

# Tint & NER

*Rachele Sprugnoli – [rachele.sprugnoli@unicatt.it](mailto:rachele.sprugnoli@unicatt.it)*

Centro Interdisciplinare di Ricerche per la Computerizzazione  
dei Segni dell'Espressione (CIRCSE)



UNIVERSITÀ  
CATTOLICA  
del Sacro Cuore

# COSA FAREMO

6 MARZO: Introduzione teorica e uso di una pipeline (demo online)

**13 MARZO: Uso di una pipeline per l'italiano + focus su NER**

20 MARZO: Estrazione di temi e parole chiave

27 MARZO: Estrazione del sentiment

# Tint - The Italian NLP Tool

Pipeline per i task di:

- sentence splitting
- tokenizzazione
- PoS tagging
- lemmatizzazione
- analisi morfologica
- dependency parsing
- NER
- analisi dei verbi composti
- keyphrase extraction
- analisi dei derivati
- leggibilità

Sito web: <http://tint.fbk.eu/>

# USIAMO Tint

- Apriamo il terminale: ok PowerShell e CMD
- Decomprimere la cartella `Tint` scaricata
- Cambiamo il nome del file `jar` in `tint.jar`
- Usare il comando `cd` per andare nella cartella `Tint`
- Digitare il seguente comando e premere Invio:

```
java -Dfile.encoding=UTF-8 -jar tint.jar -c  
default-config.properties -i prova-news.txt  
-o out-conll.conll -f conll
```

- Digitare il seguente comando e premere Invio:

```
java -Dfile.encoding=UTF-8 -jar tint.jar -c  
default-config.properties -i prova-news.txt  
-o out-json.json
```

# USIAMO Tint

- Leggiamo il comando

```
java -Dfile.encoding=UTF-8 -jar tint.jar -c  
default-config.properties -i prova-news.txt -o  
out-conll.conll -f conll
```

1. `java`: diciamo al computer che il programma è scritto in Java
2. `-Dfile.encoding=UTF-8`: specifichiamo l'encoding del testo in input: fondamentale per l'italiano!
3. `-jar`: specifichiamo estensione programma
4. `-c`: specifichiamo il file di configurazione
5. `-i`: specifichiamo il nome/percorso del file in input
6. `-o`: specifichiamo il nome/percorso del file in output
7. `-f`: specifichiamo il formato del file in output (default json)

```
java -jar tint.jar -h
```

# L'OUTPUT DI Tint

- Apriamo il file `out-conll.conll` con un editor di testo

Che task include il formato `conll`?

- Apriamo il file `out-json.json` con un editor di testo

Che task include il formato `json`?

# L'OUTPUT DI Tint: CoNLL

1	In	in	E	0	2	case
2	Italia	Italia	SP	LOC	5	nmod
3	"	[PUNCT]	FB	0	2	punct
4	la	la	RD	0	5	det
5	circolazione	circolazione	S	0	9	nsubj
6	dei	del	E+RD	0	7	case
7	virus	virus	S	0	5	nmod
8	influenzali	influenzale	A	0	7	amod
9	inizia	iniziare	V	0	0	ROOT
10	ad	ad	E	0	11	mark
11	intensificarsi	intensificare	V+PC	0	9	xcomp
12	"	[PUNCT]	FB	0	11	punct
13	e	e	CC	0	9	cc
14	si	si	PC	0	15	expl:impers
15	avvicina	avvicinare	V	0	9	conj
16	l'	l'	RD	0	17	det
17	inizio	inizio	S	0	15	dobj
18	del	del	E+RD	0	19	case
19	periodo	periodo	S	0	17	nmod
20	epidemico	epidemico	A	0	19	amod
21	.	[PUNCT]	FS	0	9	punct

Deprel: <https://universaldependencies.org/u/dep/>

# L'OUTPUT DI Tint: CoNLL

## COLONNE:

- 1) ID, identificativo numerico del token, riparte da 1 per ogni nuova frase. Le frasi sono separate da una riga vuota
- 2) token
- 3) lemma
- 4) PoS: [http://medialab.di.unipi.it/wiki/Tanl\\_POS\\_Tagset](http://medialab.di.unipi.it/wiki/Tanl_POS_Tagset)
- 5) Named Entity
- 6) ID della testa della parola nel parsing a dipendenze
- 7) etichetta della relazione a dipendenze:  
<https://universaldependencies.org/u/dep/>

# L'OUTPUT DI Tint

- Lemma, PoS, morfologia,NER...

PoS tagset:

[http://medialab.di.unipi.it/wiki/TanI\\_POS\\_Tagset](http://medialab.di.unipi.it/wiki/TanI_POS_Tagset)

```
{
  "index": 2,
  "word": "Piemonte",
  "originalText": "Piemonte",
  "lemma": "Piemonte",
  "characterOffsetBegin": 750,
  "characterOffsetEnd": 758,
  "pos": "SP",
  "featuresText": "",
  "ner": "LOC",
  "full_morpho": "Piemonte",
  "selected_morpho": "",
  "guessed_lemma": true,
  "features": {},
  "contentWord": true,
  "literalWord": true,
  "hyphenation": "Pie-mon-te",
  "difficultyLevel": 4,
  "easyWord": true
},
```

# L'OUTPUT DI Tint

- Lemma, PoS, morfologia, NER...

PoS tagset:

[http://medialab.di.unipi.it/wiki/TanI\\_POS\\_Tagset](http://medialab.di.unipi.it/wiki/TanI_POS_Tagset)

```
{
  "index": 7,
  "word": "virus",
  "originalText": "virus",
  "lemma": "virus",
  "characterOffsetBegin": 31,
  "characterOffsetEnd": 36,
  "pos": "S",
  "featuresText": "Gender\u003dMasc|Number\u003dPlur",
  "ner": "O",
  "full_morpho": "virus virus+n+m+plur virus+n+m+sing",
  "selected_morpho": "virus+n+m+plur",
  "guessed_lemma": false,
  "features": {
    "Gender": [
      "Masc"
    ],
    "Number": [
      "Plur"
    ]
  },
  "contentWord": true,
  "literalWord": true,
  "hyphenation": "vi-rus",
  "difficultyLevel": 4,
  "easyWord": true
},
```

