Semantic Matching: KG Integration and Construction

Part VI: Matching & Information Extraction

Information Extraction Tasks

Named Entity Recognition & Linking

(Clear Forest Gnosis Mozilla Plugin)

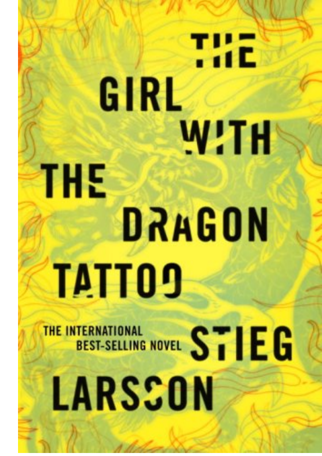
Named Entity Recognition (NER)

- Estrazione di entità significative e classificazione per tipi (persona, azienda, luogo, etc)

Named Entity Linking (NEL)

- Linking di entità estratte a entità descritte in una KB (e.g., Coldplay is linked to dbr:Coldplay

Relation Extraction: extraction of facts
(e.g., *machine reading*)



It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at hi[…] on the ti[…] of Hedeby. The old man […] Blomkvist in by promising solid evidence against Wennerström. Blomkvist agrees to spe[…] writing the Vanger family history as a cover for the real assignment: the disappearance of Vanger's niece Harr[…] 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist becomes acquainted with […]rs of the extended V[…]mily, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Henrik.

Af[…]vering that Salander has hacked into his comp[…] ades her […] him with research. They eventu[…]vers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.

A 24-year-old computer hacker sporting an […] of tattoos and body piercings supports her[…] doing deep background investiga[…] Dragan Armansky, who, in turn, worries that Lis[…]lander is "the perfect victim for anyone who wished her ill."

same · same · same · owns · uncleOf · hires · same · affairWith · affairWith · enemyOf · same · headOf · same

O. Etzioni, M. Banko, M.J. Cafarella: Machine Reading, AAAI ,06
T. Mitchell et al.: Populating the Semantic Web by Macro-Reading Internet Text, ISWC'09

NER, NEL & Relation Extraction

- Different techniques
  - NLP with deep parsing (e.g., syntactic analysis of sentences)
  - Machine learning (e.g., statistical analysis of frequent patterns)
  - Vocabulary-based (e.g., match against a KB)
- Open Information Extraction vs. Closed Information Extraction
  - extraction of unknown entities and relations (e.g., without a dictionary) vs. extraction of known entity mentions and facts based on a fixed number of relations
- Often combined, also depending on the features of the input text
  - Well-formed natural laguage (e.g., news articles) vs. ill-formed or colloquial natural language (e.g., tweets, product descriptions)

Examples in this presentation

- Open NER + NEL from Tweets (difficult research research)
- Relation extraction in a partially closed domain (eCommerce, industry-ready approach)

Matching Information Extracted from Social Media
Entity Linking for Twitter

Bridging the Gap between Short Texts and Knowledge Graphs

Slides credits:
Pikakshi Manchanda (PhD candidate in CS)
Fausto Ristagno (CS Master Student)

Social media: Entities-Emojis-Events Express

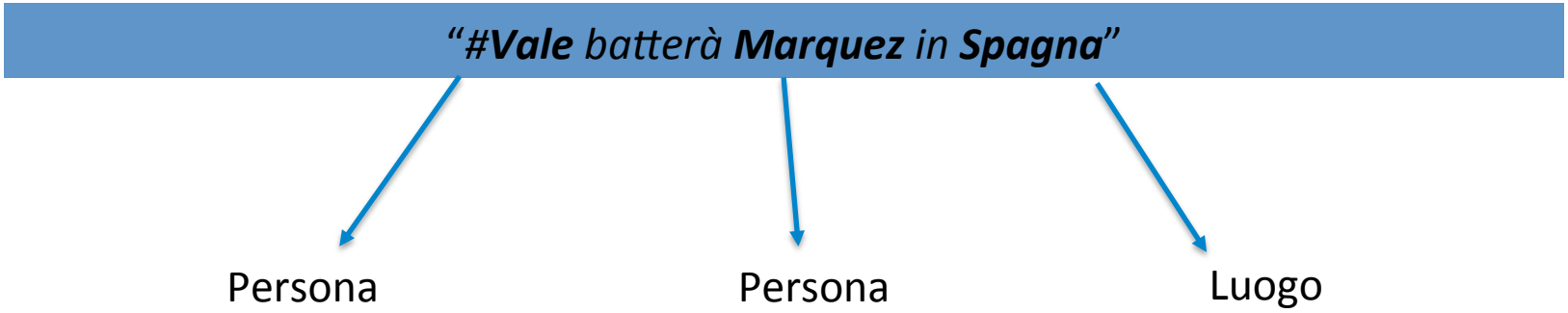
- People communicate and share important news/ announcements through social media platforms -- Constant production & dissemination of *'user generated content'*

  - ✓ Product launches, marriage invites, expressing grief and sorrow
  - ✓ Live commentary/commenting over public events such as election debates, football/ soccer tournaments are few examples

- Fresh information emerging in real-time on social media platforms primarily

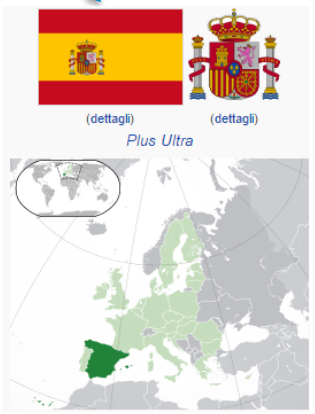
  - ✓ New (relevant/popular) entities (Book Launch)
  - ✓ New events (Emmy Awards, Orlando Shooting)
  - ✓ Factual information (Death of Muhammad Ali)
  - ✓ New relations (Clinton becomes President …. Speaking hypothetically at the moment 😉)

Named Entity Recognition



Entity Linking

"*#**Vale** batterà **Marquez** in **Spagna***"



wikipedia.org/wiki/Valentino_Rossi

wikipedia.org/wiki/Marc_Márquez

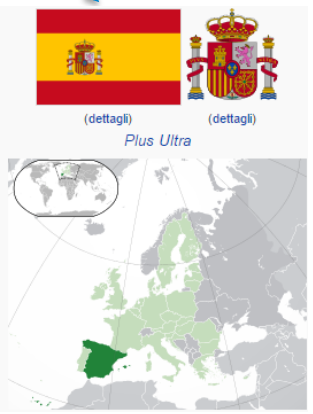
wikipedia.org/wiki/Spagna

Entity Linking

"*#**Vale** batterà **Marquez** in **Spagna***"

wikipedia.org/wiki/Rodolfo_Valentino

wikipedia.org/wiki/Álex_Márquez



wikipedia.org/wiki/Spagna

Making Sense of Micropost challenge

#Microposts2016

La sfida pone i seguenti obiettivi:

- Estrazione e **classificazione** delle Named Entity nei tweet (NER);
- **Linking** di ogni Named Entity all'entità di Dbpedia o NIL, se non ci sono corrispondenze (EL);
- Clustering delle Named Entity NIL che si riferiscono alla stessa entità reale.

