EVENTO 2018
BIGDATATECH

DATA FOR HUMAN

LUCIA PASSARO

CoLing Lab
Computational Linguistics Laboratory

SEM Il Chattadino

SEM il Chattadino: Outline

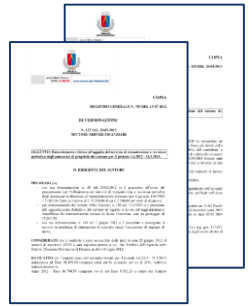
- I Big (text) data: cosa sono e cosa offrono
- Dal testo grezzo all'indicizzazione semantica
- Il TAL al servizio delle persone
- SEM: interrogare i dati via chat
- Tecnologie coinvolte
  - Focus su tecniche di Linguistica Computazionale
- SEM in azione

Perché i Big (Text) Data?

- La rivoluzione dell'informazione ci ha esposti a una mole di dati immensa che spesso non siamo in grado di processare
- Molti open data sono in formato non strutturato (solo testuali)
- Necessità di sintetizzare: dal testo alle informazioni
- La circolazione delle informazioni alimenta la trasparenza della PA

Big (Text) Data

- Dati non strutturati ad alto tasso di variabilità
- Contenuti informativi impliciti
- L'estrazione delle informazioni richiede la comprensione linguistica del testo
- Ricchi di entità (persone, luoghi, organizzazioni…), eventi e relazioni intra ed extra-testuali
- Fonti molto eterogenee



I big (text) data nella PA

La nuova **legislazione sulla trasparenza** ha obbligato le PA a pubblicare i propri documenti in forma elettronica

- Formato non omogeneo
- Nessun metadato obbligatorio
- Non adatto a **ricerca** e **analisi**

Le nuove tecniche di **TAL** e **Information Extraction** permettono di:

- Identificare trend nell'attività della PA stessa
- Individuare irregolarità
- Semplificare l'accesso alle informazioni

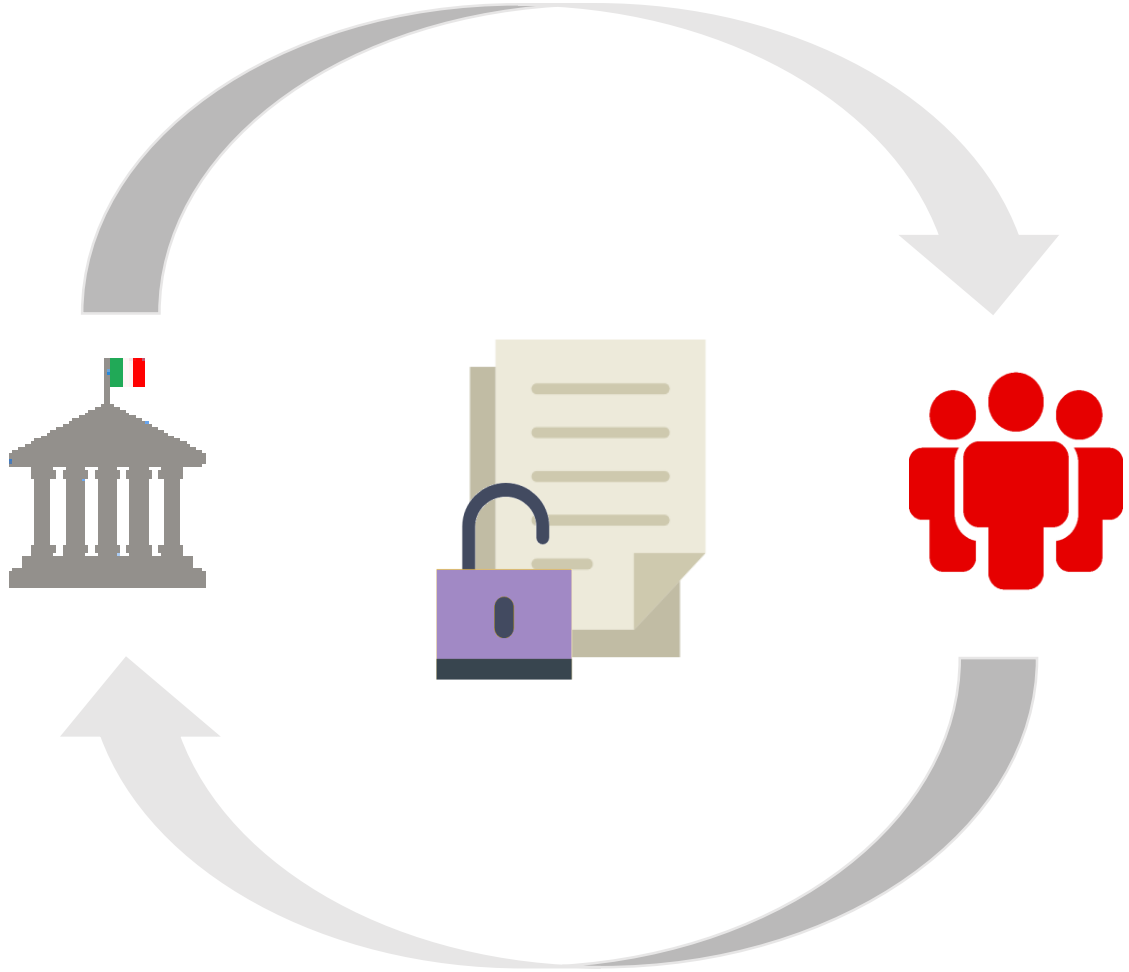
Cui prodest?

