Microdispensa

Laboratorio di informatica specialistica per Scienze dell'Antichità, A.A. 2012/13, 2. semestre, Dr. Paolo Monella, lezione del 22/03/2013.

Indici e concordanze

KWIC KeyWord In Context
Esempio (IntraText):

```
1  1,  1| facilis. Patrem autem habebam   litteris aliquantulum imbutum antequam

2  1,  1|    quoscumque filios haberet,   litteris antequam armis instrui disponeret.
```

KWOC KeyWord Out (of) Context
Esempio (inventato):

```
1  1,  1| litteris

2  1,  1| litteris
```

String matching

Confronto tra stringhe (cercare una stringa).

Struttura generale di un documento XML/TEI

- TEI.2
  - teiHeader
  - text
    - front
    - body
    - back

Tokenization

Procedimento tramite il quale si distinguono le 'parole' in un testo digitalizzato. 'Parola', in questo contesto, una stringa delimitata da due 'whitespace'.

Whitespace

Un carattere digitale costituito da uno spazio, un segno di interpunzione o un'andata a capo.

Lemmatizzazione

Procedimento tramite il quale ogni token (parola) di un testo digitalizzato viene ricondotto ad un lemma.

Lemmatizzazione in TEI

<w lemma="rex">Rex</w>
<w lemma="rex">regum</w>

Type/token

Vd. Platone.

Type: l’entità astratta. Token: una sua istanza.
Esempio: “Ciao ciao”. Due token, ma un solo type.

Stemming

Procedimento digitale tramite cui, partendo da una stringa (token), si ottiene la radice della parola (stem).

Collocations / clusters

Gruppi di parole che ricorrono spesso insieme (ad es. “bandito il concorso”).