

Progetto di Linguistica Computazionale II

A.A. 2022/2023

Linee guida

Obiettivo:

Sviluppare un classificatore (o un regressore) di testi basato sui diversi approcci studiati a lezione. Confrontare i risultati dei diversi sistemi applicati ad uno specifico compito da selezionare tra questi task estratti dalla campagna di valutazione dei sistemi per l'analisi dell'italiano, chiamata EVALITA:

- Da EVALITA 2020: AcCompl-it - Acceptability & Complexity evaluation (<https://sites.google.com/view/accompl-it/home-page?pli=1>)
- Da EVALITA 2018: HaSpeeDe - Hate Speech Detection (<http://www.di.unito.it/~tutreeb/haspeede-evalita18/data.html>)
- Da EVALITA 2018: IronITA - Irony Detection in Twitter (<http://www.di.unito.it/~tutreeb/ironita-evalita18/index.html>)

I dati e la descrizione dei diversi task possono essere trovati all'interno delle pagine web elencate sopra. I task possono essere composti da diversi sotto-task, è sufficiente sceglierne uno a piacere e non affrontarli tutti.

I classificatori dovranno essere implementati in Python sfruttando le diverse librerie di machine learning introdotte a lezione.

Fasi realizzative:

Dopo aver scelto il compito da affrontare e aver recuperato i dati necessari per farlo, sviluppare i classificatori richiesti sotto, o regressori in base al compito selezionato, e condurre le seguenti analisi:

1. Sviluppare un classificatore basato su **SVM lineari** che prende in input una rappresentazione del testo basata solo su **informazioni linguistiche non lessicali** estratte utilizzando il sistema **Profiling-UD**. Riportare i seguenti risultati:
 - valutazione del sistema con un processo di **5-fold cross validation** condotto sul training set;
 - valutazione del sistema sul test set ufficiale del task;
 - elenco delle **15 feature più importanti** per la classificazione.
2. Sviluppare un classificatore basato su SVM lineari che prende in input una rappresentazione del testo basata su n-grammi di caratteri, parole e part-of-speech. Riportare i seguenti risultati:
 - testare diverse rappresentazioni del testo che variano rispetto alla lunghezza degli n-grammi utilizzati e/o rispetto al tipo di informazione utilizzata all'interno degli n-grammi (forme, lemmi, caratteri, part-of-speech) e valutare i diversi sistemi con un processo di **5-fold cross validation** condotto sul training set.
 - valutazione sul test set ufficiale del miglior sistema rispetto ai risultati ottenuti con il processo di **5-fold cross validation** del punto sopra.
3. Sviluppare un classificatore basato su SVM lineari che prende in input una rappresentazione del testo costruita attraverso l'uso dei word embedding (<http://www.italianlp.it/resources/italian-word-embeddings/>). Riportare i seguenti risultati:
 - testare diverse rappresentazioni del testo che variano rispetto al modo di combinare gli embedding delle singole parole e/o rispetto alle categorie grammaticali delle parole

prese in considerazione. Valutare i diversi sistemi con un processo di *5-fold cross validation* condotto sul training set.

- valutazione sul test set ufficiale del miglior sistema rispetto ai risultati ottenuti con il processo di *5-fold cross validation* del punto sopra .
1. Dopo aver scelto un Neural Language Model tra quelli visti a lezione, condurre un processo di fine-tuning per 5 epoche. Riportare i seguenti risultati:
 - riportare le curve di loss di training e di validation;
 - per ogni epoca valutare il sistema sul validation set;
 - alla fine dell'ultima epoca, riportare la valutazione del sistema sul test set ufficiale.

Risultati del progetto:

Perché il progetto sia giudicato idoneo, devono essere consegnati:

- a. i corpora utilizzati per fare gli esperimenti;
- b. i programmi ben commentati scritti in Python. È accettata la consegna dei programmi sia in forma di file script in python, che in forma di file jupyter notebook;
- c. la relazione in forma di report che descriva: il compito affrontato, i sistemi sviluppati e i risultati degli esperimenti richiesti. La relazione deve avere una lunghezza approssimativa che va dalle 5 alle 10 pagine.

Date di consegna del progetto:

il progetto deve essere consegnato per posta elettronica a felice.dellorletta@ilc.cnr.it, simonetta.montemagni@ilc.cnr.it e giulia.venturi@ilc.cnr.it almeno dieci giorni prima dell'orale di ogni appello per poter essere considerato valido per l'appello.

NB: il progetto **DEVE** essere svolto individualmente.