ENTITY RECOGNITION

- Entity Recognition has been studied across various textual genres over the years: journalistic, scientific, informal
  - ✓ Ritter et al., 2011, Liu et al., 2011
- **Lu et al., 2015**, address identification of indirect entity references in microposts by modelling dependency among referred entities by a Conditional Random Field (CRF) model

Systems/ Tools	Approach	Domain	Entity Types/ Classes	Taxonomy
ANNIE	Gazetteers & FSM	Newswire	7	(adapted) MUC
Stanford NER	CRF	Newswire	4, 3 or 7	CoNLL, ACE
Alchemy API	Machine Learning	Unspecified	324	Alchemy
T-NER (Ritter et al., 2011)	CRF, LLDA	Twitter	3 or 10	CoNLL, ACE
NERD-ML	KNN & Naïve Bayes	Twitter	4	NERD
Liu et al. 2011	KNN & CRF	Twitter	4	CoNLL, ACE

Derczynski et al., 2015

ENTITY LINKING

Tools	Taxonomy	Approach/ Features used	Domain
DBpedia Spotlight (Mendes et al., 2011)	DBpedia, Freebase, Schema.org	Gazetteers and Similarity Metrics	Unspecified
TAGME (Ferragina and Scaiella, 2010)	Wikipedia	Wikipedia anchor texts and the pages linked to those anchor texts	Short texts
YODIE (Damljanovic and Bontcheva, 2012)	DBpedia	Similarity metrics and URI frequency	Twitter
Babelfy (Moro et al., 2014)	BabelNet semantic network	Graph-based approach, semantic signatures	Short text
Meij et al., 2012	Wikipedia	n-gram features, concept features, and tweet features	Twitter
Habib et al, 2012	YAGO	Unsupervised approach for improving entity extraction using disambiguation results	Twitter
Guo et al., 2013	Wikipedia	Structural SVM	Twitter
Yamada et al., 2015	Wikipedia	Supervised (String matching, n-grams)	Twitter
Basile et al, 2015	DBpedia	Unsupervised/supervised approach for NER (POS-tags, n-grams); Dist. Lesk Algorithm for NED & NEL	Twitter
Manchanda et al., 2015	DBpedia	Unsupervised approach for re-classification of entities using recognition and linking results	Twitter
Greenfield et al., 2016	DBpedia	Supervised approach (RandomForest) for entity linking using features such as Commonness, tf-idf, re-directs..	Twitter

Challenges

- Concise and idiosyncratic expressions
  - ✓@tamaraholder #WakeUpAmerica #FactsNotEmotion #DebateIt
- Abbreviated, misspelled or #hastagged entities

Recognition Challenges:

- Did I recognize it correctly?
  - ✓ @981THEBULL OMG. I want to win the **meet and greets** so[illegible]'ve been a **Swiftie** for 10 yrs.
- Out Of Vocabulary (OOV) entity mention identification problem
  - ✓ The Big Bang Theory being referred as *TBBT*
- Out of Knowledge base (OOKB) entity problem

Linking Challenges:

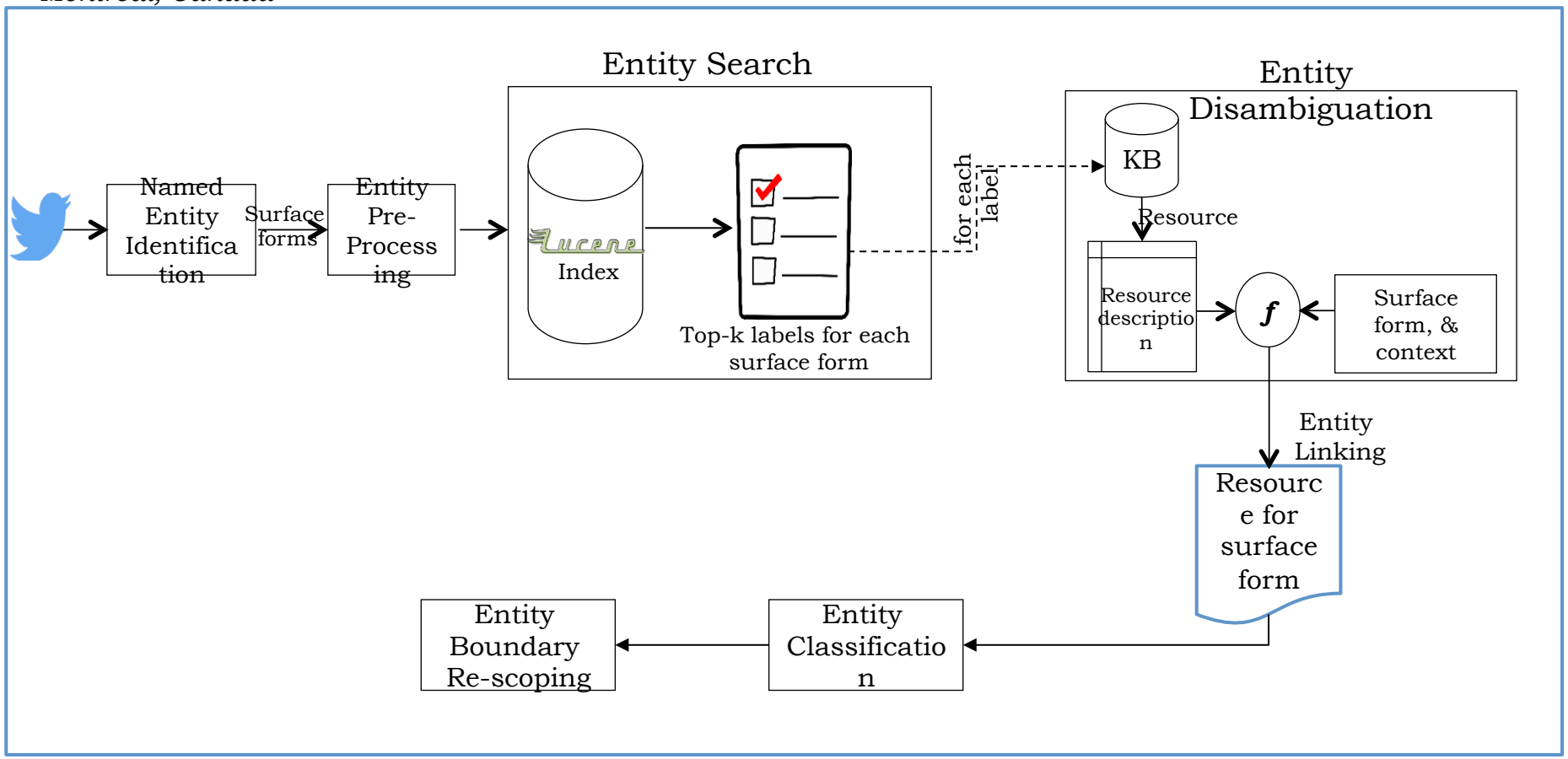
- Unlinkable entity -- Is it new? Is it mis-identified/wrongly identified?
- What comprises new knowledge? Can we apply some measures for validating the credibility of new knowledge?

Classification Challenges:

- Ambiguous named entities (e.g., Starbucks)
- Entity mention mis-classification
- Different NER systems have different classification ontologies. (e.g., Harry Potter – Person vs Character)
- New categories of named entities emerging in real-time on social media streams. ➔ Classifying them using an existing ontology? Updating the current ontology? Mapping it to a new ontology?

NEEL@Unimib - Pipeline

#Used in the NEEL Challenge, at Making Sense of Microposts (#Microposts2016), World Wide Web Conference, Montreal, Canada



- Davide Caliano, Elisabetta Fersini, Pikakshi Manchanda, Matteo Palmonari, Enza Messina. "UniMiB: Entity Linking in Tweets using Jaro-Winkler Distance, Popularity and Coherence". In 6th Workshop on Making Sense of Microposts (#Microposts2016).

SYSTEM DESCRIPTION

- Named Entity Linking: using an unsupervised, greedy approach to link an entity with a given DBpedia resource with the highest knowledge-base score, ***KB(e_j, c_k)***.

$$KB(e_j, c_k) = (\boxed{?} * lex(e_j, l_{ck}) + (1 - \boxed{?}) * (cos_k(e^*_j, a_{ck}))) + R(c_k)$$

($\boxed{?}$ = 0.7)

- $lex(e_j, l_{ck})$ is defined as follows:

$$lex(e_j, l_{c_k}) = lcs(e_j, l_{c_k}) + W_D \left(\frac{JW(e_j, l_{c_k})}{W_D + 1} \right)$$

W_D = 3.0 (boosting coefficient)

