A1. Codifica e processazione di un testo: istruzioni

Linguistica e Filologia Digitale (Simone Ciccolone)

a.a. 2019/2020

Questa **prima attività laboratoriale** è **OBBLIGATORIA per tutti gli studenti frequentanti** (o che intendono sostenere l'esame da frequentanti, ovvero senza l'integrazione nel programma d'esame di una delle letture di approfondimento proposte).

Quest'attività è **propedeutica alle due attività laboratoriali successive (a scelta)** che verranno proposte nel corso della seconda metà delle lezioni:

- **A2. Analisi dello spettro di frequenze di un testo**
- **A3. Codifica TEI-XML di un testo (o parte di un testo)**

Entrambe queste attività laboratoriali richiedono di partire da un **testo ripulito**, con un livello di codifica zero coerente e affidabile, e **tokenizzato**, ovvero segmentato per unità pertinenti (frasi, o ancor meglio parole).

Lo scopo di questa attività laboratoriale è quindi quello di concentrarsi su queste **fasi preliminari di codifica e processazione di un testo**. Saltando la "fase 0" di effettiva digitazione del testo, l'attività laboratoriale si concentrerà sulle seguenti fasi:

1. individuazione del testo da processare
2. controllo e ripulitura del livello di codifica zero
3. tokenizzazione
4. processazione/etichettatura del testo tokenizzato

Modalità di consegna

Gli studenti sono invitati a consegnare un **elaborato scritto con il resoconto sintetico** (anche in forma di appunti) dello svolgimento dell'attività laboratoriale (come descritta di seguito), **insieme ai materiali elaborati nel corso delle diverse fasi dell'attività**. Il resoconto dovrà contenere almeno un brevissimo commento sulle scelte e le procedure adottate, nonché sugli aspetti di maggiore criticità incontrati.

I materiali, preferibilmente in formato elettronico, potranno essere consegnati:

- **tramite la piattaforma di e-learning (elearning.unica.it)**, caricando i file nell'attività **"A1. Consegna del resoconto dell'attività"**;
- per **e-mail a simone.ciccolone@unica.it**;
- durante gli **orari di ricevimento**;
- a inizio o fine lezione.

Risultato atteso

Quest'attività laboratoriale può essere svolta in **due modi differenti**, applicando una serie di procedure molto simili che, in parallelo, permettono di ottenere **due diversi risultati finali**:

- **a. Lista di parole**: una lista di tutte le parole presenti nel testo (per l'analisi delle frequenze)
- **b. Testo etichettato**: una versione ricodificata ed etichettata della struttura del testo

Nel primo caso, dovremo ottenere un testo emendato da tutte le sue parti da non conteggiare come parole: punteggiatura, spazi, varianti grafiche (maiuscole, enfasi), ma anche numeri di capitoli e indici. Eventualmente anche righe di sommario o didascalie possono essere eliminate, se ritenuto opportuno. Sta alle scelte dello studioso (o alle indicazioni del gruppo di lavoro) scegliere cosa escludere e cosa preservare all'interno del documento da processare; l'importante è che tale scelta sia motivata e adottata in modo sistematico e coerente.

Nel secondo caso, piuttosto che eliminare le parti "superflue" sarà necessario circoscriverle e separarle in modo significativo, per permettere una processazione semi-automatica dell'etichettatura. Lo scopo sarà quello di avere un testo etichettato e delimitato sulla base delle sue diverse sezioni (capitoli, titoli, didascalie, paragrafi), ma rendere allo stesso tempo possibile un'etichettatura semi-automatica delle frasi (ed eventualmente delle parole), che altrimenti sarebbe molto costosa in termini di tempo se fatta in modo totalmente manuale.

Ciò che quindi verrebbe eliminato nel primo flusso di lavoro (la lista di parole), dev'essere invece trattato a parte nel secondo flusso di lavoro (codifica ed etichettatura della struttura del testo); in entrambi i casi, si dovranno **operare scelte molto simili** su cosa includere, cosa escludere e come; si dovrà **ricorrere a regular expressions** che permettano l'individuazione sistematica dei punti nel documento da trattare; si dovranno **affrontare alcuni problemi cruciali delle prime fasi di codifica e processazione digitale di testi** allo scopo di successive analisi linguistiche e filologiche.

