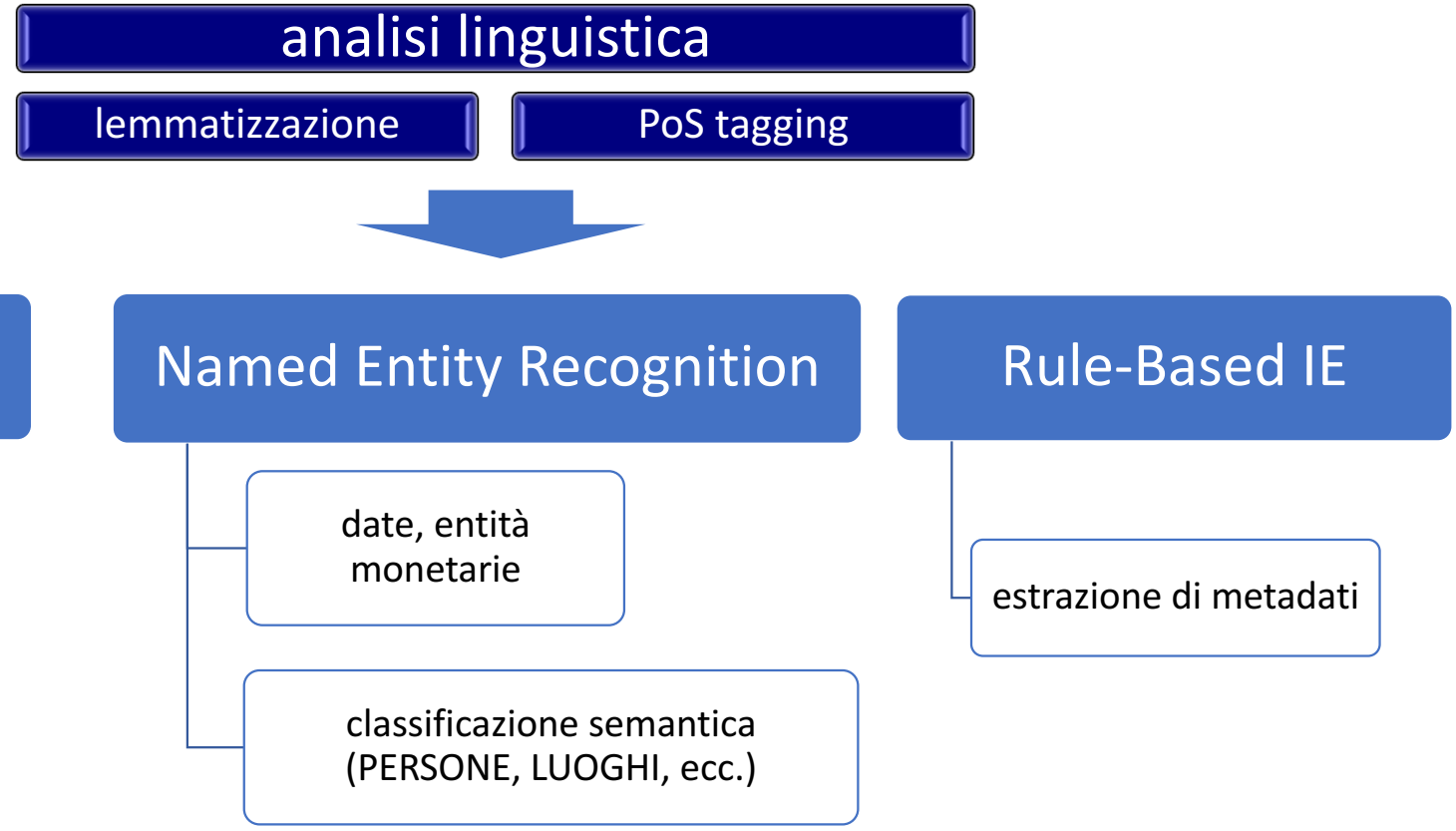
DOCUMENTI

LEGGI
Le piu' citate

COMUNI
Con piu' documenti

ARGOMENTI
I piu' usati

Analisi semantica con sistemi di TAL



Named Entity Recognition (dominio PA)