Amminsitrazione

- Migliorare l'efficienza
- Coinvolgere i cittadini
- Rispettare i vincoli sulla trasparenza
- Snellire l'attività di front office

Cittadini

- Accedere ai servizi in modo semplice
- Partecipare attivamente alla comunità
- Controllo dell'attività amministrativa



Dai dati strutturati alle informazioni

Dai dati strutturati alle informazioni

Quando è aperto l'Ufficio Anagrafe?

date
organizzazioni
comune

Dai dati strutturati alle informazioni

Quali sono aziende con cui il comune ha lavorato di più nel 2017?

aziende e organizzazioni
fatture e importi
comune

Dai dati strutturati alle informazioni

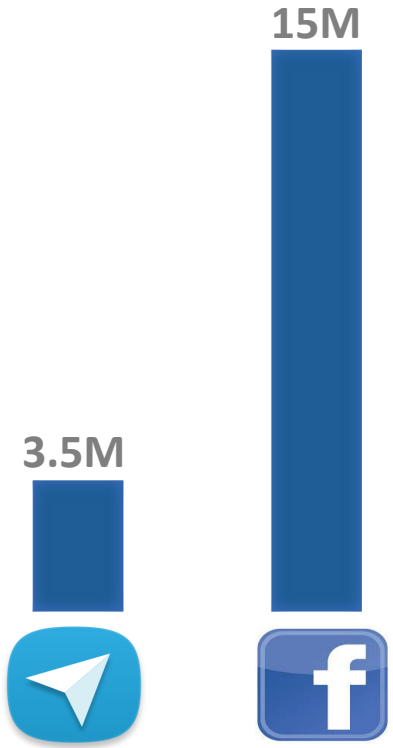
Dove si trova la piscina comunale?
Quando è aperta?

organizzazioni
date
luoghi
comune

Chatbot {chat + robot}

- Programma in grado di **simulare una conversazione umana**
- Tecnologia legata all'**Industria 4.0**
- In Italia le chat sono usate dal **78% di utenti** [Media UE 60%]

- Telegram è stata la prima app di messaggistica a lanciare i ChatBot (2015) seguita da Skype, iMessage, WeChat e Facebook Messenger