Il **risultato finale atteso** dai due flussi di lavoro sarà:

- **a. Lista di parole:** un documento di testo composto da una serie di righe di una sola parola, corrispondente al numero di parole del testo
- **b. Testo etichettato:** un documento di testo in cui il testo integrale, o una sua porzione consistente, dovrà essere etichettato tramite tag XML che ne evidenzino e delimitino la strutturazione interna, in parti di titolo e di corpo, in capitoli, paragrafi, frasi (ed eventualmente parole)

Vista la complessità di queste procedure, non vi preoccupate se non ottenete un risultato perfetto: se preferite, provate su una porzione di testo, anche procedendo manualmente, in modo da poter passare alla fase successiva. **Potete inoltre decidere di lavorare su uno stesso testo in gruppi di 2-4 studenti, distribuendovi le diverse fasi di lavoro e discutendo insieme sulle procedure da adottare**.

Un suggerimento: ogni volta che fate una modifica sostanziale al testo, **create una nuova copia del file e numerate i file in base alle fasi di lavorazione**:

0. testo di partenza
1. testo da processare (parti di paratesto eliminate)
2. testo (ri)codificato correttamente
3. testo tokenizzato
4. lista di tokens / testo con prima fase di etichettatura
5. lista di types, frequenze / codifica TEI-XML del testo

Questa prima attività laboratoriale si concentrerà sulle **fasi 1-4 (con la fase 4 opzionale)**, ma potrete continuare a usare lo stesso sistema di riferimento per i nomi dei file nelle attività laboratoriali successive.

Le diverse fasi verranno ora descritte nel dettaglio, dando indicazioni di massima su come procedere per entrambi i percorsi previsti.

Fase 1: Testo da processare

1.1. Scaricate un testo (formato TXT) dal sito del Progetto Gutenberg (www.gutenberg.org).

1.2. Modificate il nome del file facendo precedere all'identificativo numerico (già presente, e seguito dal codice "-0") il vostro cognome. Ad es. il file con "Le avventure di Pinocchio" scaricato dal sito del Progetto Gutenberg si chiama "52484-0.txt": "52484" è l'identificativo numerico, "-0" è il prefisso che indica la codifica di livello zero. Lo rinomino in "Ciccolone 52484-0".

1.3.Eliminate le parti di paratesto (come indici, frontespizio, ma soprattutto declaratorie iniziali e finali del progetto Gutenberg). **NON sovrascrivete il file originale**, ma salvate il testo emendato in una **copia del file rinominata con "-1" finale**.

1.4. Tenete da parte i metadati, per un'eventuale annotazione successiva (titolo completo, autore, data di pubblicazione etc.).

Fase 2: Ricodifica/preparazione

2.1. Controllate la struttura del testo: sono presenti capitoli, sezioni? Come sono indicati? Sono presenti didascalie, note o altre parti di testo che fuoriescono dall'ordine lineare principale? Segnate tutte le informazioni e i dettagli che vi permetteranno di rintracciare tutti gli eventuali titoli, sottotitoli, didascalie, sommari e altre forme di segmentazione nella strutturazione del testo.

2.2. Controllate la codifica del testo: sono presenti caratteri particolari, diacritici o modificatori aggiunti ad altri grafemi? Sono presenti elementi in più lingue, o addirittura in più sistemi di scrittura? Sono presenti marche convenzionali specifiche, adottate da chi ha digitalizzato il testo, per evidenziare le porzioni di testo relative a titoli, didascalie, parti in corsivo o in enfasi? Osservate attentamente il testo (scorrendolo e soffermandovi su alcuni punti a campione) e segnate tutti gli indizi e i dettagli che vi potranno essere utili per l'eventuale correzione o ricodifica del testo.

2.3. Scegliete quali operazioni di ricodifica effettuare e in che ordine. Alcune operazioni devono essere eseguite prima di altre, per evitare modifiche a catena. Ad es. se volete eliminare gli a capo forzati (perché tagliano su più righe una stessa frase), dovete prima trovare un modo per mantenere separate le diverse sezioni del testo (titolo del capitolo dal corpo del testo), altrimenti non sarà più possibile riconoscerle in automatico.

