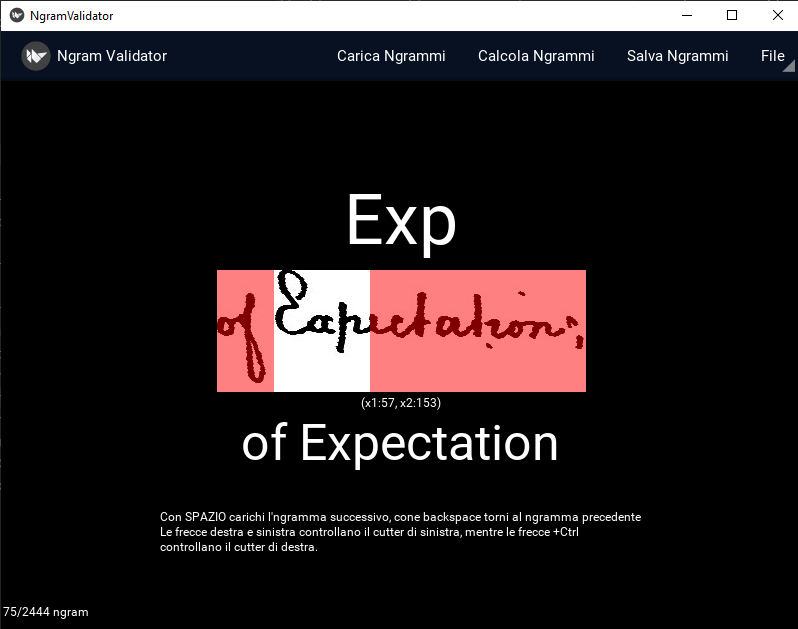
# Interfaccia per Validazione N-grammi

Il codice è disponibile su github:

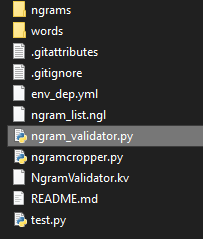
<https://github.com/ZeePPe/Ngram-Validator-Tool>



L’interfaccia permette di estrarre gli N-Grammi dall’output dell’algoritmo di allineamento MiM, di visualizzarli tutti insieme all parola dalla quale sono stati ritagliati e di correggerli.

## Operazioni Preliminari

L’interfaccia non permette di scegliere una cartella di lavoro. Bisogna allora, prima di utilizzarla, preparare la cartella contenente il file di avvio ‘*ngram\_validator.py*’. Questo file permette di avviare l’interfaccia e deve essere contenuto una cartella contenente anche le cartelle “ngrams” e “words” e il file “ngram\_list.ngl”



Prima di avviare l’interfaccia bisogna copiare all’interno della cartella “words” le parole correttamente etichettate ottenute dall’applicazione dell’algoritmo di allineamento MiM.

Bisogna a questo punto assicurarsi che il file “ngram\_list.ngl” sia vuoto. Questo file mantiene una rappresentazione degli N-grammi calcolati. Se il file non è vuoto, all’avvio dell’interfaccia il programma utilizzerà questo file per costruire la struttura dati rappresentativa degli N-Grammi.

Per questo motivo, se non è la prima volta che si usa l’interfaccia, e se non sono stati modificati i file nella cartella “words”, questo file permette di mantenere lo stato del programma e si può riavviare il programma in qualsiasi momento senza perdere i progressi fatti.

**È importante fare attenzione al formato del nome dei file delle parole.**

Il formato del nome dei file da copiare nella cartella “word”, deve rispettare il seguente standard:

ID-DOCUMENTO\_NUM-RIGA\_POSIZIONE\_TRASCRIZIONE.ESTENSIONE

Ad esempio, un nome corretto potrebbe essere:

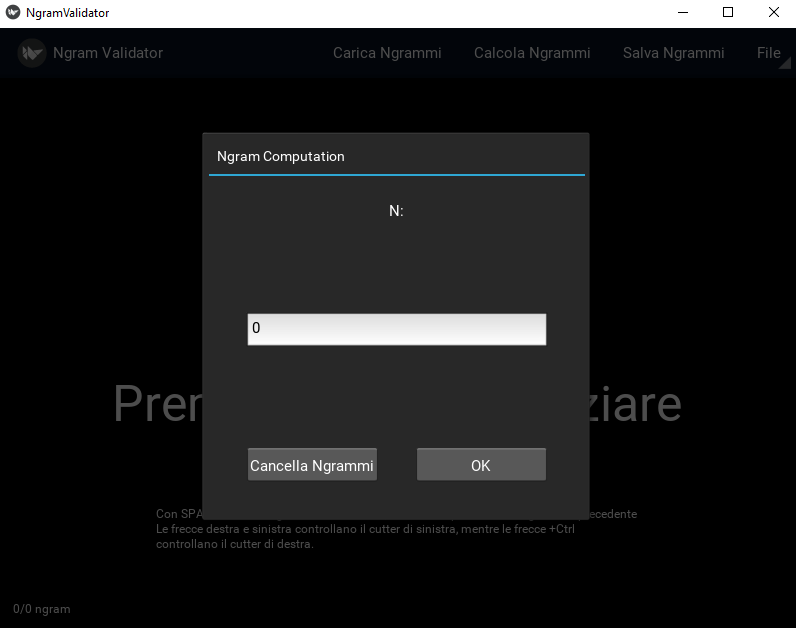
03\_04\_6\_office.png

Questo file è una immagine png della parola ‘office’ che è presente nel documento ‘03’ alla riga numero ‘04’, ed è il sesto allineamento della riga partendo da sinistra.