CoLing Lab NER (Passaro et al. 2017)

- Organizzazioni
  - *Università di Pisa*; *Acque S.p.A.*
- Luoghi
  - *Via S. Maria n. 36, PISA*
- Persone
  - *Mario Rossi*
- Leggi
  - *art. 134 comma 4 del D. Lgs. 267/2000*

- Atti [Tipo, Numero, data]
  - deliberazione di Giunta n. 111 di 1° febbraio 2000
    - [**Delibera di giunta**, **111**, **01/02/2000**]
  - Provv. Dir.2014/DD/00253
    - [**Determina**, **253**, **2014**]
  - Ordinanza n° 277 di 09.08.13
    - [**Ordinanza**, **277**, **09/08/2013**]

- Espressioni monetarie
  - 4.576 € → 4.576 €; 30.000 euro → 30.000 €; EUR640.43 → 640.43€
- Date
  - 1/9/90 → 01/09/1990
  - primo settembre del 1990 → 01/09/1990

Big text data e sentiment analysis

- I dati testuali sono una fonte di informazione sempre più cruciale per analizzare il “sentiment”

<doc user="corrieredellasera">

Ogni giorno esce fuori una merdata nuova e la CEI continua a rompere le palle sulle unioni civili e sui separati! Iniziassero a guardare il marcio in casa loro prima di fare la morale al mondo!

<doc user="corrieredellasera>

Forse i ns. giudici si stanno preparando all arrivo dell isis. Tanto per quei selvaggi le donne sono meno di niente. Gli sono solo utili quando vogliono fare i maiali schifosi (chiedo scusa ai suini) per il resto zero. Così saranno già dalla parte della legge. Che vomito!!!!!!!

<doc user="corrieredellasera">

L'Italia e tutta bella e meravigliosa. Questa terra e la sua gente mi piacciono molto e da sempre!!!E non si dovrebbe démenticare la cucina italiana!! Per me e la cucina migliore nel mondo!!😃🐝🍕

<doc user="IlFattoQuotidiano">

Non si vive in eterno. Quel poco di tempo che passiamo qui cerchiamo di viverlo godendoci del buon cibo e bevendo ottimo vino. Poi si morirà. Come tutti d'altronde

Emozioni: i colori delle parole

vacanza

stipendio

emozioni

parole

tasse

funerale

amore

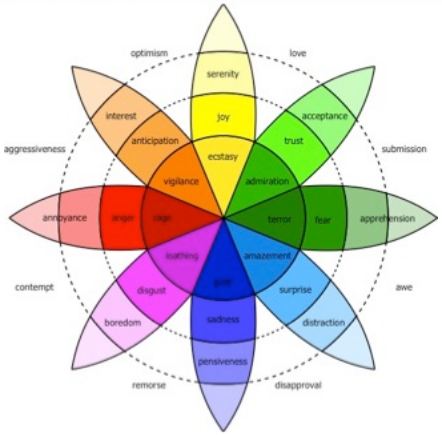
deprimente

putrido

oscuro

violenza

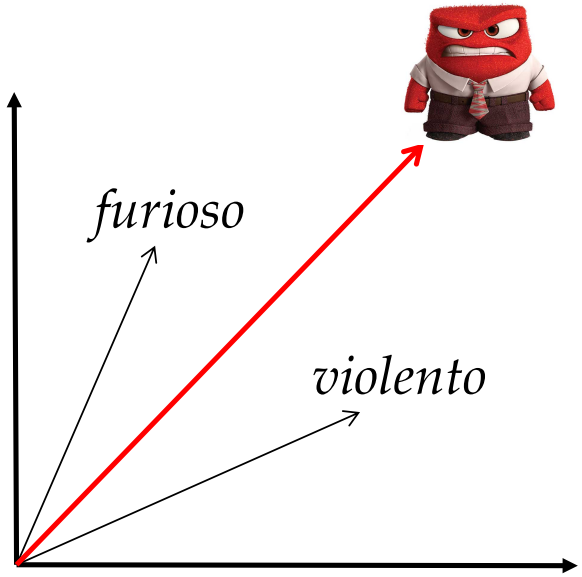
Parole, emozioni e vettori



- Ipotesi distribuzionale emotiva
  - Una parola è associata a un'emozione se ricorre negli stessi contesti linguistici in cui ricorrono altre parole associate con quell'emozione