# L'OUTPUT DI Tint

- Verbi composti:
  - “è stata superata”

```
"verbs": [  
  {  
    "tokens": [  
      25,  
      26,  
      27  
    ],  
    "isPassive": true,  
    "tense": "PrPast",  
    "mood": "Ind",  
    "person": 3,  
    "gender": "Fem",  
    "number": "Sing"  
  }  
]
```

# L'OUTPUT DI Tint

- Parole chiave

```
{
  "keyphrase": "incidenza",
  "frequency": 2,
  "score": 11.834,
  "idf": 1.0000000000751452,
  "score_boost": 1.0,
  "pattern_boost": 0.0,
  "chain_length": 1,
  "lemmas": [
    "incidenza"
  ],
  "stems": [
    "incident"
  ],
  "synonyms": [],
  "tokens": [
    "incidenza"
  ],
  "posList": [
    "S"
  ]
},
```

# L'OUTPUT DI Tint

- Forme derivate

Informazione estratta  
dal **derIvaTario**:

<http://derivatario.sns.it/>

Esempio: *influenzale*

```
"derivation": {
  "baseLemma": "fluire",
  "baseType": "presp",
  "phases": [
    {
      "affix": "2in",
      "allomorph": "in",
      "mt": "mt1",
      "ms": "ms2b",
      "type": "affixation"
    },
    {
      "conversionType": "v_a",
      "type": "conversion"
    },
    {
      "affix": "nza",
      "allomorph": "nza",
      "mt": "mt6",
      "ms": "ms2b",
      "type": "affixation"
    },
    {
      "affix": "ale",
      "allomorph": "ale",
      "mt": "mt1",
      "ms": "ms1",
      "type": "affixation"
    }
  ]
}
```

# L'OUTPUT DI Tint

- Leggibilità
  - **Level1**: 500 parole più facili
  - **Level2**: 2500 parole più facili
  - **Level3**: le 5000 parole più facili
  - **TTR**: type/token ratio
  - **Density**: #content words / #words
  - **Deep\***: profondità albero sintattico (# sintagmi)
  - **SubordinateRatio**: #subordinate / #proposizioni

N.B. Parole tratte dal “Vocabolario di Base dell’Italiano” di De Mauro

```
"readability": {  
  "level1WordSize": 19,  
  "level2WordSize": 41,  
  "level3WordSize": 47,  
  "language": "it",  
  "contentWordSize": 86,  
  "contentEasyWordSize": 75,  
  "wordCount": 145,  
  "docLenWithSpaces": 909,  
  "docLenWithoutSpaces": 769,  
  "docLenLettersOnly": 746,  
  "goodSentenceCount": 6,  
  "sentenceCount": 6,  
  "tokenCount": 168,  
  "hyphenCount": 317,  
  "hyphenWordCount": 141,  
  "ttrValue": 0.78,  
  "density": 0.593103448275862,  
  "deepAvg": 4.833333333333333,  
  "deepMax": 6.0,  
  "subordinateRatio": 0.0,  
  "deeps": {  
    "0": 4,  
    "1": 6,  
    "2": 5,  
    "3": 5,  
    "4": 4,  
    "5": 5  
  },  
}
```

# L'OUTPUT DI Tint

- Leggibilità: 
$$89 + \frac{300 * (\text{numero delle frasi}) - 10 * (\text{numero delle lettere})}{\text{numero delle parole}}$$
- **Main = GULPEASE**, basato su numero frasi, lettere e parole. 100 massima leggibilità, 0 minima leggibilità
  - < 80 difficile per chi ha la licenza elementare
  - < 60 difficile per chi ha la licenza media
  - < 40 difficile per chi ha un diploma superiore

```
"forms": {},  
"measures": {  
  "main": 49.96551724137931,  
  "level1": 25.333333333333332,  
  "level3": 54.651162790697676,  
  "level2": 47.674418604651166  
},  
"labels": {  
  "main": "Gulpease"  
},
```

N.B. Vale per l'italiano contemporaneo in prosa. Formula di Flesch per l'inglese.

# L'OUTPUT DI Tint

- Leggibilità
  - **Level1:** quanto è difficile per un lettore che conosce solo le 500 parole più facili dell'italiano
  - **Level2:** quanto è difficile per un lettore che conosce solo le 2500 parole più facili dell'italiano
  - **Level3:** quanto è difficile per un lettore che conosce solo le 5000 parole più facili dell'italiano

```
"forms": {},  
"measures": {  
  "main": 49.96551724137931,  
  "level1": 25.333333333333332,  
  "level3": 54.651162790697676,  
  "level2": 47.674418604651166  
},  
"labels": {  
  "main": "Gulpease"  
},
```

N.B Più il valore è basso, più è difficile.

# L'OUTPUT DI Tint

- Statistiche sui POS tags
  - Categorie granulari
  - Macro-categorie

PoS tagset:

[http://medialab.di.unipi.it/wiki/TanI\\_POS\\_Tagset](http://medialab.di.unipi.it/wiki/TanI_POS_Tagset)

```
"posStats": {  
  "support": {  
    "CC": 4,  
    "FF": 13,  
    "A": 9,  
    "B": 2,  
    "E": 17,  
    "DI": 1,  
    "VA": 6,  
    "FS": 6,  
    "N": 10,  
    "V+PC": 1,  
    "RD": 13,  
    "S": 34,  
    "PC": 1,  
    "V": 13,  
    "E+RD": 13,  
    "FB": 4,  
    "SP": 21  
  }  
},
```

```
"genericPosStats": {  
  "support": {  
    "P": 1,  
    "A": 9,  
    "R": 13,  
    "B": 2,  
    "S": 55,  
    "C": 4,  
    "D": 1,  
    "E": 30,  
    "F": 23,  
    "V": 20,  
    "N": 10  
  }  
}
```

# RICONOSCIMENTO DEI NOMI PROPRI

- È il riconoscimento di nomi propri all'interno del testo e la loro classificazione in un insieme predefinito di categorie di interesse
- Named Entity = tutto ciò a cui ci si può riferire usando un nome proprio
- Nome proprio = designatore rigido che identifica uno specifico elemento all'interno di una categoria

Named Entity Recognition and Classification

=

NERC o solo NER

# QUANTE E QUALI CLASSI

- Tre classi universalmente riconosciute:
  - Persone
  - Luoghi: GPE versus LOC
  - Organizzazioni
- Altre classi comunemente usate: espressioni numeriche (%,\$), date, orari, indirizzi mail, URL...
- Entità specifiche di dominio: nomi di medicine, tipi di proteine, condizioni mediche, nomi di navi, nomi di armi, nomi di veicoli...