2.4. Già in questa fase dovete **scegliere quale risultato finale produrre**: nel caso della lista di parole, dovrete scegliere cosa eliminare e come ricodificare il testo in modo da ottenere solo le parole da conteggiare; nel caso del testo etichettato, dovrete scegliere come codificare ed etichettare il testo in base alla sua strutturazione interna. Già da questo punto potrete (o meglio, dovreste!) cominciare a utilizzare **regular expressions** per cercare di applicare in modo sistematico e omogeneo le vostre scelte di ricodifica.

Esempio: titoli dei capitoli

Formato nel testo di partenza: [numero]. [titolo] (es.: "1. Inizio")

Lista di parole: vogliamo eliminare tutti i titoli di capitoli. Testiamo una regular expression che ci permette di individuare tutte le possibili sequenze target (numeri, punto, spazio, sequenza di caratteri fino a fine riga); verifichiamo di non avere "falsi positivi" prima di fare una sostituzione automatica! Sistemiamo gli eventuali falsi positivi o, se sono tanti, cerchiamo di migliorare la regular expression. Dopodiché, facciamo la sostituzione automatica senza inserire nulla nel campo "Replace" (o inserendo un a capo: `\n`).

Testo etichettato: procediamo come per la lista di parole, ma invece di sostituire con nulla, inseriamo due tag XML all'inizio del titolo del capitolo, uno per separare i capitoli e l'altro per evidenziare il titolo: `<chapter> <title>`. Dobbiamo modificare leggermente la regular expression per fare in modo di non sostituire il testo del titolo, ma di posizionarci prima del titolo. Replichiamo la query per posizionarci alla fine del titolo e inserire il tag di chiusura `</title>`. Ricordiamoci poi di inserire anche il tag di chiusura del capitolo `</chapter>`.

Fase 3: Tokenizzazione

3.1. Scegliete come individuare tutti i punti di "confine" dei segmenti da tokenizzare. Come trattare la punteggiatura? Come trattare l'apostrofo?

3.2. Provate a segmentare in frasi (opzionale). Seguite le indicazioni nell'**esercitazione 03C**.

3.3. Procedete alla tokenizzazione per parola.

Nel caso della **lista di parole**, potete semplicemente eliminare tutti gli elementi che non vi interessano e sostituirli con un a capo (`\n`). Se avete deciso di mantenere la punteggiatura (per conteggiarla in quanto elemento strutturale del testo scritto), scegliete come processarla (attenzione ad es. alle sequenze di caratteri di punteggiatura! sono "parole" autonome o solo giustapposizioni di segni da tenere distinti?).

Nel caso del **testo etichettato**, fate attenzione alla possibile sovrapposizione con i livelli gerarchici superiori: frase, paragrafo, capitolo (oppure verso, strofa, componimento). Per questo sottotipo di attività, data la complessità di questa procedura, potete decidere di lasciar perdere la segmentazione ed etichettatura per parola, dando priorità alla segmentazione ed etichettatura della struttura del testo (paragrafi, strofe etc.) e delle frasi (o dei versi).

Fase 4: (Preparazione alla) processazione successiva

Lista di parole

4.1. Create una lista di parole ordinata alfabeticamente. Usate le funzioni di SublimeText.

4.2. Create una lista dei types, eliminando i duplicati di forme identiche. Scegliete se usare SublimeText o un software di elaborazione di fogli elettronici (Excel, Calc, Google Sheets).

4.3. Confrontate le due liste (opzionale): quante occorrenze (tokens) ci sono in totale nel testo? quanti tipi (types) diversi? qual è la parola più frequente?

4.4. Avviate la lemmatizzazione (opzionale). Usate la lista dei types per aggiungere le forme base dei rispettivi lemmi (cfr. sezione 5 dell'**esercitazione 03C**).

Testo etichettato

4.1. Inserite/verificate l'etichettatura della struttura del testo. Se avete inserito dei tag tramite regular expressions, verificate che siano nel punto giusto e che siano correttamente chiusi. Verificate che le diverse parti del testo siano ben evidenziate, e che l'organizzazione dei tag rispecchi la strutturazione gerarchica del testo (capitolo, titolo del capitolo, corpo del capitolo, paragrafi etc.).

4.2. Aggiungete gli attributi alle etichette XML inserite (opzionale). Potete inserire attributi con informazioni testuali sulle frasi o i paragrafi, oppure aggiungere gli attributi di lemma o POS alle parole.

Simone Ciccolone (30/10/2019)