Linguistica Computazionale

Tokenizzazione

Sai Tokenizzare (**~contare** :-))

iniziamo giocando … poi lavoriamo

- Quanti token (~parole) nella frase

C’era una volta un pezzo di legno.

Sai Tokenizzare (**~contare** :-))

iniziamo giocando … poi lavoriamo

- Quanti token (~parole) nella frase

C’era una volta un pezzo di legno.

C’era | una | volta | un | pezzo | di | legno.

C’ | era | una | volta | un | pezzo | di | legno | .

Preparazione del testo

- I testi digitali possono contenere varie forme di "rumore"
  - errori di conversione
    - caratteri spuri, ecc.
  - errori nella digitalizzazione
  - codici di markup
- Vari aspetti del testo legati alla sua fruizione umana, possono ostacolarne l'elaborazione computazionale
  - convenzioni ortografiche
  - diversi sistemi di scrittura

Preparazione del testo

Fase preliminare in cui il testo viene reso compatibile con il formato richiesto dagli strumenti di analisi computazionale

Il testo nel computer

- Organizzazione logica di un testo
  - lettere
  - parole
  - frasi
  - paragrafi
- Organizzazione fisica di un testo in formato “machine-readable”
  - caratteri
  - sequenze di caratteri
  - righe

Problema

Le due modalità di organizzazione sono ortogonali e non direttamente compatibili

Analizzare un testo

- Per analizzare computazionalmente un testo è necessario insegnare al computer a riconoscere gli elementi che lo compongono
  - quante parole?
  - quante frasi?
  - ma …
    - *cosa è una parola? cosa è una frase?*
- La nozione di parola per un computer è molto diversa da quella per gli esseri umani
  - basata sulla conoscenza dell'organizzazione linguistica (morfologica, semantica, ecc.)
    - questione ancora più complessa nel parlato
  - l'informazione grafica non è sempre sufficiente

Tokenizzazione

- Passo preliminare di qualsiasi elaborazione computazionale del testo è la sua tokenizzazione
- Tokenizzare un testo significa dividere le sequenze di caratteri in unità minime di analisi dette “token”
  - parole, punteggiatura, date, numeri, sigle, ecc.
  - i token possono essere anche entità strutturalmente complesse (es. date), ma sono comunque assunte come unità di base per i successivi livelli di elaborazione (morfologico, sintattico ecc.)
- A seconda del tipo di lingua e sistema di scrittura può essere un task estremamente complesso
  - in lingue dove i confini di parola non sono marcati esplicitamente nella scrittura la tokenizzazione viene anche chiamata word segmentation

Token

- Quanti tokens ci sono in questo testo?

  Dopo essere sceso, l'uomo si allontanò.

  - 6 tokens
    - se consideriamo gli spazi come delimitatori dei tokens
  - … in realtà 9 tokens
    - se consideriamo i tokens come unità linguistiche (es. parole, punteggiatura, ecc.)
- I token di un testo NON corrispondono alle sequenze di caratteri delimitate da spazi:
  - convenzioni grafiche
  - organizzazione lessicale e morfologica del linguaggio, ecc.
- La nozione di token è distinta da quella di parola
  - la tokenizzazione non si basa generalmente su criteri morfosintattici o semantici
    - *mandarglielo* = 1token ma… 3 parole morfologiche (*mandare* + *gli* + *lo*)

Token
alcuni problemi

- Non ci sono spazi che separano le parole dalla punteggiatura che le segue (precede):
  - **… gli ispettori, …**
  - **… “Dobbiamo aumentare gli ispettori…**
  - **gli elementi principali (azoto e ossigeno) …**
  - **… il presidente francese a “Time”: …**
- Ci sono sequenze di caratteri non separati da spazi che corrispondono a 2 token:
  - apostrofo
    - **dell’uomo**
    - **c’è**
  - parole composte con il trattino (“-”) o con il “/”:
    - **la linea Firenze-Pisa …**

Punteggiatura e ambiguità

- La punteggiatura deve essere considerata come tokens separati

  ma …

- La punteggiatura è ambigua!!!
- Il carattere ' (apice) ha vari usi:
  - apostrofo
    - **`dell'uomo`** (2 token)
  - accento
    - **`c'e'`** (2 token)
  - virgoletta
    - **`'token'`** (3 token)

Punteggiatura e ambiguità

- Il carattere . (punto) ha vari usi:
  - punto di fine frase
    - **`Questa è una frase.`**
      - in questo caso il punto deve essere considerato un token separato
  - abbreviazioni
    - **`Sig. Rossi`**
  - acronimi
    - **`U.S.A.`**
  - separatore di cifre decimali
    - **`9.70`**
  - data
    - **`25.02.2003`**
  - indirizzi WWW o e-mail
    - **`www.unipi.it`**
- Correlato alla tokenizzazione è il sentence splitting
    - segmentazione del testo in frasi
      - presuppone la disambiguazione dei diversi usi del "."