# ESEMPI DI CLASSIFICAZIONE - 1

- MUC - Message Understanding Conference:  
[https://www-nlpir.nist.gov/related\\_projects/muc/proceedings/ne\\_task.html](https://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html)
  - ENAMEX: person, organization, location  
`<ENAMEX TYPE="LOCATION">North America</ENAMEX>`
  - TIMEX: date, time  
`<TIMEX TYPE="TIME">8:24 a.m.</TIMEX>`
  - NUMEX: money, percentage  
`<NUMEX TYPE="MONEY">$10 million</NUMEX>`

## ESEMPI DI CLASSIFICAZIONE - 2

- CoNLL-2003: <https://www.aclweb.org/anthology/W03-0419.pdf>
  - PER: persone
  - LOC: luoghi
  - ORG: organizzazioni
  - MISC: miscellanea (nomi che non sono PER, LOC, ORG)  
E.g., “1000 Lakes Rally”

# PERCHÉ È DIFFICILE?

- Variazioni: *Francesco Totti, Totti, Er Pupone, Francesco*
- Presenza di abbreviazioni: *F.T., 3 sett.*
- Cambiano nel corso del tempo: *Bombay vs. Mumbai*
- Metonimia:
  - Organisation vs. Location : *l'**Italia** ha vinto i Mondiali vs. I Mondiali si sono giocati in **Italia***
  - Facility vs. Organisation: *entrare alla **Casa Bianca** vs. la **Casa Bianca** ha approvato l'emendamento*

# PERCHÉ È DIFFICILE?

Nome	Possibili Categorie
<i>Washington</i>	Person, Location, Organization, Vehicle
<i>Downing St.</i>	Location, Organization
<i>Louis Vuitton</i>	Person, Organization, Product

- ***Washington*** nacque in Virginia
- Blair è arrivato a ***Washington*** per la sua ultima visita di stato
- ***Washington*** ha vinto la serie 7 a 5
- La ***Washington*** è lunga 333 metri

## QUANTE NE CI SONO?

ROMA - Il commissario di Alitalia, Giuseppe Leogrande, nella relazione sulla situazione economico-finanziaria inviata in agosto alle commissioni parlamentari competenti sui trasporti, ha affermato che «tra il 17 febbraio e il 25 giugno 2020 Alitalia ha provveduto a rimborsare circa 123 milioni di euro».

# QUANTE NE CI SONO?

ROMA - Il commissario di Alitalia, Giuseppe Leogrande, nella relazione sulla situazione economico-finanziaria inviata in agosto alle commissioni parlamentari competenti sui trasporti, ha affermato che «tra il 17 febbraio e il 25 giugno 2020 Alitalia ha provveduto a rimborsare circa 123 milioni di euro».

8 = 1 LOC - 1 PER - 2 ORG - 3 TIMEX - 1 NUMEX

# ANNOTAZIONE NER: RECOGITO

- Piattaforma online gratuita per l'annotazione e la pubblicazione di testi/immagini: <https://recogito.pelagios.org/>
- Linked (geo-)data

## LINKED DATA

*Linked Data is simply about using the Web to create typed links between data from different sources. (Bizer et al., 2011)*

- 3 task manuali o semi-automatici:
  - 1) riconoscimento NE (PER e LOC)
  - 2) classificazione NE (PER e LOC)
  - 3) linking a gazzettini → mapping

