Anno accademico 2021/2022
Università Ca'Foscari, Venezia

Project Work finale

DATI IN RIMA

TEXT MINING LETTERARIO

Gestione dei Dati Digitali

Cholevas Ioannis, Iori Pietro, Tolin Filippo

1.
Analisi preliminare

Scenario

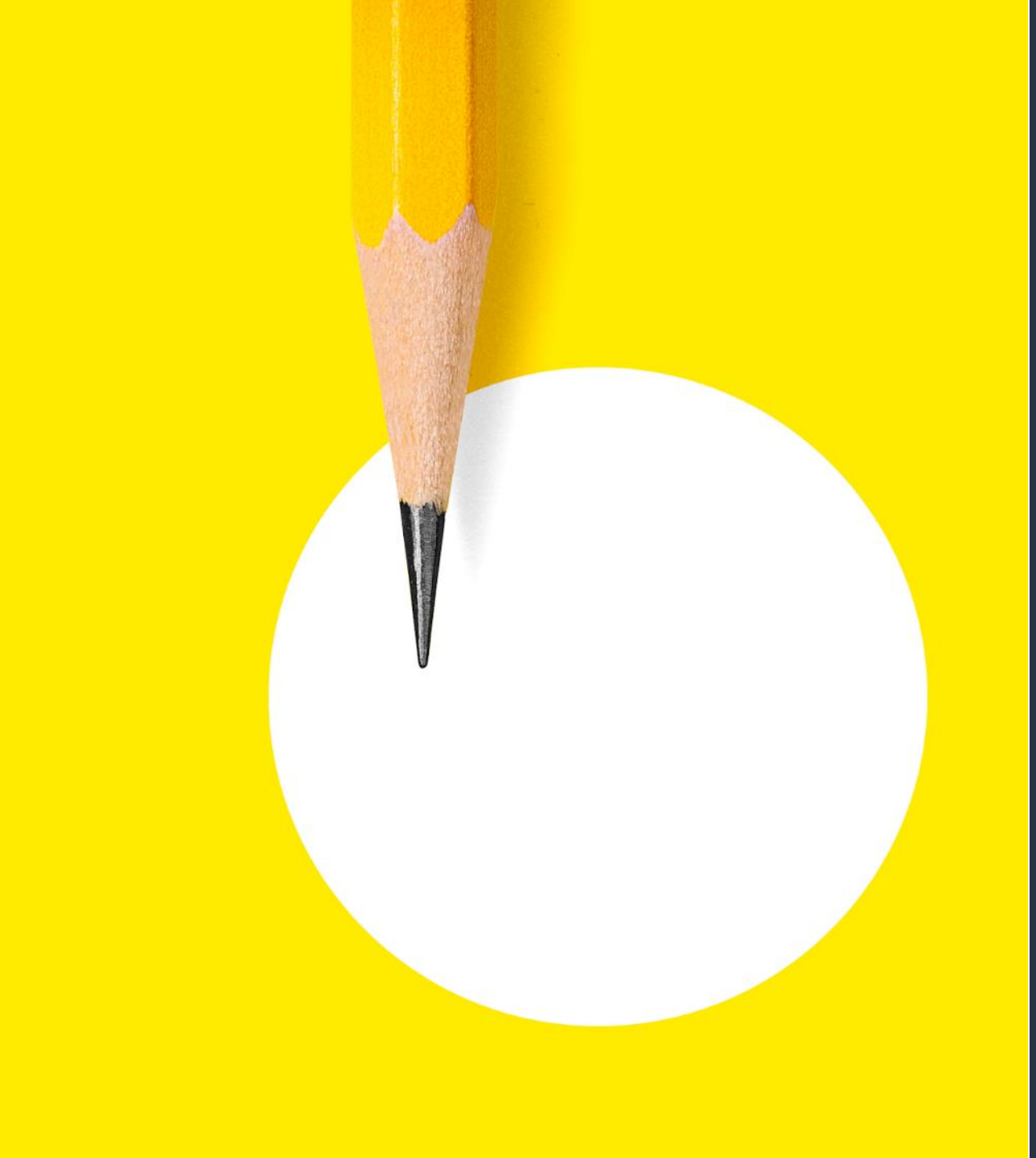
Il corpo insegnanti di una scuola superiore deve organizzare un torneo di «Indovina l'autore», una competizione letteraria nella quale agli studenti partecipanti verranno sottoposti dei testi poetici. Lo scopo della competizione è di identificarne l'autore tra i cinque possibili.

Obiettivo generale

Le finalità sono quelle di testare il campione di autori scelti: un adeguato margine di difficoltà nell'associazione poesia-autore necessita un campione di autori le cui produzioni sono sufficientemente identificabili.
Il nostro project work permette di verificare il livello di complessità nell'identificare gli autori alla luce delle loro principali produzioni.

Destinatari

Il progetto è rivolto al corpo docenti incaricato di selezionare gli autori, fungendo da strumento di supporto per testare l'idoneità e l'identificabilità degli autori scelti.