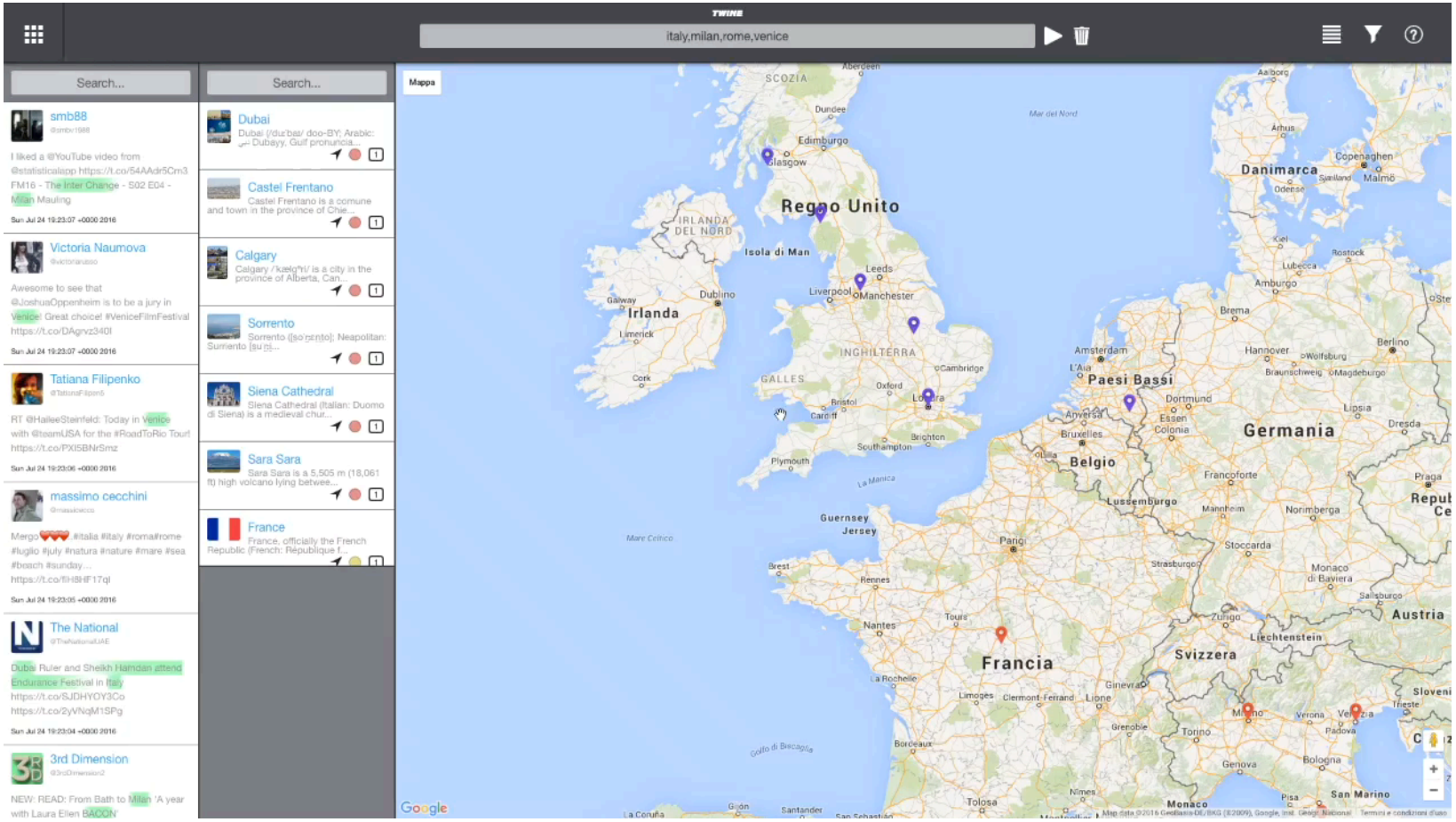
- Use of extended abstracts (rdfs:abstract) from DBpedia for the sake of calculating cosine similarity, defined as:

$$cos_k(e^*_j, a_{c_k}) = \begin{cases} cos(e^*_j, a_{c_k}) \text{ if } k = 1 \\ \dfrac{cos(e^*_j, a_{c_k})}{\log_2(k)} \; k \geqslant 2 \end{cases}$$

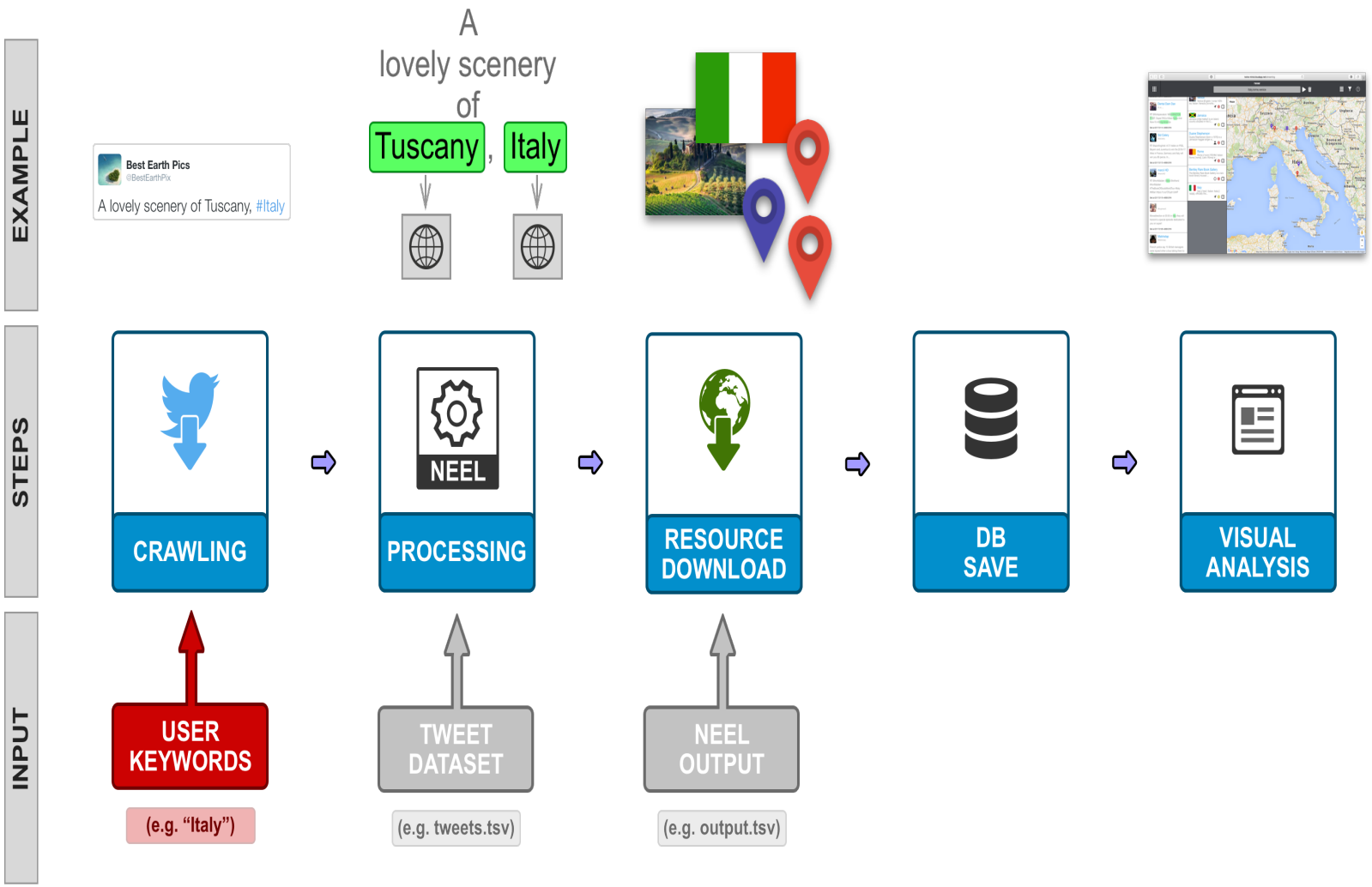
- $R(c_k)$ calculated by taking into account the popularity of a given candidate, and is computed by using the following boosted Page Rank measure:

$$R(c_k) = \beta * PR(c_k) \qquad (\beta = 0.6)$$

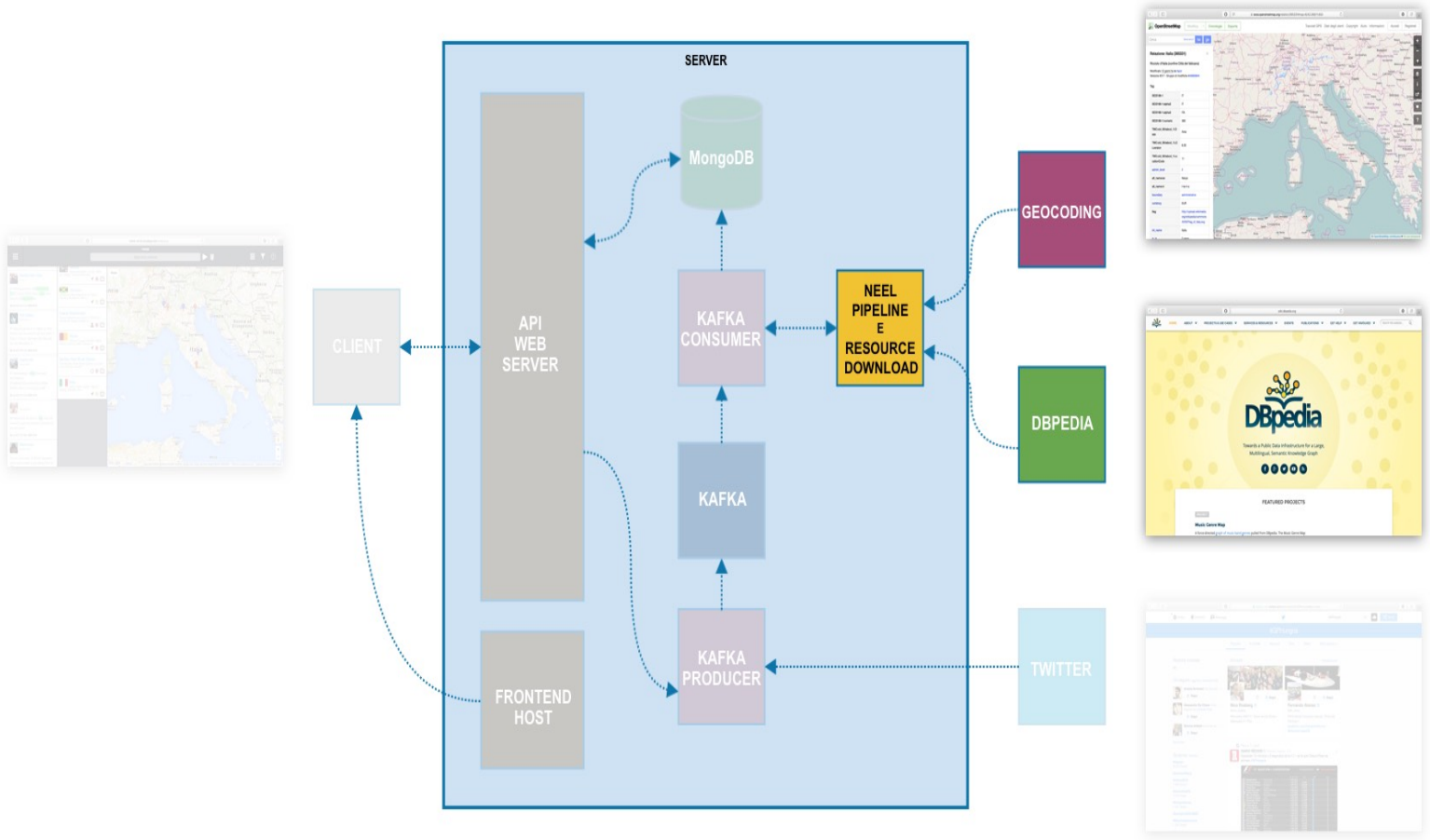
TWINE: SEMANTICS FOR SOCIAL MEDIA ANALYTICS



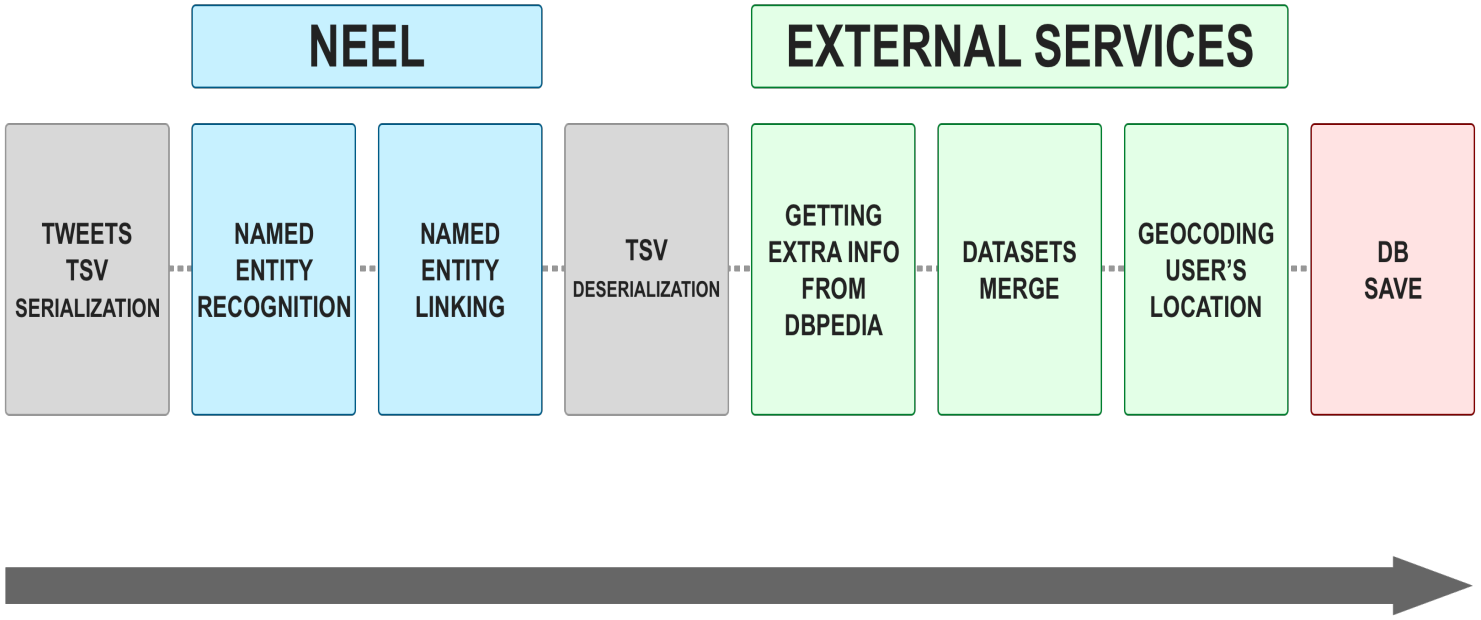
OVERVIEW DI TWINE



ARCHITETTURA



PIPELINE

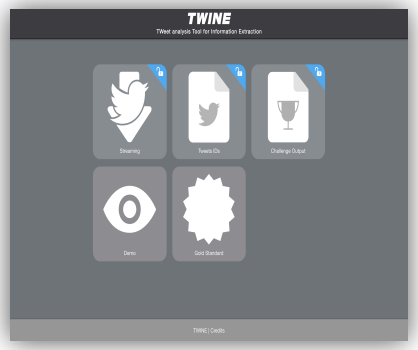

NEEL
EXTERNAL SERVICES
TWEETS TSV SERIALIZATION
NAMED ENTITY RECOGNITION
NAMED ENTITY LINKING
TSV DESERIALIZATION
GETTING EXTRA INFO FROM DBPEDIA
DATASETS MERGE
GEOCODING USER'S LOCATION
DB SAVE