Token graficamente complessi

- Sequenze di caratteri separate da spazi possono formare un solo token:
  - nomi propri composti
    - **New York**
    - **Los Angeles**
    - **Reggio Emilia**
      - **la città di La Spezia** (4 token)
      - **la spezia che viene usata** (5 token)
  - espressioni polirematiche (multiword expressions)
    - **al di là** (1 token)
    - **ad hoc** (1 token)
  - date e ore
    - **18 giugno 1815** (1 token)
  - prezzi
    - **20 euro** (1 token)
    - **20 euro e 30 centesimi** (1 token)

Maiuscole e minuscole

- Molti programmi sono "case sensitive"
  - "treno" e "Treno" sono considerate due parole diverse
- Ambiguità nell'uso della maiuscola:
  - nomi propri
    - **`Carlo Azeglio Ciampi`**
  - inizio frase
    - **`La macchina non partiva.`**
  - enfasi e titoli
    - **`… ma ATTENZIONE …`**
    - **`NUOVA STRAGE DEL TERRORE`**
- Annullare la distinzione tra maiuscole e minuscole non è sempre la soluzione ottimale
  - le maiuscole possono avere valore discriminativo e semantico
    - cf. **`NATO`** vs. **`nato`**, **`USA`** vs. **`usa`**, **`Agnelli`** vs. **`agnelli`**

Tokenizzazione
criteri generali

- Per tokenizzare un testo è necessario stabilire i criteri di identificazione delle unità di analisi
  - criteri linguistici, euristici, pratici, ecc.
    - es. *le date sono token? gli indirizzi formano un token unico?*
- La tokenizzazione è strettamente dipendente da
  - il linguaggio e sistema grafico del testo
    - diversi modi di scrivere i numeri
      - 123.725,43 italiano
      - 123,725.43 inglese
      - 123 725,43 francese
    - diverse strategie per la creazione dei composti
    - dinamiche linguistiche
      - l'uomo => l' + uomo po' (elisione) => po'
      - doesn't => does + n't
      - they're => they + 're 'ndrangheta (aferesi) => 'ndrangheta
  - gli scopi di elaborazione computazionali
- Uno dei problemi maggiori è rappresentato dallo spazio di variabilità delle convenzioni ortografiche
  - cf. **`database`**, **`data base`**, **`data-base`**

Tokenizzazione

L’aereo per gli U.S.A. atterra a New York.

```
L’aereo
per
gli
U.S.A
.
atterra
a
New
York.
```

NO!!

SÌ!!

```
L’
aereo
per
gli
U.S.A.
atterra
a
New York
.
```

Tokenizzatori

- Moduli software per la preparazione e tokenizzazione del testo tramite
  - mini-grammatiche, che specificano le forme in cui possono comparire i token
    - es. le date:
      25/02/1993
      25-02-1993
      25 febbraio 1993
      febbraio 1993
      25 febbraio
      25 feb. 1993
  - repertori e glossari
    - acronimi, nomi propri, abbreviazioni, ecc.
- Generalmente basati su espressioni regolari
  - implementati in Perl, Python, ecc.

Tokenizzatori

- Output in formato testo
  - aggiungere separatori
    - **dell'uomo > dell' uomo**
    - **'uomo' > ' uomo '**
  - unire token separati
    - **La Spezia > La_Spezia**
    - **25 febbraio 2003 > 25_febbraio_2003**
  - normalizzare le maiuscole
    - **La macchina è guasta > la macchina è guasta**
- Output XML
  - i token sono delimitati da elementi XML
    - **<t n="1">dopo</t><t n="2">essere</t><t n="3">sceso</t><t n="4">,</t><t n="5">l'</t><t n="6">uomo</t><t n="7">si</t><t n="8">allontanò</t><t n="9">.</t>**