ETI evolution, technology & innovation

bNOVA

UNIVERSITÀ DI PISA
FILOLOGIA, LETTERATURA E LINGUISTICA

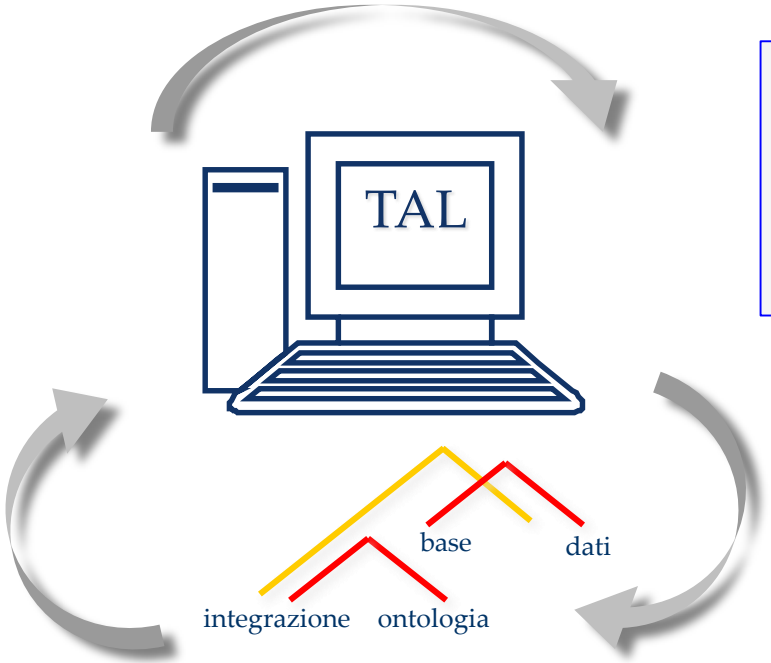
[r][e]

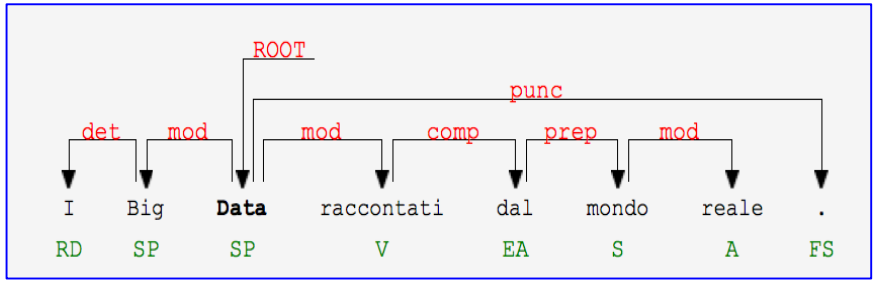
- Framework per la **creazione di ChatBot** in grado di **dialogare** via chat o a voce
- Competenze derivanti da un vasto **repository di dati strutturati e non**, riguardanti le **pubbliche amministrazioni**
- L'attività di SEM sarà monitorata attraverso una **Dashboard di Data Analytics**
- SEM sarà **raggiungibile**, dalle più diffuse app di messaggistica e dalla **propria app**
- Base dati di partenza: **SemplicePA**

Trattamento Automatico del Linguaggio (TAL)

Analisi automatica della struttura linguistica

Indicizzazione semantica
(aggiunta di metadati strutturati ai testi)