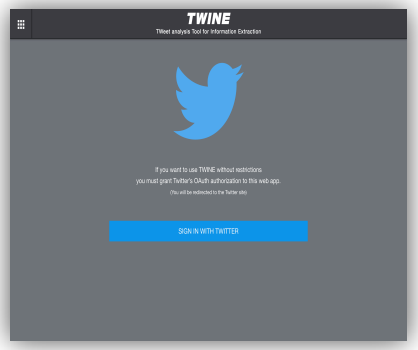
FRONTEND

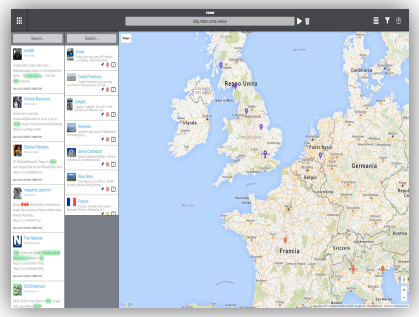
SPA

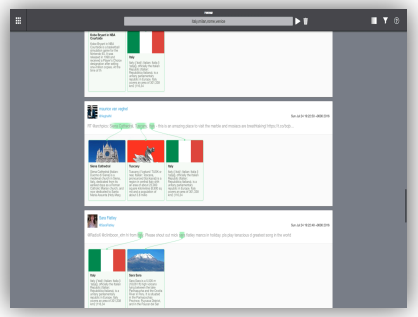
Vue.js

Flux

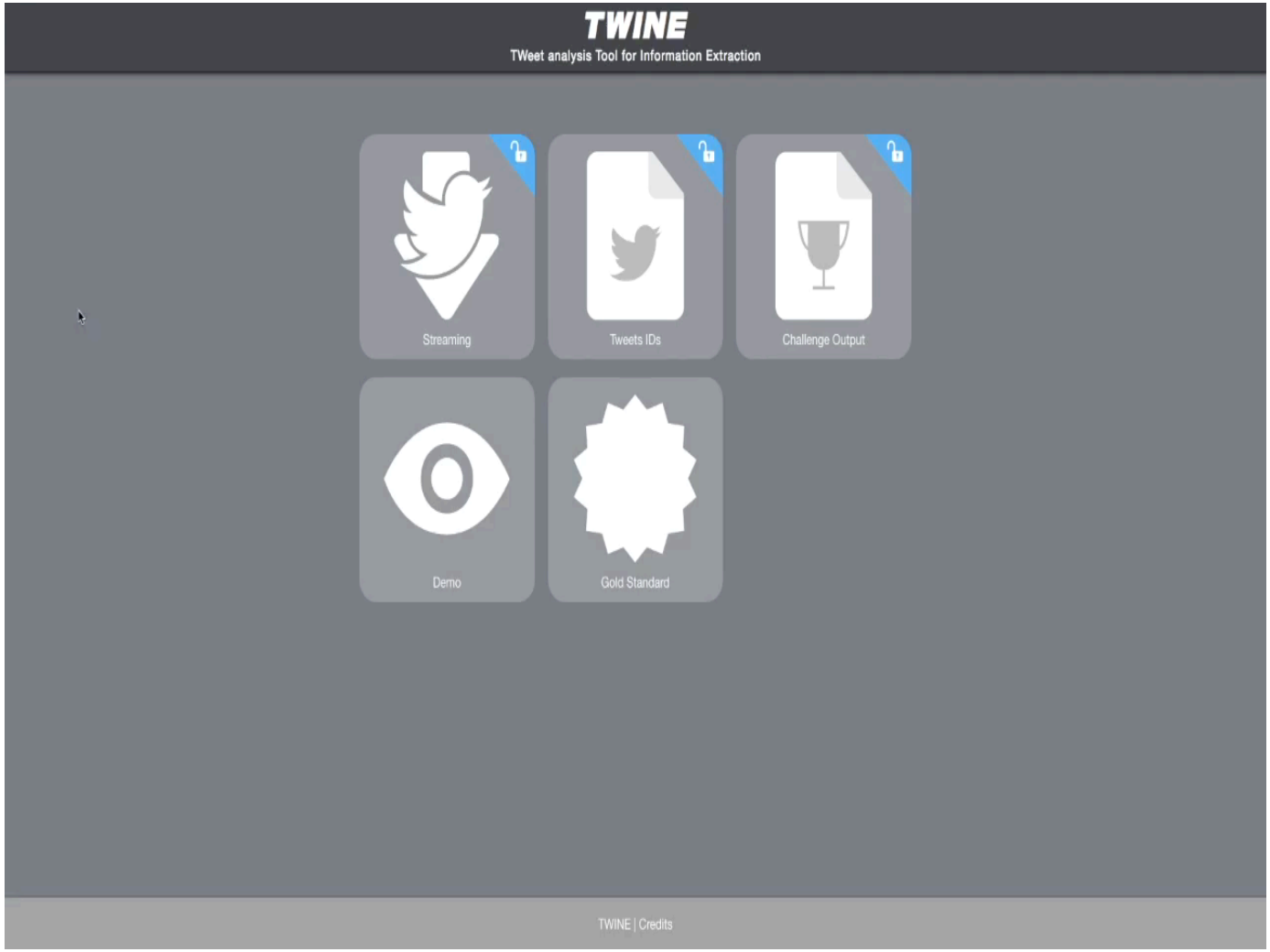




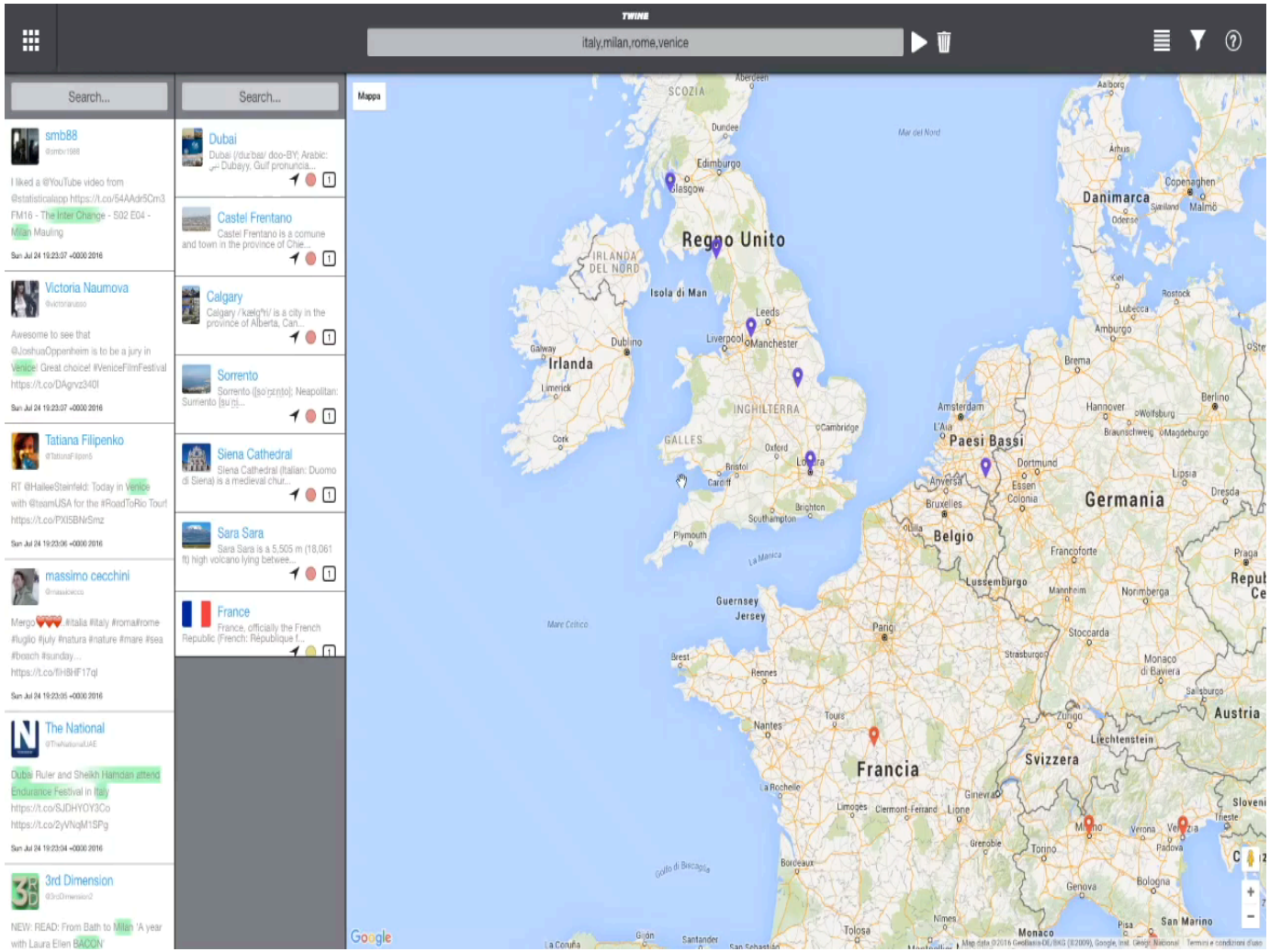




SCHERMATA PRINCIPALE



MODALITÀ LISTA



Matching for Focused Relation Extraction in the eCommerce Domain

(Master Thesis Project by Paolo Smedile)

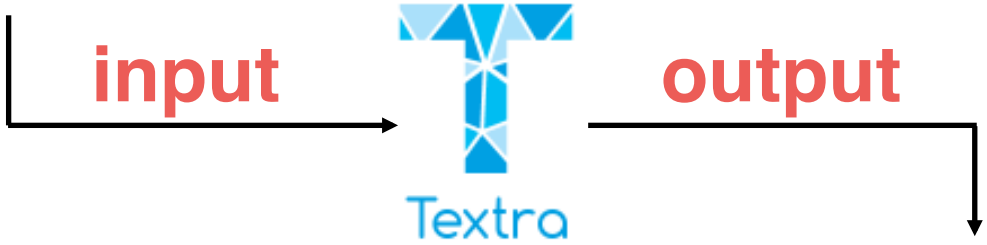
Slides credits:
Paolo Smedile (CS Master Graduate)

