Micro-dispensa

18.11.2013

Sito web del corso

http://www.unipa.it/paolo.monella/labinformatica

String matching

string = sequenza di caratteri digitali (“abc”; “ino”; “andus”).
“string matching”

I limiti dello string matching

Quel ramo del <place ref=”barrington/203”>Lago di Como</place> che volge a mezzogiorno tra due catene non interrotte di monti.

Camminammo fino al <place ref=”barrington/203”>lago</place>.

Il <w type="n">governo</w> <name>Monti</name> si è occupato del <place>Mezzogiorno d'Italia</place>.

Io <w type="v">governo</w> l'Italia anche se è impossibile.

<w type="n">
<w type="v">

XQuery
/name/”monti”

site:www.unipa.it monella
lang:eng monella

Markup di nomi propri: esempio www.emilydickinson.org

Markup di nomi di luoghi

TEI:
<place>Romae</place> sum

Esempi:

- Hestia, progetto su Erodoto (sito)
- GapVis (descrizione; sito vero e proprio)

Indici e concordanze

Abelardo: http://www.intratext.com/IXT/LAT0428/_P1.HTM

Lemmatizzazione

Ammiano Marcellino: http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3atext%3a2007.01.0081

- Corpora non lemmatizzati
- Corpora lemmatizzati

Lemmatizzazione tramite markup:
<w lemma=”vis”>Vis</w> animi te impellit, si <w lemma=”volo”>vis</w>, ut hoc facias

Corpora lemmatizzati e non

Non lemmatizzati:

- PHI5.3 (latino) e...
- TLG versione “E” (Cd-rom): entrambi non lemmatizzati, però Diogenes permette di fare ricerche lemmatizzate
- Intratext

Lemmatizzati:

- Perseus

Tokenization

Distinzione del testo in parole (token)

- token (istanza)
- type (categoria)

“I **cani** mi hanno morso il polpaccio, porco **cane**!”

Due token diversi di due type diversi: uno del type “cani”, l’altro del type “cane”
(parola flessa)

Due token diversi dello stesso type “cane” (lemma)

Applicazioni in ambito storico-artistico ed archeologico

GIS - Geographical Information System
GIS di scavo