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

Architettura e partner del progetto SEM

Base di conoscenza dinamica

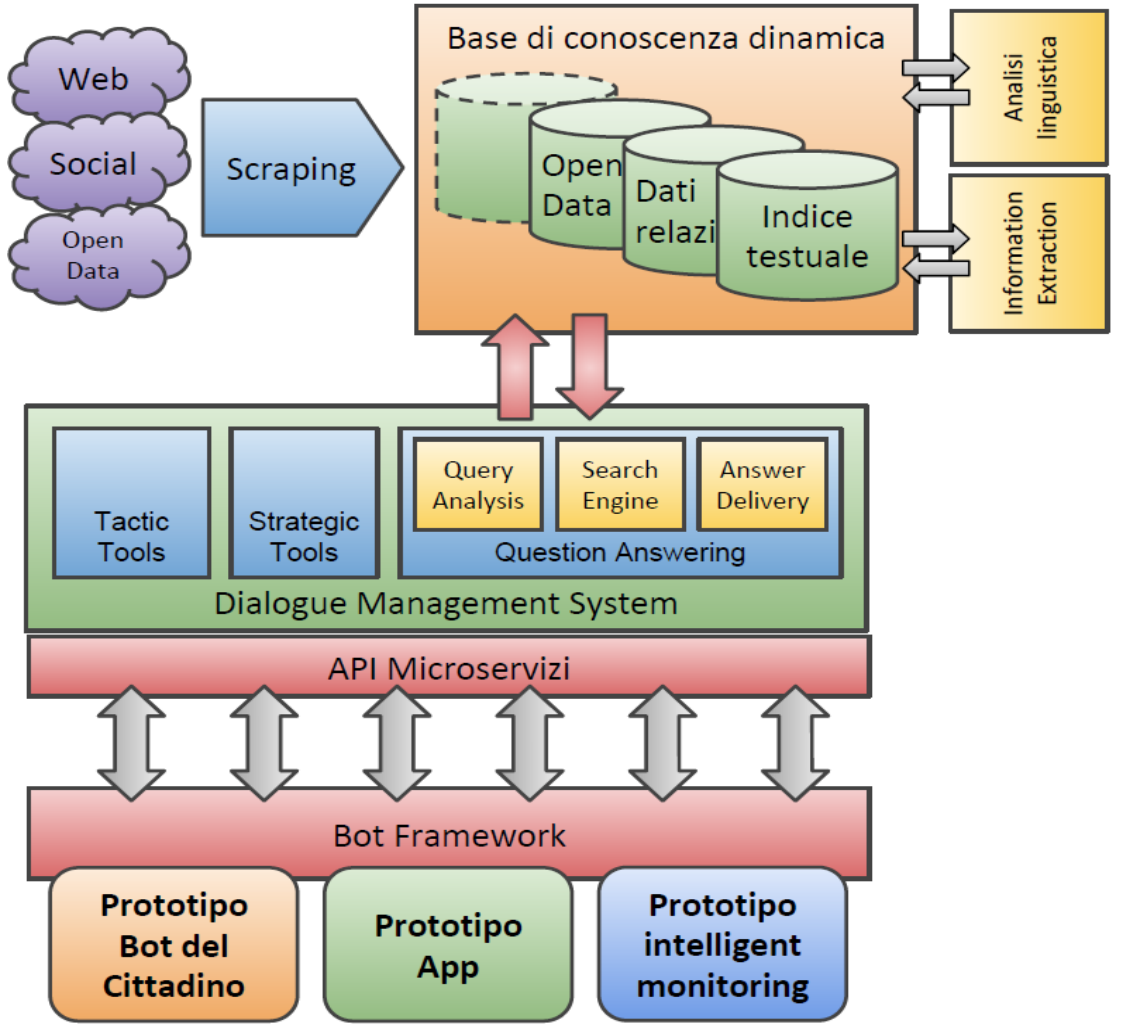
UNIPI e Bnova

Question Answering & Dialogo

ETI3 e Unipi

App e monitoraggio

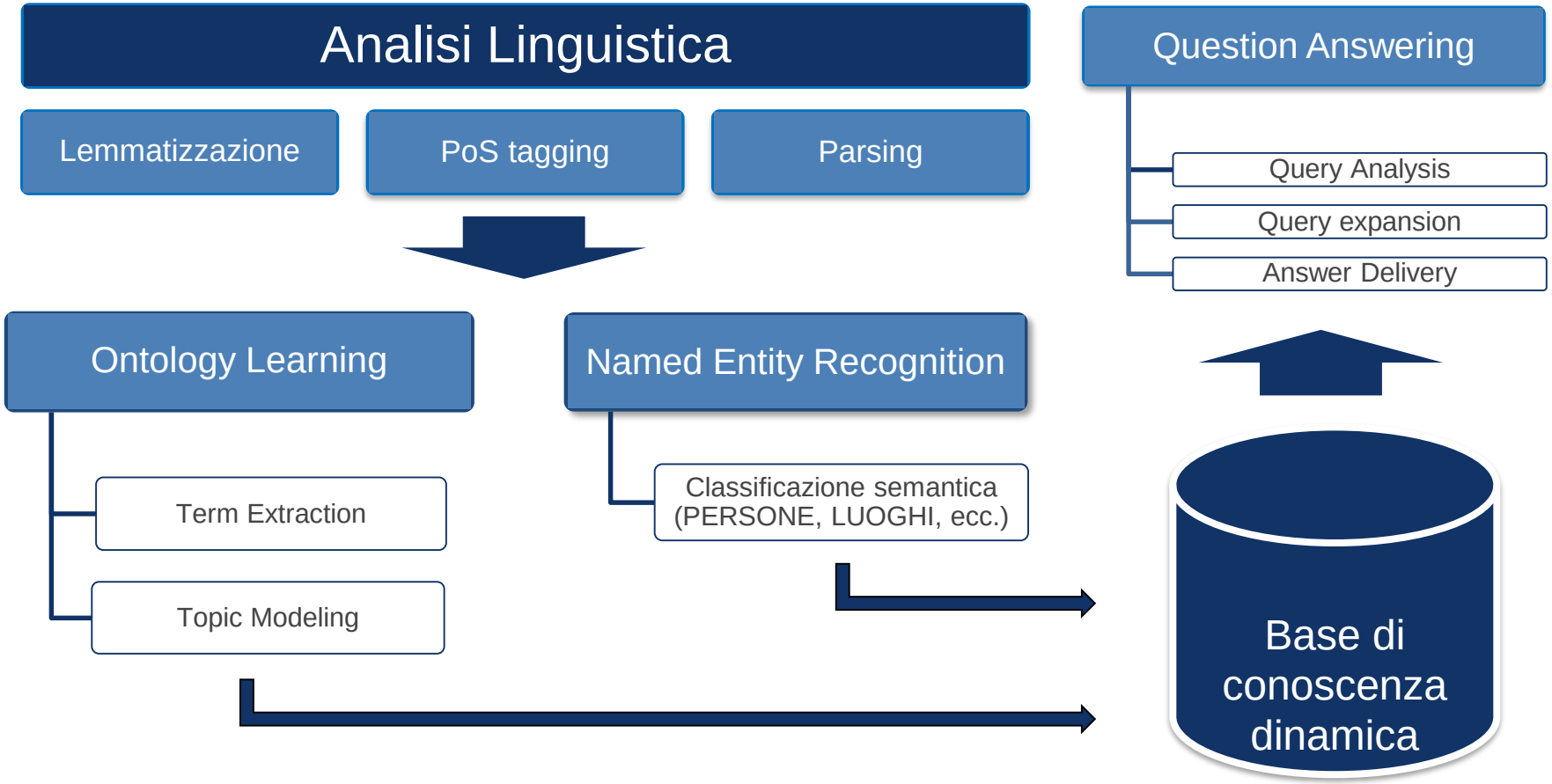
ETI3, Rigel, Bnova e UNIPI



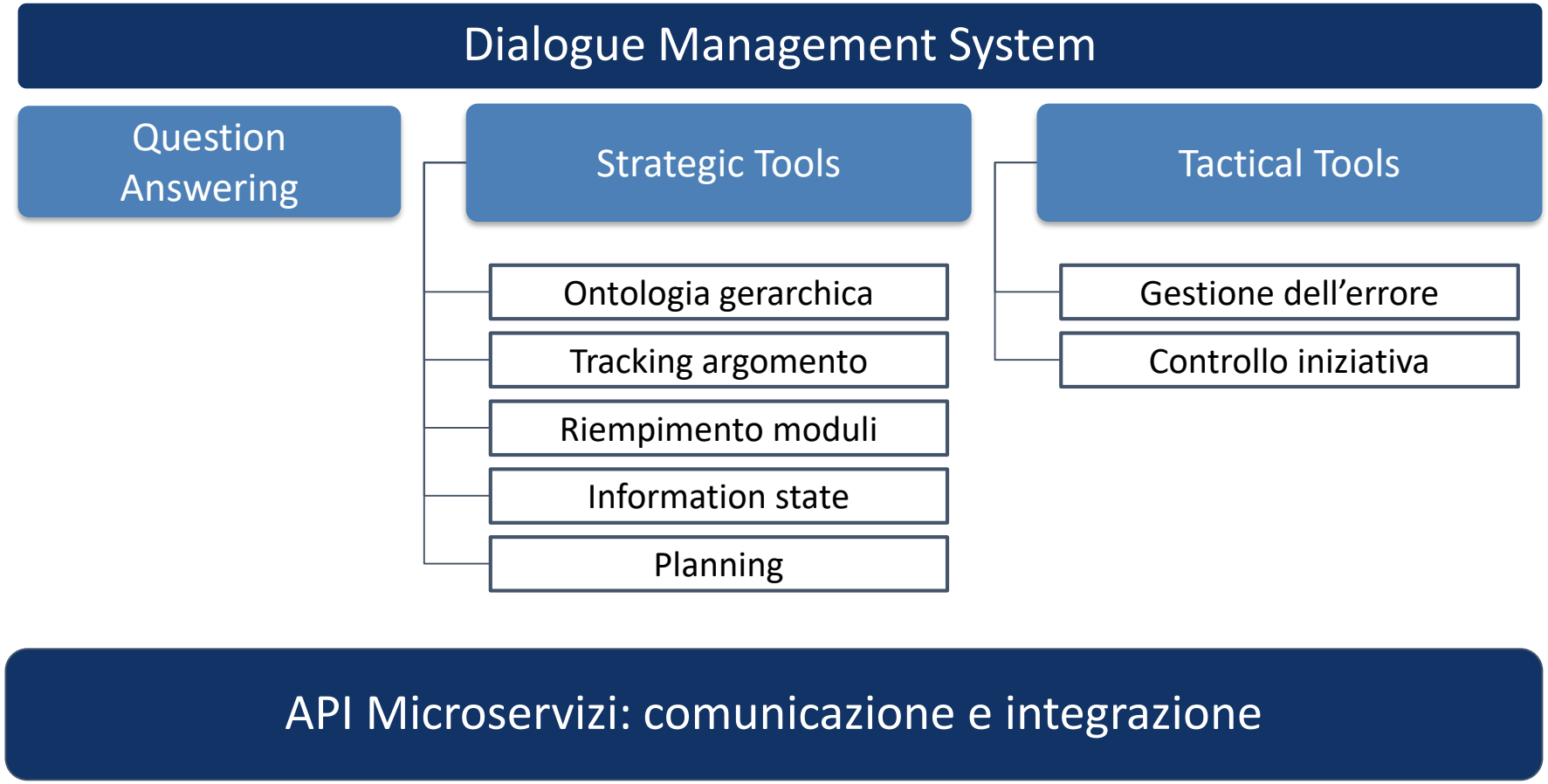
Principali Tecnologie coinvolte

- **Analisi Semantica**
- **Question Answering e Dialogue Management**
- **Data analytics**

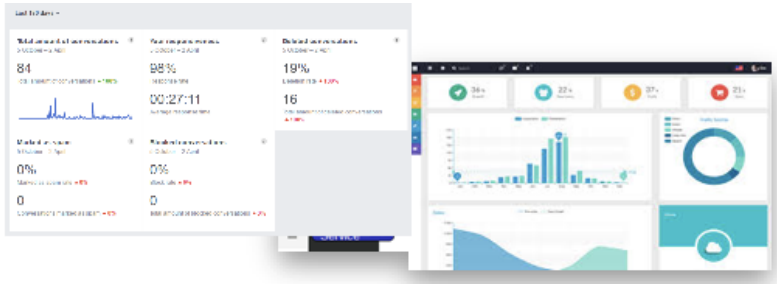
Analisi Semantica



Dialogue Management



Data Analytics



pentaho

Scraping

Web Data Source | Social Network | Open Data