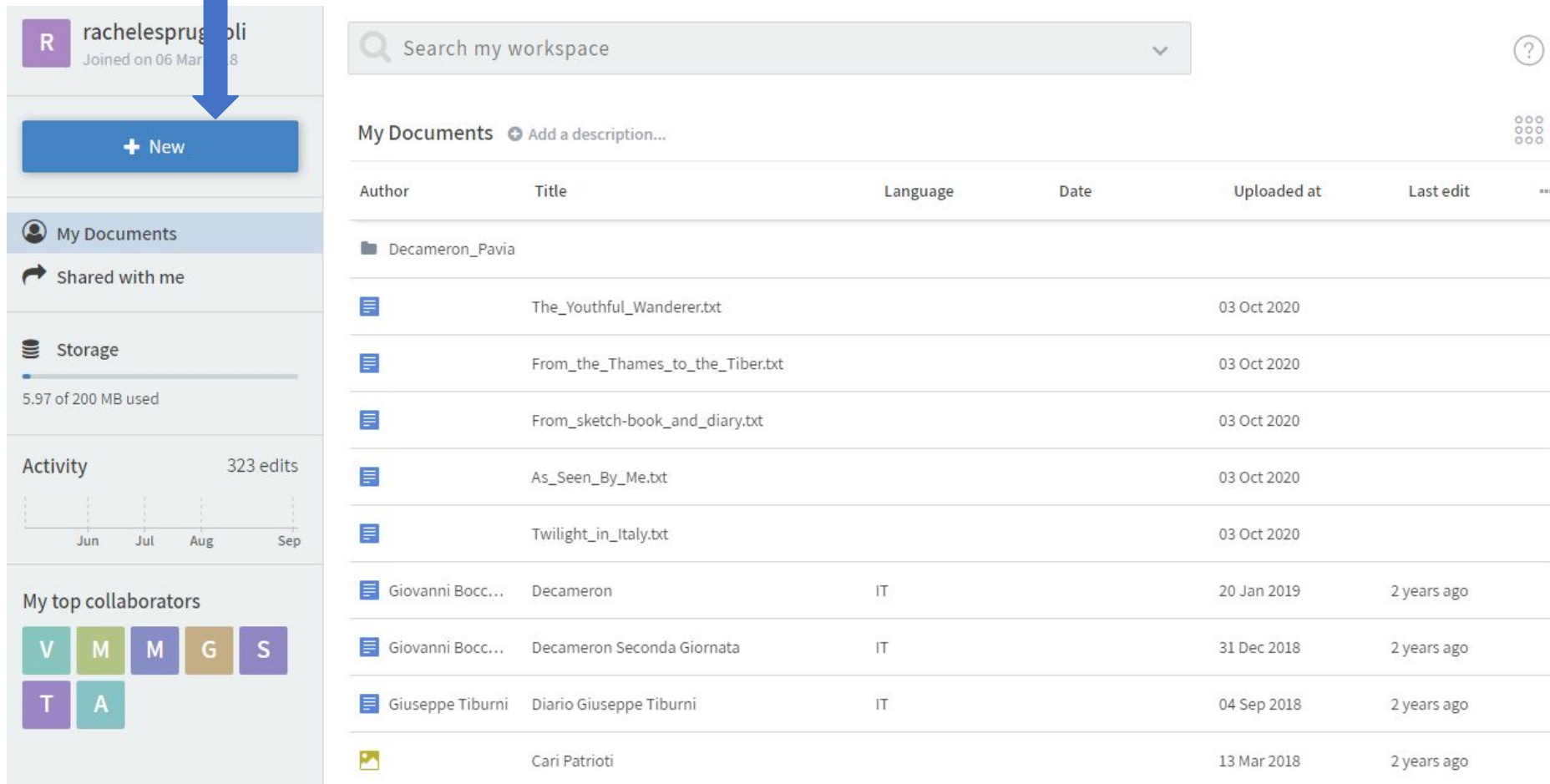
# RECOGITO

- Annotazione di testi (raw - txt files) o immagini (JPEG, PNG, TIFF)
  - Testi:  
<https://recogito.pelagios.org/document/7drvngnipl50zx/part/1/edit>  
<https://recogito.pelagios.org/document/tjrrsqn4dwmgep/part/1/edit>
  - Immagini di manoscritti:  
<https://recogito.pelagios.org/document/gayytt1t2zgzeq/part/1/edit>

# RECOGITO

Caricare un nuovo file: *The\_Youthful\_Wanderer.txt*

<http://www.gutenberg.org/ebooks/10638>



**Left Sidebar:**

- User: **rachelesprugoli** (Joined on 06 Mar 2018)
- + New**
- My Documents**
- Shared with me**
- Storage**: 5.97 of 200 MB used
- Activity**: 323 edits (Jun, Jul, Aug, Sep)
- My top collaborators**: V, M, M, G, S, T, A

**Main Area:**

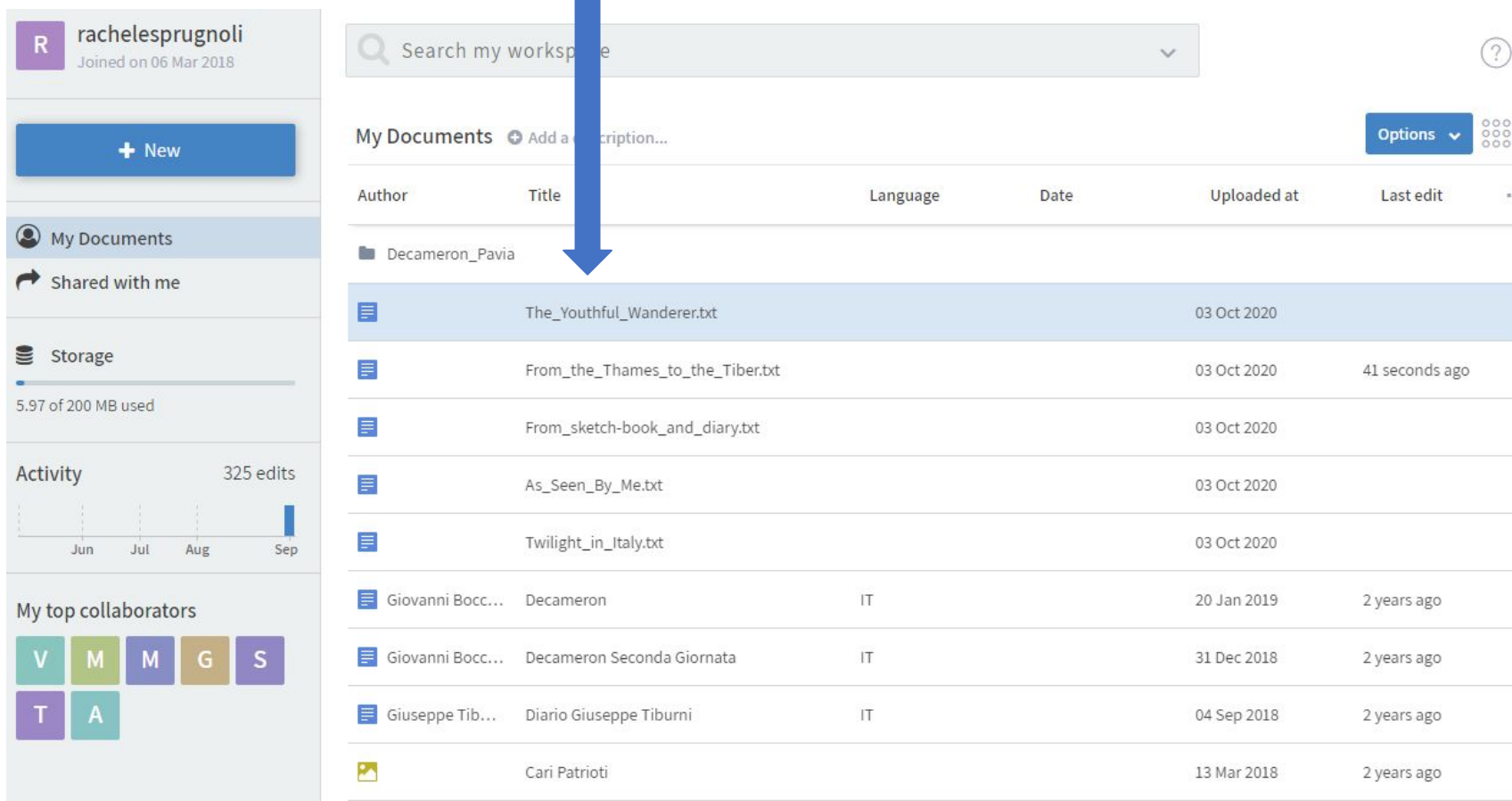
Search my workspace

**My Documents** Add a description...

Author	Title	Language	Date	Uploaded at	Last edit	...
Decameron_Pavia						
	The_Youthful_Wanderer.txt			03 Oct 2020		
	From_the_Thames_to_the_Tiber.txt			03 Oct 2020		
	From_sketch-book_and_diary.txt			03 Oct 2020		
	As_Seen_By_Me.txt			03 Oct 2020		
	Twilight_in_Italy.txt			03 Oct 2020		
Giovanni Bocc...	Decameron	IT		20 Jan 2019	2 years ago	
Giovanni Bocc...	Decameron Seconda Giornata	IT		31 Dec 2018	2 years ago	
Giuseppe Tiburni	Diario Giuseppe Tiburni	IT		04 Sep 2018	2 years ago	
Cari Patrioti				13 Mar 2018	2 years ago	