- Otto emozioni di base
  - GIOIA, TRISTEZZA, RABBIA, PAURA, DISGUSTO, FIDUCIA, SORPRESA, ATTESE
- Per ogni emozione, 60 soggetti hanno prodotto parole "seed" fortemente associate a quell'emozione

Parole, emozioni e vettori



- Ogni parola è rappresentata con un vettore che codifica la sua distribuzione statistica nei contesti linguistici
- Ogni emozione è rappresentata con un vettore medio dei vettori delle sue parole "seed"

$$= \frac{\overrightarrow{violento} + \overrightarrow{furioso} + \cdots + \overrightarrow{s_n}}{n}$$

- Ad ogni nuova parola, viene assegnato un punteggio emotivo, misurando la vicinanza del suo vettore con il vettore dell'emozione

Emozioni e spazi semantici distribuzionali

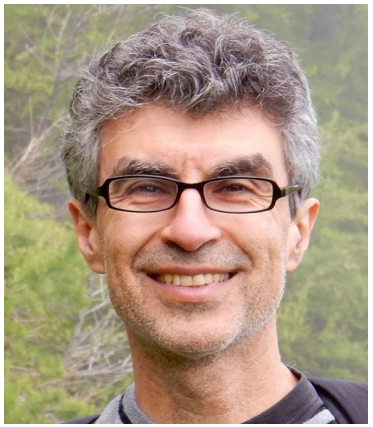
- Vettori addestrati su FB-NEWS15, un corpus scaricato automaticamente da Facebook (Passaro et al. 2016)
  - post, commenti e risposte pubblicate sulle pagine Facebook dei maggiori quotidiani online dal 1 gennaio al 31 dicembre 2015

Quotidiani on line	Post	Commenti	Risposte	Token
rainews.it	32.982	267.562	69.290	7.685.122
avvenire.it	6.726	62.684	22.414	2.593.732
ilGiornale	44.315	2.971.825	481.470	63.799.437
liberonews	29.400	2.151.188	255.658	40.841.255
HuffPostItalia	18.246	1.090.003	443.793	32.340.049
corrieredellasera	37.015	2.709.861	807.090	63.704.038
repubblica	24.829	3.248.123	1.285.877	95.326.364
IlFattoQuotidiano	25.217	3.628.701	1.246.396	98.279.833
TOTALE	**218.730**	**16.129.947**	**4.611.988**	**404.569.830**

Deep Learning? La *vera* novità sono i Big Data

- “We don’t have better algorithms. We just have more data.”

  Google’s Research Director Peter Norvig



- “The learning algorithms reaching human performance on complex tasks today are nearly identical to the learning algorithms that struggled to solve toy problems in the 1980s, though the models we train with these algorithms have undergone changes that simplify the training of very deep architectures […] The age of “Big Data” has made machine learning much easier”

  Goodfellow, Bengio, Courville (2016), Deep Learning, MIT Press

Big Data, AI e TAL: verso nuove sinergie

learning

analytics

Conclusioni e prospettive

- I Big Text Data sono una “miniera di informazioni” ancora largamente inesplorata
- Le tecnologie per il trattamento automatico della lingua sono necessarie per arrivare a una “comprensione profonda” dei testi

MUSE

(MUltimodal Semantic Extraction)

Estrarre semantica dai testi e dalle immagini

CoLing Lab
Computational Linguistics Laboratory

bNOVA

The COmputational LINGuistics LABoratory

colinglab.fileli.unipi.it/

CoLing Lab
Computational Linguistics Laboratory

Aree di ricerca:

- Natural Language Processing (Tools and Resources)
- Domain-specific Information Extraction, Term Extraction
- Opinion Mining, Sentiment Analysis, Affective Computing
- Computational Lexicons and Annotated Corpora

I nostri partner

bNOVA

ETI
evolution, technology & innovation

GruppoMeta