## Estrazione N-grammi

Una volta avviata l’interfaccia il primo passo dovrebbe essere quello di generare gli N-grammi (e popolare quindi il file “ngram\_list-ngl”.

Cliccando sul bottone “*Calcola Ngrammi*” presente nella barra in alto, si aprirà una finestra che permetterà di calcolare gli N-grammi:



Con il bottone “Cancella Ngrammi” si cancellano tutti gli n-grammi che l’interfaccia a calcolato in precedenza.

Per calcolare dei nuovi N-grammi basta indicare la lunghezza desiderata nel campo di testo (2→bigrammi, 3→trigrammi, …) e premere sul tasto “OK”. Un messaggio avviserà del completamento dell’operazione.

Per chiudere le finestre di Popup, basta cliccare al di fuori della finestra.

## Correzione e Validazione

Una volta estratti gli N-grammi, cliccando sul bottone “Carica Ngrammi”, l’interfaccia mostrerà tutti gli N-grammi estratti uno alla volta, come mostrato in figura 1. Al centro dell’interfaccia sarà mostrato l’N-gramma insieme alla parola dalla quale questo è stato estratto. Al di sopra dell’immagine sarà visualizzata la trascrizione dell’N-gramma, mentre immediatamente sotto sarà visualizzata la trascrizione dell’intera parola.

L’N-gramma è delimitato dalle due zone di taglio evidenziate in rosso. È possibile modificare le zone di taglio utilizzando le frecce della tastiera. Utilizzando quindi le frecce della tastiera “destra” e “sinistra” è possibile modificare il confine di taglio di destra. Tenendo premuto il tasto “Ctrl” mentre si pigia sulle frecce, si può invece modificare il confine di taglio di sinistra.

Una volta delineati i corretti confini di taglio, con la barra spaziatrice si conferma la correzione, e l’interfaccia mostrerà l’N-gramma successivo. Se per qualche motivo si volesse tornare indietro agli N-grammi precedenti si può utilizzare il tasto “BACKSPACE” per navigare la lista degli N-grammi al contrario.

Per eliminare l’N-gramma correntemente visualizzato, si può utilizzare il tasto “DELETE” della tastiera. ATTENZIONE! Questa operazione non è per il momento reversibile!

In basso a sinistra si può visualizzare lo stato di avanzamento, visualizzando l’indice dell’N-gramma corrente e il numero totale di N-grammi calcolati.

## Salvare le immagini degli N-Grammi

Una volta validate tutti gli N-grammi, il passo successivo è quello di generare le immagini degli N-grammi.

Cliccando sul bottone “*Salva Ngrammi*”, verranno generate tutte le immagini degli N-grammi. Per generare le immagini verranno applicati i confini di taglio definiti in fase di validazione.

**Formato del nome dei file degli N-grammi.**

Il formato del nome dei file degli N-grammi sarà simile a quello delle parole:

ID-DOCUMENTO\_NUM-RIGA\_POSIZIONE\_ID\_NGRAMMA.ESTENSIONE

Ad esempio, un nome corretto potrebbe essere:

03\_04\_1\_off.png

Questo file è una immagine png dell’N-gramma ‘off’ che è presente nel documento ‘03’ alla riga numero ‘04’, ed è il sesto secondo N-gramma della classe ‘off’ poiché il suo ID è ‘1’

(gli ID partono da 0)

# Test su 5 Pagine di Training

Ho definito un insieme di Training composto da 5 pagine delle 20 pagine selezionate dalla Bentham Collection. Le pagine che compongono l’insieme sono:

* 002\_080\_001
* 002\_541\_001
* 027\_029\_001
* 035\_320\_001
* 071\_010\_002

Ho selezionato tutti gli allineamenti corretti ottenuto da queste pagine, e ho generato gli N-grammi di lunghezza 3. Dopo la fase di validazione effettuata con l’interfaccia è stato possibile ottenere un insieme di 3-grammi composto da 2444 elementi, divisi in 742 classi (La classe si un N-gramma è definita dalla sua corretta trascrizione).

Con lo stesso insieme di pagine, l’estrazione precedente di N-grammi senza la fase di validazione effettuata con la GUI, aveva permesso di ottenere 1907 elementi divisi in 648 classi. In questo caso inoltre, alcuni N-grammi presentavano dei tagli significativi nella parte finale o iniziale dell’Ngramma, oppure riportavano elementi spuri non appartenenti all’N-gramma.

|  |  |  |
| --- | --- | --- |
|  | **WITHOUT GUI** | **WITH GUI** |
| *Classes* | 648 | 742 |
| *items* | 1907 | 2444 |

Per la fase di validazione degli N-grammi è stato impiegato un tempo pari a:

|  |  |
| --- | --- |
|  | **VALIDATION TIME** |
| *HH:mm:ss* | 02:40:34 |