# RECOGITO

Cliccare una volta su un file



The screenshot displays the RECOGITO workspace interface. On the left, a sidebar shows the user profile 'rachelesprugnoli', a '+ New' button, and sections for 'My Documents', 'Shared with me', 'Storage' (5.97 of 200 MB used), 'Activity' (325 edits), and 'My top collaborators'. The main area features a search bar and a table of documents. A blue arrow points to the first document in the list.

Author	Title	Language	Date	Uploaded at	Last edit
Decameron_Pavia	The_Youthful_Wanderer.txt			03 Oct 2020	
	From_the_Thames_to_the_Tiber.txt			03 Oct 2020	41 seconds ago
	From_sketch-book_and_diary.txt			03 Oct 2020	
	As_Seen_By_Me.txt			03 Oct 2020	
	Twilight_in_Italy.txt			03 Oct 2020	
Giovanni Bocc...	Decameron	IT		20 Jan 2019	2 years ago
Giovanni Bocc...	Decameron Seconda Giornata	IT		31 Dec 2018	2 years ago
Giuseppe Tib...	Diario Giuseppe Tiburni	IT		04 Sep 2018	2 years ago
Cari Patrioti				13 Mar 2018	2 years ago

# RECOGITO

Scegliere tra le  
opzioni



R

rachelesprugnoli

Joined on 06 Mar 2018

+ New

My Documents

Shared with me

Storage

5.97 of 200 MB used

Activity

325 edits

Jun Jul Aug Sep

My top collaborators

V M M G S

T A

Search my workspace

My Documents + Add a description...

Author	Title	Language	Date
Decameron_Pavia			
	The_Youthful_Wanderer.txt		
	From_the_Thames_to_the_Tiber.txt		
	From_sketch-book_and_diary.txt		
	As_Seen_By_Me.txt		
	Twilight_in_Italy.txt		03 Oct 2020
Giovanni Bocc...	Decameron	IT	20 Jan 2019 2 years ago
Giovanni Bocc...	Decameron Seconda Giornata	IT	31 Dec 2018 2 years ago
Giuseppe Tib...	Diario Giuseppe Tiburni	IT	04 Sep 2018 2 years ago
	Cari Patrioti		13 Mar 2018 2 years ago

Options

Open

Open in new tab

Move to

Duplicate

Delete

Share

Explore network

Named Entity Recognition

# RECOGITO

- Annotazione automatica basata su NER (solo sui testi raw)

## Recognition Engines

<input checked="" type="radio"/> Example Kima NER Plugin	he	An attempt to use Kima with the Recogito NER plugin interface.
<input type="radio"/> Stanford CoreNLP	en	The standard engine with the default English language model
<input type="radio"/> Stanford CoreNLP	fr	The standard engine with the default French language model
<input type="radio"/> Stanford CoreNLP	de	The standard engine with the default German language model
<input type="radio"/> Stanford CoreNLP	es	The standard engine with the default Spanish language model
<input type="radio"/> Herodotus Latin NER		An experimental Latin NER plugin by Alex Erdmann