pentaho Data Integration

Staging Area

pentaho Data Integration

DWH

Intelligent Monitoring

- Churn Analysis
- Sentiment Analysis
- Data Mining

Dialogue Management System

Analisi linguistica

- Sentence Splitter
- Tokenizer
- Morphological analysis
- PoS-tagger
- Dependency Parser
- Term extraction
- Named Entity Recognition

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
3	un	un

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	

VISTO l'atto di concessione prot. n. 137145/L.11.1 rilasciato in data 20/11/2014 dall'Ufficio Manutenzione Strade e Infrastrutture del Comune di Arezzo a "Telecom Italia S.p.A." al fine di far effettuare lavori di scavo per l'allacciamento elettrico di collegamento con gli armadi Telecom in Loc. Olmo, Loc. Madonna di Mezzastrada, Loc. Il Matto e Loc. Ripa dell'Olmo, alla ditta "T.T.E. S.p.A".; RITENUTO di consentire lo svolgimento dei lavori ed allo stesso tempo

Information Extraction

Termini semplici rilevanti

- Imposta, scadenza, ufficio

Termini complessi

- [nome+prep+aggettivo] ordine del giorno, bando di gara
- [nome+aggettivo] casa farmaceutica, verde pubblico

Entità nominate

- [PER] Mario Rossi
- [ACT] Delibera di giunta n. 23 del 25/10/2013
- [ORG_PA] Ufficio Anagrafe, Servizio Finanziario

