Predictive-Emoji-Keyboard

Membri del gruppo: Somma Francesco 681735, Paparella Leonardo 682008
Link Repository: https://github.com/FrancescoSomma/Predictive-Emoji-Keyboards

Il progetto consiste in una piccola applicazione capace di consigliare emoji data una frase in input. Nello specifico,l'applicazione funziona con frasi in lingua inglese.

Funzionamento
L'utente inserisce una frase in lingua inglese e clicca il tasto **Invia**.
Il sistema prende in input la frase ed esegue la correzione sintattica di tutte le parole che la compongono. Dopodichè il sistema consiglia 3 emoji per ogni parola principale (quindi escluse stopwords e punteggiatura).
L'utente può cliccare sull'emoji che vuole utilizzare per inserirla all'interno della frase al posto della parola corrispondente.
Nel caso la frase contenga parole che il sistema non conosce, il sistema non effettua alcuna predizione per queste ultime ma,all'utilizzo successivo, il sistema chiederà all'utente se desidera aggiornare la conoscenza del sistema inserendo le nuove parole nel modello. In caso di risposta affermativa, partirà il training con le nuove parole.
E' inoltre possibile visualizzare graficamente il modello di clustering cliccando sul tasto **Mostra cluster**

Scelte di progettazione
Per lo sviluppo dell'applicazione è stato scelto il linguaggio Python.
Come sorgente di conoscenza è stato scelto un file .csv contenente numerose recensioni di film sul sito IMDB.
Dopodichè sono state effetuare operazioni di text processing come conversione in minuscolo,tokenizzazione, rimozione della punteggiatura e rimozione delle stopwords.
Al termine di questo processo, le frasi tokenizzate sono state utilizzate come input per creare il modello di apprendimento. E' stato utilizzato un approccio di apprendimento non supervisionato, nello specifico soft-clustering.
Per fare ciò ad ogni parola è stata applicata una trasformazione Word2Vec, che appunto trasforma ogni parola in un vettore di n dimensioni (nel nostro caso 100 dimensioni). Dopo aver addestrato il modello, il sistema è in grado di calcolare la similarità tra due parole in base alla loro distanza nel modello.
Il modello viene serializzato e salvato su file.

La fase successiva è stata creare il dataset di 100 parole con le emoji corrispondenti. Ogni emoji viene rappresentata utilizzando un apposito codice Unicode.

Infine è stata creata una interfaccia grafica con cui è possibile utilizzare l'applicazione.

La correzione sintattica delle parole avviene controllando la distanza sintattica che le stesse hanno da una lista 466000 parole della lingua inglese, contenute nel file *word.txt*
Le nuove parole inserita vengono salvato sul file *nuoveparole.txt*, il cui contenuto viene controllato ad ogni avvio.
Il grafico del modello viene creato nel momento dell'addestramento e viene salvato sul file *grafico.html*
Il tempo impiegato per l'addestramento varia da 60 a 80 secondi, mentre il tempo impiegato per la creazione del grafico è di circa 30 secondi, per un totale di 90-110 secondi circa.
La stima dell'incertezza delle emoji è stata limitata proponendo all'utente 3 emoji differenti,le più simili alla parola data

Librerie utilizzate
* *Tkinter* per la creazione dell'interfaccia grafica * *Pandas* per le operazioni di lettura da file .csv * *Nltk* per il download del pacchetto di stopwords e le operazioni di text processing * *Gensim* per la creazione del modello Word2Vec * *Levenshtein* per la funzione di correzione sintattica * *Pickle, os* e *time* per utilities * *Sklearn* e *bokeh* per la creazione del grafico