# RECOGITO

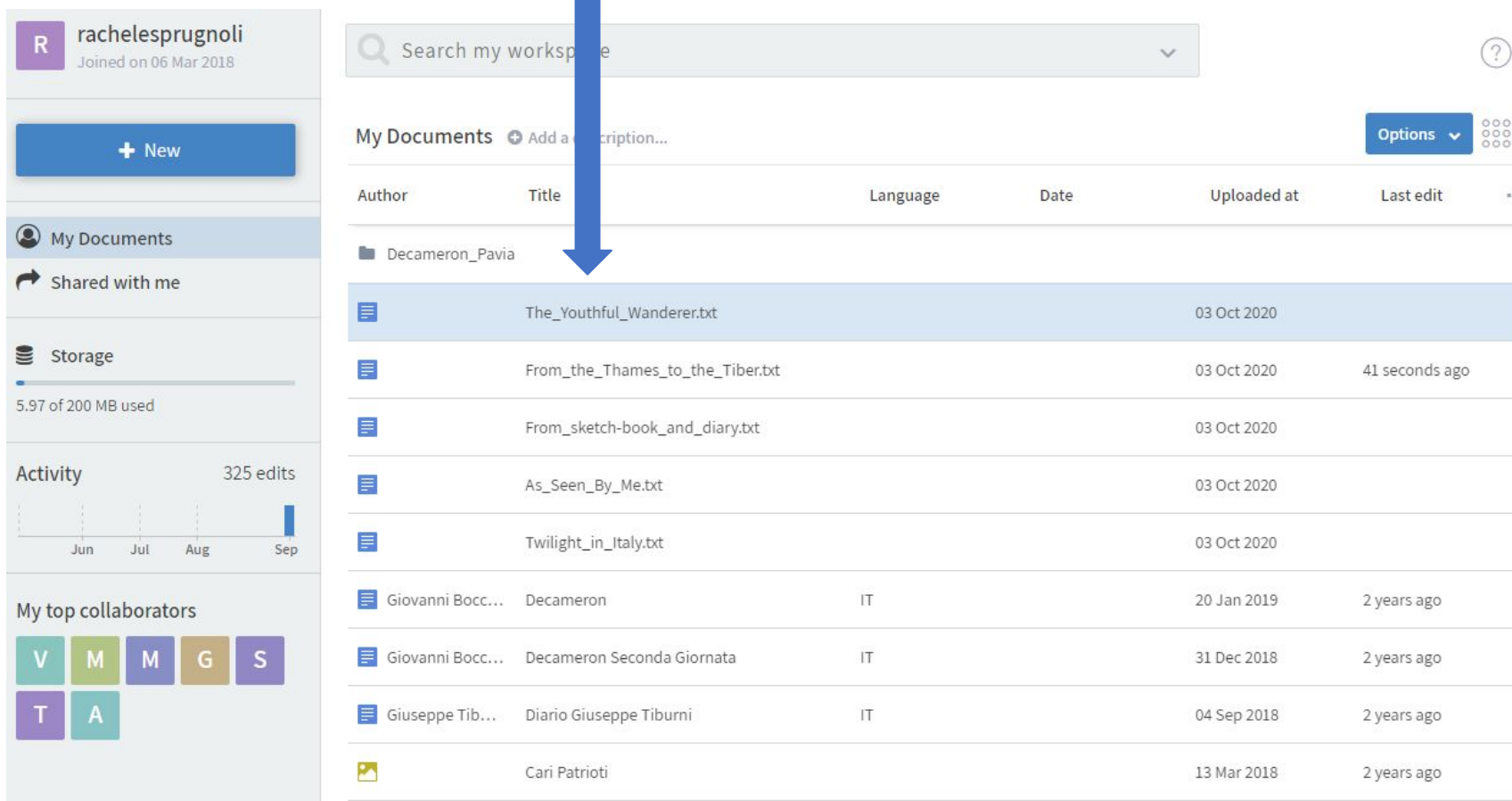
- Uso di gazzettini

<input checked="" type="checkbox"/>	<b>Pleiades</b>	Pleiades Gazetteer of the Ancient World
<input checked="" type="checkbox"/>	<b>CHGIS</b>	China Historical GIS
<input checked="" type="checkbox"/>	<b>DPP Places</b>	Places from the Digitizing Patterns of Power project
<input checked="" type="checkbox"/>	<b>DARE</b>	Digital Atlas of the Roman Empire
<input checked="" type="checkbox"/>	<b>MoEML</b>	Map of Early Modern London
<input checked="" type="checkbox"/>	<b>HGIS de las Indias</b>	Historical-Geographic Information System for Spanish America (1701-1808)
<input checked="" type="checkbox"/>	<b>GeoNames</b>	A subset of GeoNames populated places, countries and first-level administrative divisions
<input checked="" type="checkbox"/>	<b>Kima</b>	Kima Historical Gazetteer - place names in the Hebrew script

- Document settings  → Annotation Preferences

# RECOGITO

Cliccare due volte su un file per aprirlo



The screenshot displays the Recogito web interface. On the left is a sidebar with the user profile 'rachelesprugnoli', a '+ New' button, and sections for 'My Documents', 'Shared with me', 'Storage' (5.97 of 200 MB used), 'Activity' (325 edits), and 'My top collaborators'. The main area features a search bar and a table of documents. A blue arrow points to the first document in the table.

Author	Title	Language	Date	Uploaded at	Last edit	
Decameron_Pavia	The_Youthful_Wanderer.txt			03 Oct 2020		
	From_the_Thames_to_the_Tiber.txt			03 Oct 2020	41 seconds ago	
	From_sketch-book_and_diary.txt			03 Oct 2020		
	As_Seen_By_Me.txt			03 Oct 2020		
	Twilight_in_Italy.txt			03 Oct 2020		
Giovanni Bocc...	Decameron	IT		20 Jan 2019	2 years ago	
Giovanni Bocc...	Decameron Seconda Giornata	IT		31 Dec 2018	2 years ago	
Giuseppe Tib...	Diario Giuseppe Tiberni	IT		04 Sep 2018	2 years ago	
	Cari Patrioti			13 Mar 2018	2 years ago	





# RECOGITO



## Annotazione con un semplice doppio-click

rachelesprugnoli  
Joined on 6 Mar 2018

The\_Youthful\_Wanderer.txt



405 Annotations · No Other Contributors

ANNOTATION MODE: **NORMAL** QUICK ▾ RELATIONS COLOUR: **BY ENTITY TYPE** BY VERIFICATION STATUS BY FIRST TAG

Chapter XVI. **Geneva** to **Turin**.

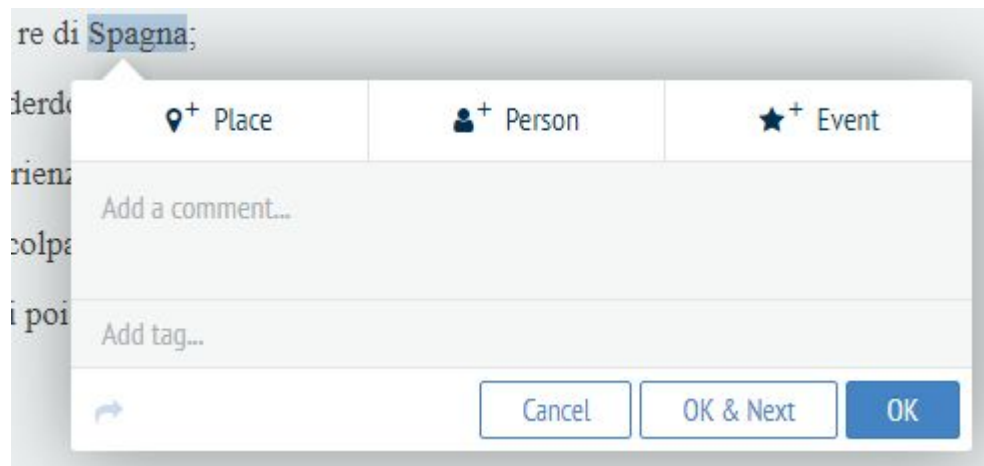
**Switzerland** has two national languages, the **German** and the **French**, both of which are recognized by the Government. **Geneva** is **French**, so I had some trouble in getting my information and procuring a ticket for **Italy**. I left **Geneva** at 6:40 a.m., September 10th; and after passing through a number of tunnels, one of which required 5-1/2 minutes of moderate railway speed, we arrived at **Bellegarde**, on the **French** border, and passed muster. From 9:00 to 10:00 o'clock we were detained at **Culoz**, and by noon we saw the snow-covered Alps again. At 3:30 p.m., we arrived at **Modane** and passed muster for **Italy**. **Mont Cenis Tunnel**. We entered the mouth of this great tunnel, over 8 miles in length, at 4:58-1/2 p.m., and were exactly 26 minutes in the very bowels of the earth, where absolute darkness reigns. Temperature in the middle, 59° Fahrenheit. **Italy**. We now come to a country which contrasts as strangely with the nations of western **Europe**, as those do with **America**, or as **Alpine Switzerland** does with the rest of the world. When I parted at **Paris** with my **New York** friend, he bound for **Rome**, I for the north, we still had our school-boy ideas of **Germany**, **Switzerland** and **Italy**; and I shall never forget the remark which he then made, and which embodied my notions and anticipations perhaps as well as his own. He said, "I suppose we have now seen the brightest side of the picture, the trouble is that scenes will now become tamer as we advance toward the cradle of humanity." I had been pleasantly disappointed almost every time that I entered a new country, but now, as I was entering **Italy**, I expected that I would surely not see much to interest me except her

