
Torino Scala Programming &
Big Data Meetup
6 maggio 2015

# Workshop
# Scala vs Java

Raffaella Ventaglio

r.ventaglio@gmail.com

https://github.com/vraffy

http://www.celi.it

# whoami

- appassionata di NLP, Information Extraction, Sentiment Analysis, Machine Learning
- con una spiccata allergia alla scrittura di documenti di qualunque genere
- ma sempre a caccia di qualche complicato problema da risolvere
- e da qualche anno... MOOC dipendente :)

# Cosa NON è questo Workshop

- un corso di introduzione a Java
- un corso di introduzione a Scala
- un corso di introduzione a NLP

# Cosa è allora?

un tentativo di mostrare un utilizzo di concetti funzionali di base in un contesto “reale”

un confronto tra due approcci diversi per risolvere uno stesso problema

# NLP for dummies

# Perché?

- informazioni strutturate
  - DB, DWH, Analytics
- la “vecchia” Internet: blog, forum, ...
  - passaparola, marketing virale
- arrivano i Social Network: twitter, Facebook, ...
  - Customer Care, Real Time Marketing
- testo non strutturato:
  - come estrarre “informazioni”?

# Come?

- analisi di base
  - Tokenizzazione, riconoscimento di frasi
- riconoscimento di “entità”
  - luoghi, persone, marchi, concetti, ...
- analisi grammaticale
  - morfologia, disambiguazione, chunking
- analisi “semantica”
  - analisi del mood, estrazione di opinioni, ...

# Analisi di base: tokenizzazione

- è facile, basta separare le parole
  - un momento: cosa è una “parola”?
- il mondo là fuori è una giungla
  - xché chiamare casseurs dei volgari vandali? .
  - #MatchForExpo grazie a tutti. Emozioni a non finire□□❤😆 .
- non esiste solo l'italiano!
  - Wahlkampfkostenrückerstattungsgesetz
    - (la legge sul rimborso delle spese della campagna elettorale!)
  - কারচুপির অভিযোগে ৩ সিটিতেই বিএনপির নির্বাচন বর্জন

# Analisi di base: separazione frasi

- questa è facile davvero: c'è la punteggiatura
  - davvero?
- in teoria sì, ma sui SN non scrive Manzoni!
  - Buongiorno e buon inizio settimana .....una domanda vorrei fare un regalo a mio marito che è rimasto come cellulare al nokia mattoncino😝....vorrei uno smartphone ma senza spendere un capitale....cosa mi consigliate?
  - RT @gayit: Luciana Littizzetto incontra Stefano #OmofobiaStop Condividete e diffondete! @chetempochefa http://t.co/OpmIf0YTLW http://t.co/8pyVnGXqvg .

# Riconoscimento di entità

- ricerca di match con liste predefinite
  - elenchi di luoghi, persone, marchi...
- algoritmi di Machine Learning
  - utilizzo di esempi noti per “apprendere” come identificare entità “sconosciute”

# Analisi grammaticale

- sì, proprio la cara vecchia analisi grammaticale
  - la – articolo femminile singolare
  - mamma – sostantivo femminile singolare
- ma poi c'è la solita “giungla”
  - pesca – sostantivo femminile singolare
  - pesca – voce del verbo pescare ind. pres. 3°p.s.
- a volte il contesto aiuta, a volte no
  - la vecchia porta la sbarra

# Analisi semantica

- riconoscimento di “espressioni polari”
  - mi piace, mi fa schifo, fa cagare...
    - (eh sì, la gente scrive di tutto)
  - non posso dire che mi dispiaccia :O
- d'accordo, ho un'opinione positiva, ma per chi?
  - interazione con le entità riconosciute
  - riconoscimento del “target” (o dei target)
    - preferisco xyz a zyx
- di che umore siamo oggi?
  - analisi del “mood” complessivo di un documento anziché delle singole opinioni

# Domande?

# References

- https://github.com/vraffy/scalaVsJavaMeetup
- Functional Programming in Scala
  http://www.manning.com/bjarnason/
- Java 8 in Action
  http://www.manning.com/urma/
- Coursera: corsi su Scala, Reactive Programming, Big Data, NLP … nutrizione equina ;)