2.
Base dati

Base dati

La base dati è una selezione di poesie di cinque autori italiani:

- Alda Merini (poesie varie - aldamerini.it);
- Giovanni Pascoli (*Myricae* – wikisource.org);
- Giosuè Carducci (*Juvenilia* – wikisource.org);
- Gabriele D'Annunzio (*Alcyone* – wikisource.org);
- Giacomo Leopardi (*Canti* ed *Idilli* – wikisource.org).

Le poesie sono state selezionate attraverso un processo di webscraping effettuato con uno script python. Ogni poesia è stata salvata in un file .txt differente, preservando la struttura dei versi.

Qualità della base dati

Vista la natura dei dati in oggetto, i criteri principali utilizzati per valutare la qualità dei dati sono:

1. Tutela della struttura originale del testo poetico;
2. Affidabilità del sito utilizzato come fonte.

Le maggiori criticità riscontrate riguardano l'inesistenza di un ampio catalogo di opere di Alda Merini in *wikisource.org*, motivo per il quale si è utilizzato *aldamerini.it*.

3.
Metodo di lavoro

Modellizzazione del sistema

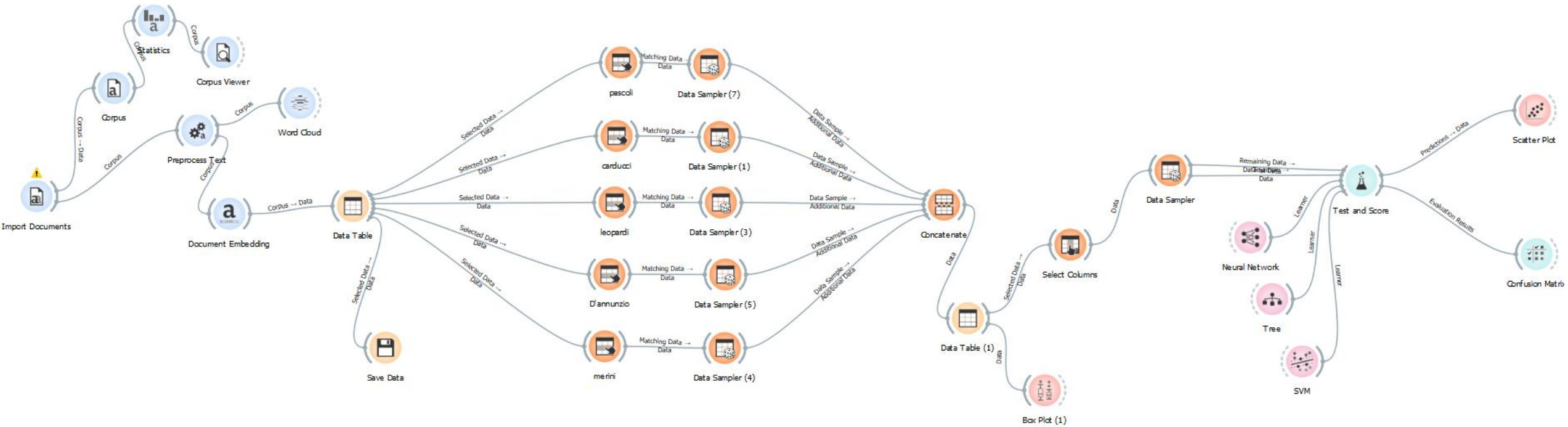
L'obiettivo del sistema è la costruzione di un modello in grado di attribuire un testo poetico ad un autore tra cinque possibili.
Il modello in esame è una rete neurale a 350 neuroni, la cui struttura reticolare ottimizza l'organizzazione dei dati.

- Input: cartella contenente poesie in file .txt divisi per autore.
- Fase di preprocess: il sistema filtra le parole non significative.
- Document embedding: i valori vengono scomposti in token e quindi trasformati in numeri.
- Campionamento riorganizzativo: per livellare lo squilibrio del numero di testi per autore, i dati sono stati campionati in maniera semi-automatica.
- Test data: l'efficienza del modello è stata allenata sull'80% dei dati disponibili, così da conservare un ampio margine per testare l'efficienza effettiva.

Problemi di Machine Learning

Il modello deve rispondere a un problema di classificazione fondato sulla text analysis.
Basandosi sul supervised learning, al modello vengono mostrati corpora differenti dall'attribuzione nota. Una volta sottopostogli un testo di uno degli autori campionati, il modello in questionepuò dirsi efficiente se è in grado di compiere la corretta attribuzione.