# RECOGITO



Annotazione con un semplice doppio-click

1. Doppio click su una parola o selezionare espressione multitoken (e.g. Castel Guglielmo)
2. Si apre una finestra di annotazione: scegliere se si tratta di luogo o persona.
3. Aggiungere un tag (sotto-classificazione): e.g., maschio vs. femmina



# RECOGITO



Annotazione con un semplice doppio-click

4. Scegliendo “Place” viene suggerito un linking: se ok cliccare su “Confirm” altrimenti su “Change”

di Spagna;

Place Person Event

Spain  
geonames:2510769  
Spanish State, Spanien, Espagne, Espanha, Estado E...

Automatic Match Confirm Change

Add a comment...

Add tag...

Cancel OK & Next OK

# RECOGITO



Annotazione con un semplice doppio-click

5. Cliccare su “OK”: la stessa annotazione può essere automaticamente applicata ad altre menzioni uguali

**Re-Apply**

There is 1 un-annotated and 1 annotated occurrence of «Spagna» in the text.  
Do you want to re-apply this annotation?

**YES & merge existing annotations**

NO, don't re-apply

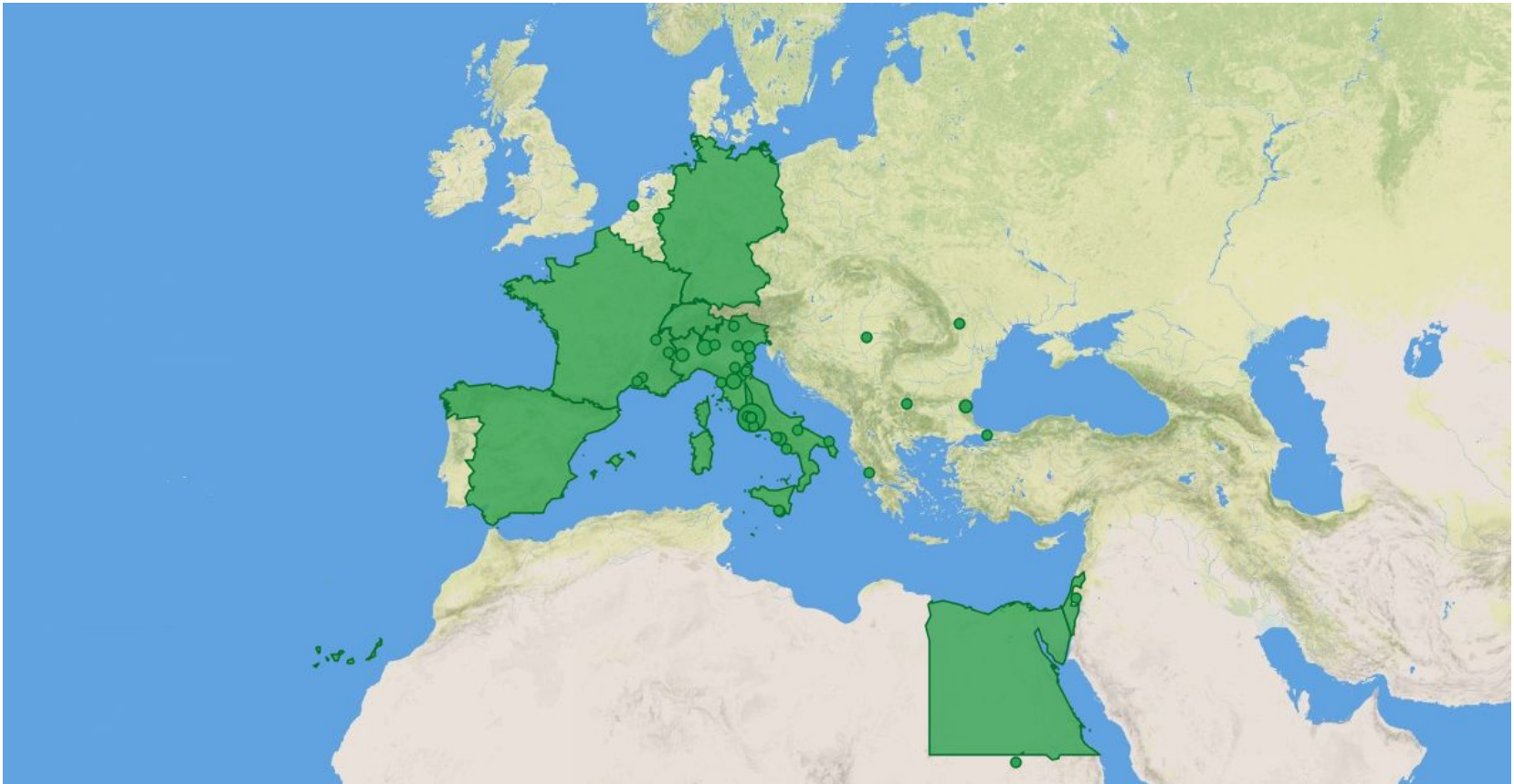
[Show advanced options...](#)