Information Extraction

Termini semplici rilevanti

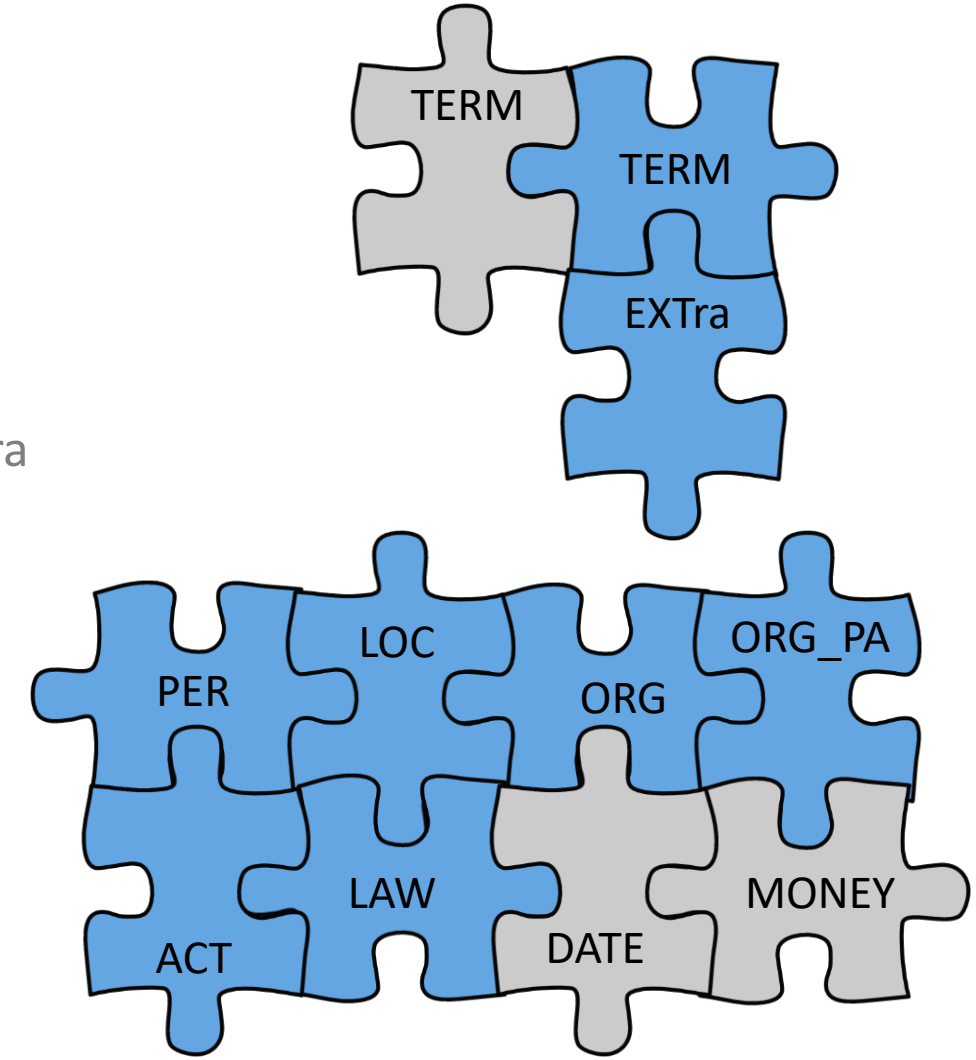
- Imposta, scadenza, ufficio

Termini complessi

- [nome+prep+aggettivo] ordine del giorno, bando di gara
- [nome+aggettivo] casa farmaceutica, verde pubblico

Entità nominate

- [PER] Mario Rossi
- [ACT] Delibera di giunta n. 23 del 25/10/2013
- [ORG_PA] Ufficio Anagrafe, Servizio Finanziario



Question Answering

Analisi della domanda

- Processing della domanda per estrarre le informazioni rilevanti
  - termini, entità, topic
- Classificazione della focus richiesta
  - dove, quando, come ecc.

Query expansion

- Espansione dei termini rilevanti mediante tecniche di semantica distribuzionale

Answer delivery

- Ranking dei risultati e preparazione della risposta
  - link, snippet, breve testo di risposta

Text Object detection (indicizzazione)

Consente di collegare dati strutturati (entità e termini) per formare degli «oggetti» sui quali SEM potrà fornire delle risposte puntuali

- Eventi
- Servizi
- Tributi
- Segnalazioni/interventi
- Enti/Uffici

Text Object detection (indicizzazione)

*"Terre di Pisa Food & Wine Festival" settima edizione - un viaggio di gusto alla riscoperta di prodotti tipici e antiche ricette delle "**Terre di Pisa**" - ti aspetta **dal 19 al 21 ottobre 2018** alla **Stazione Leopolda di Pisa**, **ingresso libero**.*