ARCHITETTURA DEL PROCESSO IN ORANGE



4.
Risultati

Valutazione finale

Risultati della valutazione (report di Orange):

1. Operatore "Test and Score"
2. Operatore "Confusion Matrix"
3. Operatore "Scatter Plot"

Criteri di valutazione

- Classification Accuracy
- Precision
- Recall

Report di Orange

Test and Score Sun Jul 24 22, 15:32:26

Settings

Sampling type: No sampling, test on testing data
Target class: None, show average over classes

Scores

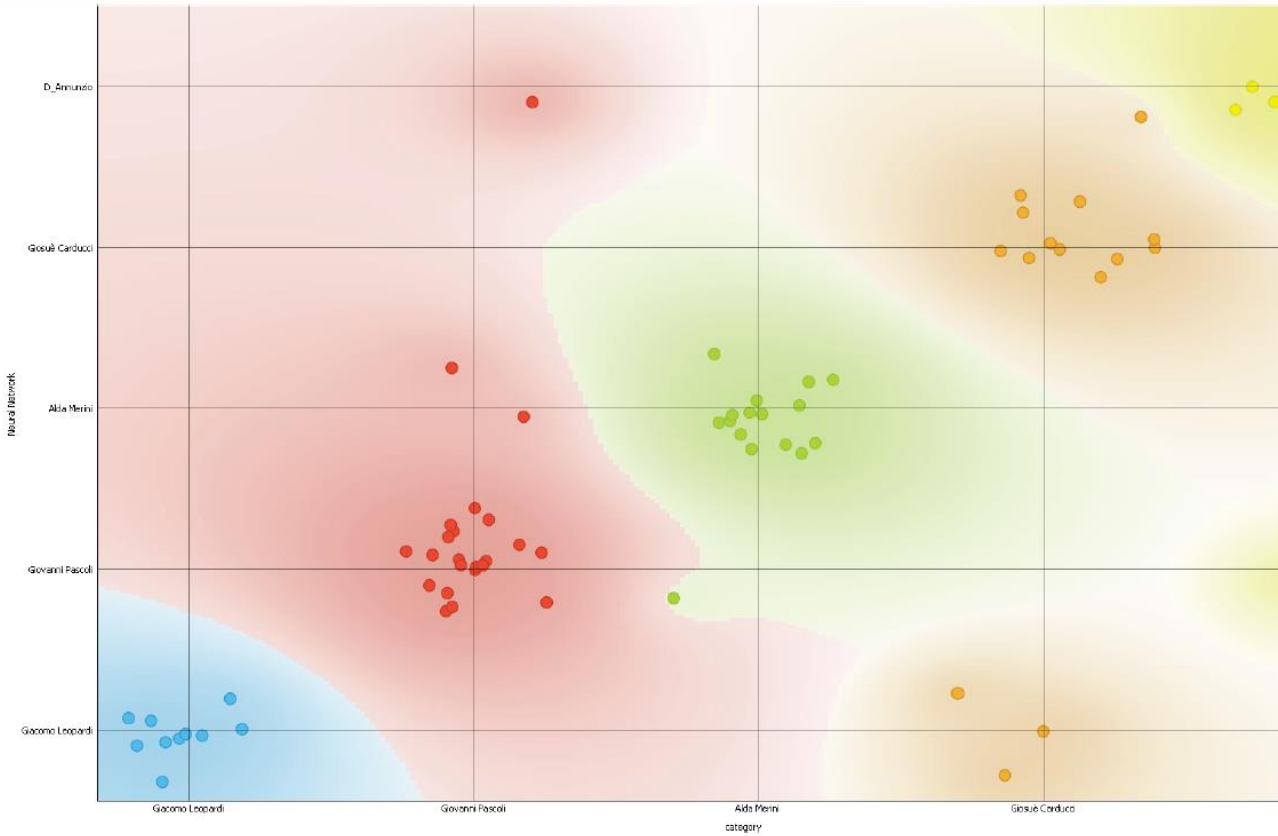
Model	AUC	CA	F1	Precision	Recall
Tree	0.7816129032258063	0.6233766233766234	0.6185412117709935	0.6216244073386931	0.6233766233766234
SVM	0.9847311827956988	0.8961038961038961	0.893466477500091	0.8980257116620753	0.8961038961038961
Neural Network	0.98	0.8831168831168831	0.8822521352165622	0.8943104260140166	0.8831168831168831

Confusion Matrix Sun Jul 24 22, 15:36:08

Confusion matrix for Neural Network (showing proportion of predicted)

		Predicted					
		Giacomo Leopardi	Giovanni Pascoli	Alda Merini	Giosuè Carducci	D_Annunzio	Σ
Actual	Giacomo Leopardi	76.9 %	0.0 %	0.0 %	0.0 %	0.0 %	10
	Giovanni Pascoli	0.0 %	90.9 %	11.8 %	0.0 %	7.1 %	23
	Alda Merini	0.0 %	4.5 %	88.2 %	0.0 %	0.0 %	16
	Giosuè Carducci	23.1 %	0.0 %	0.0 %	100.0 %	7.1 %	15
	D_Annunzio	0.0 %	4.5 %	0.0 %	0.0 %	85.7 %	13
	Σ	13	22	17	11	14	77

Scatter Plot Sun Jul 24 22, 15:42:38



Color: category, **Jittering:** 10