Estrazione Descrizione Prodotto da Testo

APPLE iPad mini Retina Wi-Fi 16GB Space Gray

Processore Chip A7 e Coprocessore di movimento M7 Display Retina multitouch da 7.9 pollici Memoria interna 16GB - Connettività WiFi Bluetooth 4.0 - Fotocamera iSight 5MPixel Sistema operativo iOS 7

Disponibile



Risultati

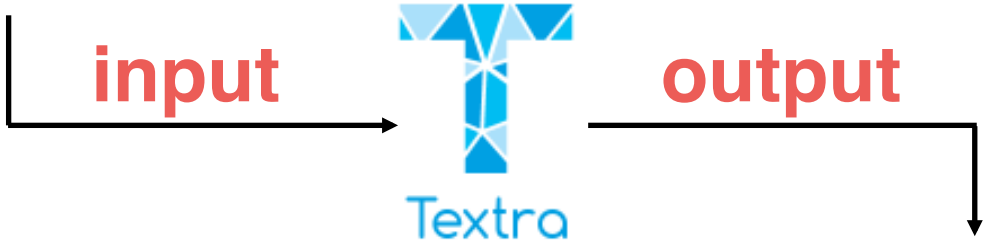
#Attributo	#Valore Catalogo	#Valore Algoritmo
#Brand	Apple	APPLE
#Model	iPad MINI	iPad mini Retina
#Sistema Operativo	Microsoft Windows 7	iOS 7
#Memoria	16 GB	16 GB [0.85509617]
#Processore	A7	A7
#Lunghezza diagonale	7.9	7.9 pollici [0.46233651]
#Tipo display	7.9" IPS TFT	Retina
#Megapixel fotocamera principale	5 Megapixel	5 MP [0.99997662]

Estrazione Descrizione Prodotto da Testo

APPLE iPad mini Retina Wi-Fi 16GB Space Gray

~~Processore Chip A7 e Coprocessore di movimento M7 Display Retina~~ multitouch da 7.9 pollici Memoria interna 16GB - Connettività WiFi Bluetooth 4.0 - Fotocamera iSight 5MPixel Sistema operativo iOS 7

Disponibile



Risultati

#Attributo	#Valore Catalogo	#Valore Algoritmo
#Brand	Apple	APPLE
#Model	iPad MINI	iPad mini Retina
#Sistema Operativo	Microsoft Windows 7	iOS 7
#Memoria	16 GB	16 GB [0.85509617]
#Processore	A7	A7
#Lunghezza diagonale	7.9	7.9 pollici [0.46233651]
#Tipo display	7.9" IPS TFT	Retina
#Megapixel fotocamera principale	5 Megapixel	5 MP [0.99997662]

Estrazione Relazioni da Titolo

Runtime

APPLE iPad mini Retina Wi-Fi 16GB Space Gray

Processore Chip A7 e Coprocessore di movimento M7, Display Retina multitouch da 7.9 pollici e Memoria Interna da 16GB - Connettività WiFi Bluetooth 4.0 - Fotocamere iSight da 5MPixel e Sistema operativo iOS 7

Disponibile

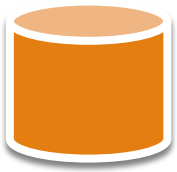
Frammenti testuali di interesse estratti

- Processore Chip A7 e Coprocessore
- e Memoria Interna da 16GB - Connettività
- Display Retina multitouch da 7.9 pollici e
- da 5MPixel e sistema operativo iOS 7
- - Fotocamere iSight da 5MPixel e Sistema

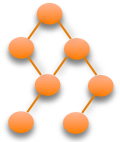
Coppie Attributo-Valore estratte

Valori Nominali	Valori Numerici Tipizzati
Processore - "A7"	Memoria Integrata - 16GB
Tecnologia Display - "Retina"	Lunghezza Diagonale - 7.9 pollici
Sistema Operativo - "iOS 7"	Megapixel Fotocamera Principale - 5MP

Design Time



Context Patterns
pattern testuali che indicano possibile relazione



Ontologia Attributi
Sinonimi, Valori, Vincoli su Codominio, Distribuzioni dei Valori

Occorrenza di Context Pattern
E.g., ***"da"***

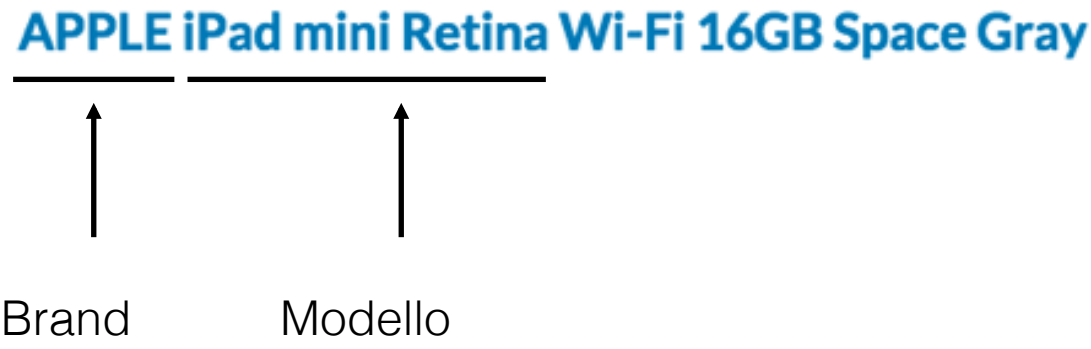
Estrazione <att,value> potenziali con match su nomi-valori atttributi
E.g., "memoria" in att
"retina" in value
match con regex valore

Selezione <att,value> estratte con confronto statistico attendibilità valore estratto (distribuzioni valori)

Estrazione Relazioni da Titolo

Premessa:

L'algoritmo è in grado di estrarre dal titolo gli attributi **brand** e **modello** che permettono di identificare univocamente un prodotto commerciale.



iPhone 6 32GB = iPhone 6 64GB

Estrazione Relazioni da Titolo

Approccio di estrazione:
Il problema dell'estrazione degli attributi brand e modello dal titolo di un annuncio viene **trattato** come un problema di classificazione.

APPLE iPad mini Retina Wi-Fi 16GB Space Gray

B	M	M	M	O	O	O	O

B —> Brand

M —> Model

O —> Other

Approccio estrazione relazioni dal titolo

Approccio di estrazione:

Approccio supervisionato: imparare a classificare in automatico un titolo arbitrario a partire da un insieme di titoli annotati manualmente

Un approccio di questo tipo prevede:

- **Scelta di un modello** da utilizzare per la classificazione
  - *Support Vector Machine*
- **Scelta di un insieme di feature** capace di caratterizzare i dati passati in input al classificatore
- **Insieme di dati annotati** manualmente (training set)

Features di classificazione

Textra utilizza 3 differenti set di feature:

- **Features di posizione**: considerano la posizione di una parola all'interno del titolo
- **Features di ortografiche**: considerano l'ortografia di una parola
- **Features di contesto**: considerano il contesto di una parola

Features di posizione

Features di posizione:

- **Feature 1**: distanza della parola dall'inizio del titolo (int)
- **Feature 2:** distanza della parola dalla fine del titolo (int)

APPLE iPad mini Retina Wi-Fi 16GB Space Gray

Esempio Token: iPad

Feature 1: valore 1
Feature 2: Valore 7

Features di ortografiche

Features di ortografiche:

- **Feature 3**: la parola contiene un numero (bool)
- **Feature 4:** la parola contiene solo numeri (bool)
- **Feature 5**: la parola è in maiuscolo (bool)
- **Feature 6**: solo la prima lettera è in maiuscolo (bool)
- **Feature 7**: lunghezza della parola (int)
- **Feature 8**: la parola è tra parentesi (int)
- **Feature 9**: la parola inizia/finisce con una parentesi (bool)

APPLE iPad mini Retina Wi-Fi 16GB Space Gray

Esempio Token: iPad

1, 7 | 0,0,0,0,4,0,0 |

Features di contesto

Features di contesto:

Basate sul catalogo

- **Feature 10**: la parola è un brand (dizionario)
- **Feature 11:** la parola precedente è un brand (dizionario)
- **Feature 12**: è un possibile modello (algoritmo)
- **Feature 13**: la parola è un colore (dizionario)
- **Feature 14**: la parola è una sigla comune (dizionario)

APPLE iPad mini Retina Wi-Fi 16GB Space Gray

Esempio Token: iPad

1, 7 | 0,0,0,0,4,0,0 | 0, 1, 0, 0, 0

Output del problema

APPLE iPad mini Retina Wi-Fi 16GB Space Gray

```
Token: APPLE Vector: 0,8,1,0,1,1,5,0,0,1,0,1,0,0
Token: iPad Vector: 1,7,0,0,0,0,4,0,0,0,1,0,0,0
Token: mini Vector: 2,6,0,0,0,0,4,0,0,0,0,0,0,0
Token: Retina Vector: 3,5,0,0,1,6,0,0,0,0,0,0,0
Token: Wi-Fi Vector: 4,4,0,0,0,1,5,0,0,0,0,0,0,1
Token: 16GB Vector: 5,3,1,0,1,0,4,0,0,0,0,1,0,0
Token: Space Vector: 6,2,0,0,0,1,5,0,0,0,0,0,1,0
Token: Gray Vector: 7,1,0,0,0,1,4,0,0,0,0,0,1,0

Token :Apple Tag: Brand
Token :iPad Tag: Model
Token :mini Tag: Model
Token :Retina Tag: Model
Token :Wi-Fi Tag: Others
Token :16GB Tag: Others
Token :Space Tag: Others
Token :Gray Tag: Others


Brand -> Apple
Model -> iPad mini Retina
Other -> Wi-Fi 16GB Space Gray
```

Risultati sperimentazione

Risultati approccio di estrazione degli attributi dal titolo							
		POS+ORTO			POS+ORTO+CONT		
	Attributo	precision	recall	f1-score	precision	recall	f1-score
	brand	0,97	0,91	0,94	**0,97**	**0,97**	**0,97**
	model	0,87	0,83	0,85	**0,93**	**0,92**	**0,92**
	other	0,92	0,95	0,93	**0,96**	**0,97**	**0,97**
Smartphone	total	0,92	0,90	0,91	**0,96**	**0,95**	**0,95**
	brand	**0,98**	0,93	0,95	0,97	**0,98**	**0,97**
	model	0,84	0,89	0,86	**0,92**	**0,91**	**0,91**
	other	0,93	0,91	0,92	**0,95**	**0,95**	**0,95**
Tablet	total	0,92	0,91	0,91	**0,95**	**0,95**	**0,95**
	brand	**0,97**	0,92	0,94	0,96	**0,97**	**0,97**
	model	0,90	0,83	0,86	**0,91**	**0,87**	**0,89**
	other	0,90	0,95	0,92	**0,94**	**0,95**	**0,94**
Frigoriferi	total	0,92	0,90	0,91	**0,93**	**0,93**	**0,93**
	brand	0,96	0,88	0.92	**0,97**	**0,94**	**0,96**
	model	0,89	0,86	0,87	**0,91**	**0,90**	**0,90**
	other	0,91	0,96	0,93	**0,94**	**0,96**	**0,95**
Lavatrici	total	0,92	0,90	0,91	**0,94**	**0,93**	**0,94**

Metodologia: Cross validation (10-fold)

Risultati sperimentazione

Risultati con tipologie di prodotti mischiate							
		POS+ORTO			POS+ORTO+CONT		
	Attributo	**precision**	**recall**	**f1-score**	**precision**	**recall**	**f1-score**
Mixed	**brand**	**0,98**	0,89	0,93	0,96	**0,97**	**0,97**
	model	0,83	0,82	0,83	**0,88**	**0,90**	**0,89**
	other	0,90	0,93	0,91	**0,95**	**0,93**	**0,94**
	total	0,90	0,88	0,89	**0,93**	**0,94**	**0,93**

Metodologia: Cross validation (10-fold)

Estrazione Relazioni da Descrizioni

Runtime

APPLE iPad mini Retina Wi-Fi 16GB Space Gray

Processore Chip A7 e Coprocessore di movimento M7, Display Retina multitouch da 7.9 pollici e Memoria Interna da 16GB - Connettività WiFi Bluetooth 4.0 - Fotocamere iSight da 5MPixel e Sistema operativo iOS 7

Disponibile

Frammenti testuali di interesse estratti

Processore Chip A7 e Coprocessore

e Memoria Interna da 16GB - Connettività

Display Retina multitouch da 7.9 pollici e

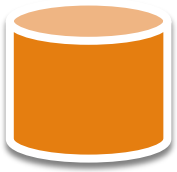
da 5MPixel e sistema operativo iOS 7

- Fotocamere iSight da 5MPixel e Sistema

Coppie Attributo-Valore estratte

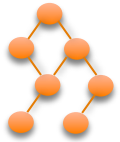
Valori Nominali	Valori Numerici Tipizzati
Processore - "A7"	Memoria Integrata - 16GB
Tecnologia Display - "Retina"	Lunghezza Diagonale - 7.9 pollici
Sistema Operativo - "iOS 7"	Megapixel Fotocamera Principale - 5MP

Design Time



Context Patterns

pattern testuali che indicano possibile relazione



Ontologia Attributi

Sinonimi, Valori, Vincoli su Codominio, Distribuzioni dei Valori

Occorrenza di Context Pattern

E.g., ***"da"***

Estrazione <att,value> potenziali con match su nomi-valori atttributi

E.g., "memoria" in att
"retina" in value
match con regex valore

Selezione <att,value> estratte con confronto statistico attendibilità valore estratto (distribuzioni valori)

Ontologia degli attributi

Descrive tutti gli attributi (54) che Textra è in grado di estrarre dalle 4 tipologie di prodotti considerati. Formalmente è definita come:

$$O = (ao_1, ao_2, \ldots, ao_n)$$

$$ao_i = \langle nome, sinonimi, tipo, dominio, codominio \rangle$$

- **Nome**: attributo di riferimento nel catalogo
- **Sinonimi**: lista di parole utilizzate per riferirsi all'attributo.
- **Tipo**: indica la tipologia di valori (nominale o numerica-tipizzata)
- **Vincolo sul dominio**: indica i domini di riferimento
- **Vincolo sul codomio**: contiene la specifica del codominio
  - **Intensionale** per gli attributi numerici-tipizzati, basata su una espressione regolare e non dipende dalla categoria di prodotto
  - **Estensionale** per gli attributi con valori nominali. Viene fornito un elenco di possibili valori estratti dal catalogo

Ontologia degli attributi

```
{
  "nome": "Memoria RAM",
  "tipo": "numerico_tipizzato",
  "domimio": "cellulari",
  "codominio": "data_size",
  "sinonimi": [
    {
      "key": "ram",
      "univoco": "True"
    },
    {
      "key": "memory",
      "univoco": "False"
      },
    {
      "key": "memoria",
      "univoco": "False"
    }
  ]
}
```

```
{
  "nome": "Processore",
  "tipo": "nominale",
  "domimio": "cellulari",
  "codominio":"Processore.txt",
  "sinonimi": [
    {
      "key": "cpu",
      "univoco": "False"
    },
    {
      "key": "processore",
      "univoco": "False"
    }
  ]
}
```

```
"data_size":{
  "value_regex":"(\d{1,4})((,|.)(\d{1,2})){0,1}",
  "unit_regex":"(KB|MB|GB)"
},
```

Context pattern e loro estrazione

Approccio di estrazione basato sui context pattern

In una descrizione ogni valore presente viene generalmente contestualizzato, ovvero, nel suo intorno occorre un frammento testuale che indica la presenza di una coppia attributo-valore. Es. “memoria RAM *da* 32gb”

Dati 1) un‘ontologia di attributi, 2) un insieme di valori noti in un knowledge graph (KG), e 3) una descrizione testuale, è possibile estrarre una serie di espressioni linguistiche dette **context pattern** che è possibile separino gli attributi dai loro rispettivi valori

Esempio. “memoria RAM da 32gb”

Context Pattern: “da”

I **context pattern** sono **generici** (e.g., “da”, “di”)
Estratti cercando sequenze di parole che separano nomi di attributi noti come specificato nella ontologia, e loro valori noti nel KG

Estrazione Frammenti di Interesse con Context Pattern

Un **frammento testuale di interesse** è definito come una stinga contenente il context pattern più le 3 parole che lo precedono e che lo seguono e che **contiene con alta probabilità la specifica** di una coppia attributo-valore per un qualche attributo.