…	**Oggetto**	**Attributo**	**Valore**
…	Evento	Titolo	Terre di Pisa
…		Data	dal 19 al 21 ottobre 2018
…		Luogo	Stazione Leopolda di Pisa
…		Costo	ingresso libero

Text Object detection (indicizzazione)

*"Terre di Pisa Food & Wine Festival" settima edizione - un viaggio di gusto alla riscoperta di prodotti tipici e antiche ricette delle "**Terre di Pisa**" - ti aspetta **dal 19 al 21 ottobre 2018** alla **Stazione Leopolda di Pisa**, **ingresso libero**.*

…	**Oggetto**	**Attributo**	**Valore**
…	Evento	Titolo	**Terre di Pisa**
…		Data	**dal 19 al 21 ottobre 2018**
…		Luogo	**Stazione Leopolda di Pisa**
…		Costo	**ingresso libero**

Text Object detection (Domanda)

Quando *si terrà* ***Terre di Pisa quest'anno****?*

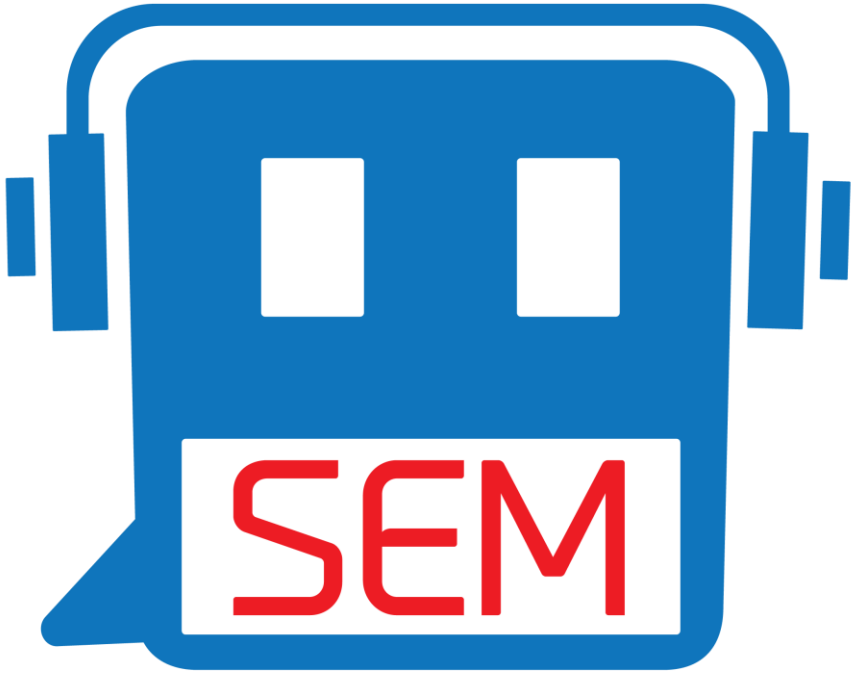
…	**Oggetto**	**Attributo**	**Valore**
…	Evento	Titolo	Terre di Pisa
…		Data	dal 19 al 21 ottobre 2018
…		Luogo	Stazione Leopolda di Pisa
…		Costo	ingresso libero

Text Object detection (Domanda)

Quando *si terrà* ***Terre di Pisa quest'anno*****?**

…	Oggetto	Attributo	Valore
…	Evento	Titolo	**Terre di Pisa**
…		**Data**	**dal 19 al 21 ottobre 2018**
…		Luogo	Stazione Leopolda di Pisa
…		Costo	ingresso libero

SEM in azione



Conclusioni e prospettive

- I Big Text Data sono una «miniera di informazioni» largamente inesplorata
- Le tecnologie per il TAL oggi possono approssimare una comprensione profonda dei testi
- SEM il Chattadino rende accessibili i dati della PA in modo semplice e diretto
  - Sfruttando una base di conoscenza indicizzata con metadati semantici
  - Verificando la soddisfazione degli utenti e l'attività del chatbot attraverso un sistema di intelligent monitoring

http://colinglab.fileli.unipi.it

CoLing Lab

Computational Linguistics Laboratory

UNIVERSITÀ DI PISA
FILOLOGIA, LETTERATURA E LINGUISTICA

ALESSANDRO LENCI 

MARTINA MILIANI 

ALESSANDRO BONDIELLI 

GRAZIE PER L'ATTENZIONE