# RECOGITO



Map view

- Visualizzazione geografica immediata

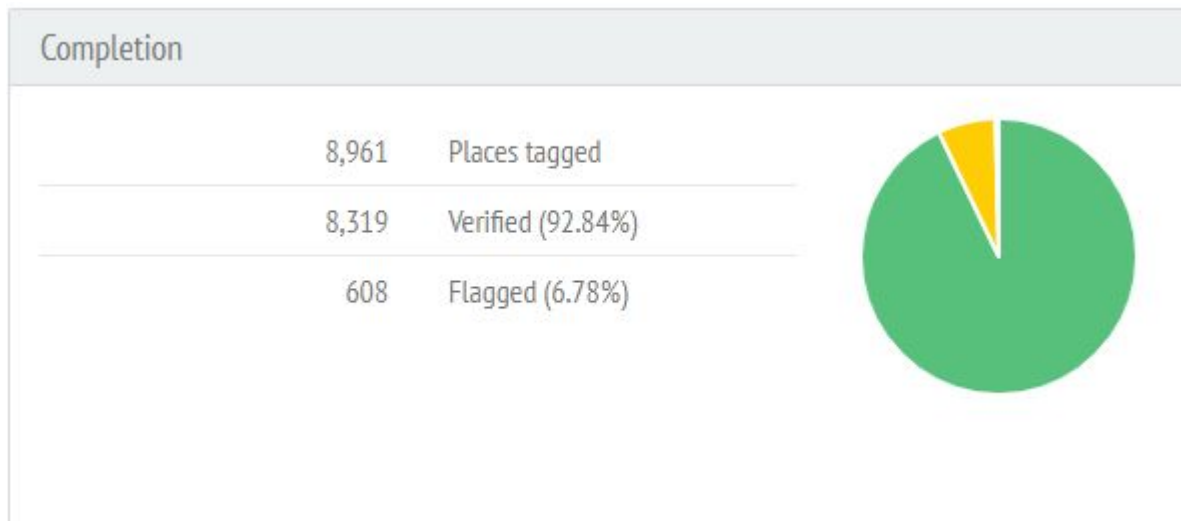


# RECOGITO



## Annotation statistics

- Statistiche



# RECOGITO



## Download options

- Download delle annotazioni in vari formati

Other

IOB BETA

The text tokenized in [IOB format](#), a common format used for machine learning training data.

IOB

Spacy JSON BETA

The text, line by line, in a JSON format usable as training data input for the [Spacy](#) machine learning library. **Work in progress!**

Spacy

# RECOGITO



## Document settings → sharing

- Pubblicazione online

### Public Access

☒ off

Only people you explicitly add as collaborators can access this document.

☐ Anyone on the Web

The document is listed on your profile page, and anyone on the Internet can find it.

☐ Anyone with the Link

The document is accessible to anyone who has the link, but it is not listed on your profile page. Link to share:

<https://recogito.pelagios.org/document/mk01xus85afugo>

Access: anyone can view the document.

Annotating requires a Recogito account. Visitors who are not logged in can view the document (if enabled), and download annotations from the download page. But they can not add annotations, even when public annotation is enabled. Blocking access to the document content will also limit some of the download options.

# VALUTAZIONE

1. Aprire `out-conll.conll`: l'annotazione delle NE è nella colonna 5
2. Aprire il Terminale (CMD, PowerShell)
3. Scrivere `cd` poi spazio poi trascinare la cartella in cui si trova il file `out-conll.conll` e infine premere Invio
4. Scrivere il seguente comando:  

```
cut -f 2,5 out-conll.conll > soloNE.conll
```

  
in questo modo estraiamo solo la seconda e la quinta colonna
5. Aprire `soloNE.conll` con un editor di testo (NO WORD)
6. Copiare il contenuto, apriamo un foglio di calcolo (Excel, Calc, Fogli) e incollarlo nel foglio di calcolo: ci devono essere due colonne
7. Annotare a mano le NE corrette nella terza colonna usando:
  - O se un token non è una NE
  - PER, ORG, LOC se è una NE
8. Copiare le tre colonne e incollarle in un editor di testo (NO WORD)

# VALUTAZIONE

9. Salvare questo file con tre colonne nella cartella “valutazione” con il nome `valutazione-news.txt`
10. Tornare sul Terminale (CMD, PowerShell)
11. Scrivere `cd` poi spazio poi trasciniamo la cartella “valutazione” e infine premiamo Invio
12. Scriviamo il seguente comando:  
`python conlleva_perl.py -r < valutazione-news.txt`
13. Dovrebbe apparire sul Terminale una scritta come questa:

```
processed 168 tokens with 28 phrases; found: 30 phrases; correct: 27.  
accuracy: 97.62%; precision: 90.00%; recall: 96.43%; FB1: 93.10  
LOC: precision: 93.75%; recall: 100.00%; FB1: 96.77 16  
ORG: precision: 85.71%; recall: 92.31%; FB1: 88.89 14
```

# VALUTAZIONE

14. Fare lo stesso procedimento per il file *valutazione-DG.txt* (Trento, 4 dicembre 1911)

*L'episodio delle contestazioni di cui è stato vittima il vescovo Endrici in visita pastorale a Bolzano suscita in De Gasperi un commento di ampia portata sul fenomeno volksbundista; considera la questione delle scuole di lingua tedesca finanziate dal Volksbund, il comportamento delle autorità bolzanine, e i rischi legati alle ascendenze protestanti di alcuni membri e capi della società pantedesca.*

15. Che differenze nella valutazione si notano?

# APPROFONDIMENTI

Tint:

<https://arxiv.org/pdf/1609.06204.pdf>

<http://ceur-ws.org/Vol-2253/paper58.pdf>

NER:

<https://web.stanford.edu/~jurafsky/slp3/18.pdf>

<https://infoscience.epfl.ch/record/277015?ln=en>

Recogito:

[https://eprints.lancs.ac.uk/id/eprint/86362/1/Linked Data Annotation Without the Pointy Brackets.pdf](https://eprints.lancs.ac.uk/id/eprint/86362/1/Linked_Data_Annotation_Without_the_Pointy_Brackets.pdf)

<https://www.modernlanguagesopen.org/articles/10.3828/mlo.v0i0.299/>

# APPROFONDIMENTI

Tool di annotazione:

<https://brat.nlplab.org/>

<https://webanno.github.io/webanno/>

Natural Language Annotation for Machine Learning (libro):

<https://www.oreilly.com/library/view/natural-language-annotation/9781449332693/>



# GRAZIE!

Email: [rachele.sprugnoli@unicatt.it](mailto:rachele.sprugnoli@unicatt.it)

Twitter: [@RSprugnoli](https://twitter.com/RSprugnoli)