$word_1$ $word_2$ $word_3$ **[context_pattern]** $word_1$ $word_2$ $word_3$

"**a 32GB 512MB di Ram display TFT**"

Context Pattern

Context prima del pattern

Context dopo il pattern

Estrazione coppie attributo-valore

Textra cercherà all'interno del Frammento testuale di interesse un riferimento ad un attributo dello schema ("memoria") e successivamente un possibile valore.

"5MP e memoria da 256GB , batteria"

(Ram - 256 GB)
(Memoria Interna - 256GB)

Verifica attendibilità valori estratti

Verifica attendibilità valori estratti

Per gli attributi con **valori nominali** Textra è in grado di estrarre esclusivamente **valori noti presenti nel catalogo** di riferimento. Questo limite è però più che accettabile considerando che:

- In campo e-commerce gli attributi aventi valori nominali si evolvono in genere più lentamente rispetto a quelli numerici.
- In contesto e-commerce è meglio sacrificare la recall piuttosto che la precision.

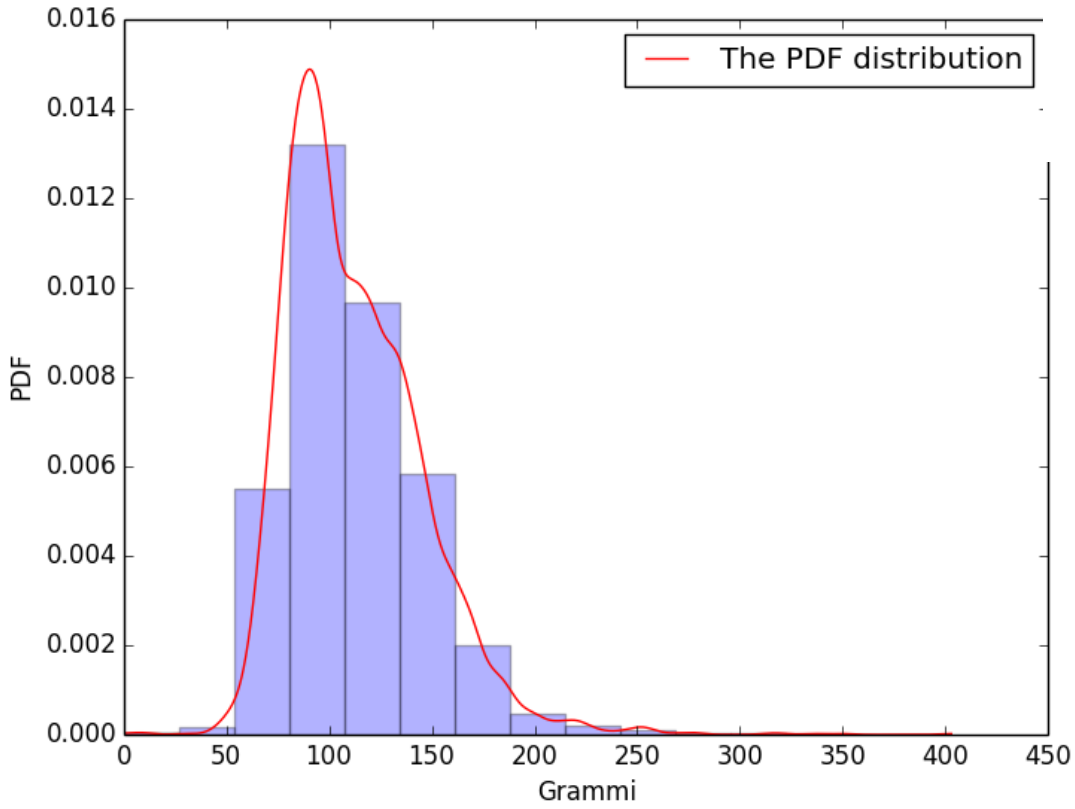
Score di ammissibilità

Lo score di ammissibilità è basato sul calcolo del **Gaussian Kernel Density** che è un metodo non parametrico di stima della densità di una variabile aleatoria. Si basa essenzialmente su una media di funzioni non negative, dette funzioni Kernel (K), centrate attorno a ciascuna osservazione campionaria x_1, ..., x_n. Il valore stimato della densità in un punto X_0 si ottiene:

$$pdf(x_0) = (\frac{1}{nh}) \sum_{x_i=1}^{n} K(\frac{x_0 - x_i}{h})$$

h è un parametro di regolarizzazione chiamato ampiezza di banda ed è molto importante perché controlla la forma della funzione di densità. Esistono alcuni metodi automatici di scelta di tale parametro e Textra utilizza il metodo di Scott.

Score di ammissibilità



Distribuzione valori attributo peso

$$pdf(x_0) = (\frac{1}{nh}) \sum_{x_i=1}^{n} K(\frac{x_0 - x_i}{h})$$

Valore	KDE
0 gr	[2.91618319e-05]
10 g	[4.10611878e-05]
50g	[0.00047986]
100 g	[0.01254165]
150gr	[0.00506272]
200gr	[0.00047968]
300 grammi	[6.38518055e-07]
800 gr	[0.]

Score di ammissibilità - Problemi Affrontati

Diversi simboli per indicare una unità di misura

Mapping simboli nel catalogo con possibili alternative

"g":["g","gr","grammi"],

Conversione multipli di una unità di misura all'unità base

Creazione di una scala di conversione

1GB —> 1024 MB

Estrazione coppie attributo-valore con eliminazione coppie non attendibili

“5MP e memoria da 256GB , batteria”

Possibili coppie attributo-valore

(Ram - 256 GB) ✓

(Memoria Interna - 256GB) ✗

Coppie attributo-valore estratte

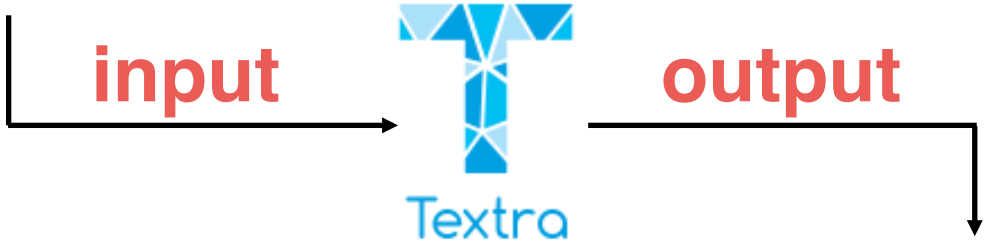
Ram - 256 GB

Estrazione Descrizione Prodotto da Testo

APPLE iPad mini Retina Wi-Fi 16GB Space Gray

Processore Chip A7 e Coprocessore di movimento M7 Display Retina multitouch da 7.9 pollici Memoria interna 16GB - Connettività WiFi Bluetooth 4.0 - Fotocamera iSight 5MPixel Sistema operativo iOS 7

Disponibile



Risultati

#Attributo	#Valore Catalogo	#Valore Algoritmo
#Brand	Apple	APPLE
#Model	iPad MINI	iPad mini Retina
#Sistema Operativo	Microsoft Windows 7	iOS 7
#Memoria	16 GB	16 GB [0.85509617]
#Processore	A7	A7
#Lunghezza diagonale	7.9	7.9 pollici [0.46233651]
#Tipo display	7.9" IPS TFT	Retina
#Megapixel fotocamera principale	5 Megapixel	5 MP [0.99997662]

Risultati Sperimentazione

	Estratti	Estraibili	Esatti	Errati	Precision	Recall	F1-Mesure
Cellulari	150	241	137	13	91,33	56,85	70,08
Tablet	128	205	121	7	94,53	59,02	72,67
Frigoriferi	82	129	80	2	97,56	62,02	75,83
Lavatrici	88	128	86	3	93,63	67,19	79,26
Total	448	703	424	25	94,43	60,31	74,50

La sperimentazione è stata effettuata su 50 annunci per ogni categoria di prodotti

Per attributi estraibili si intendono tutti quegli attributi considerati dall'algoritmo e che di conseguenza dovrebbe essere in grado di estrarre

Ata precisione, recall migliorabile

Nota:

iOS 8 (estratto) - iOS (catalogo) —> corretto
iOS (estratto) - iOS 8 (catalogo) —> errato

Textra