TEXT MINING & SEARCH

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Reviews classification

Alex Ceccotti | Stefano Fiorini | Davide Pecchia

Agenda

1. **Introduzione**
   - Obiettivo
   - Dataset
2. **Data manipulation**
3. **Metodologia**
   - Preprocessing del testo
   - Classification
4. **Risultati e Valutazioni**

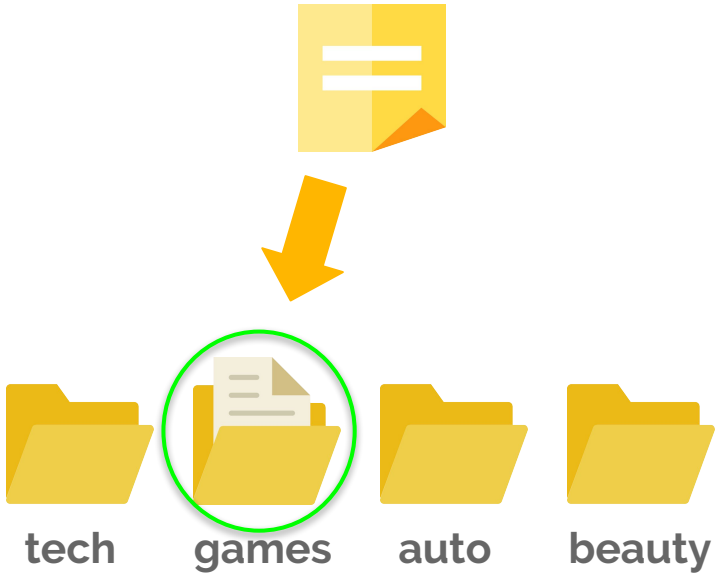
1

Introduzione

Definizione degli obiettivi e presentazione del dataset

Obiettivo

Attribuire ad una qualsiasi review pubblicata da un utente su Amazon la giusta categoria di appartenenza (ambito tech, video games, auto, beauty …)



Dataset

Il dataset è composto da una raccolta di reviews pubblicate su amazon (disponibili a questo link).

Le categorie selezionate per le analisi sono state in totale 11:

- Automotive
- Beauty
- Cell phones & accessories
- Digital music
- Grocery & gourmet food
- Office products
- Patio, lawn & garden
- Pet supplies
- Tools & home improvement
- Toys & games
- Video games

2

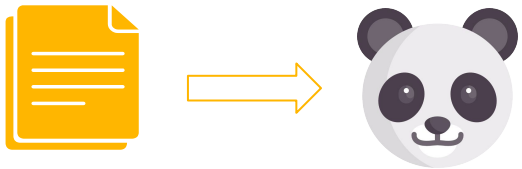
Data Manipulation

Descrizione dei processi di data manipulation

Data Manipulation

1 Creazione pandas dataframe

Essendo i dati in formato .json è stato necessario creare un dataframe pandas, per avere una migliore gestione del dato



2 Selezione variabili

Delle 11 variabili disponibili per ogni review, ne sono state selezionate 2 utili per le analisi: il testo della recensione (***reviewText***) e la categoria di appartenenza (***y***)

3

Metodologia

Preprocessing del testo e implementazione del modello di classificazione

Processo

1

Preprocessing del testo

Pulizia dei testi tramite:

- Lowercase del testo
- Eliminazione punteggiatura
- Eliminazione stopwords
- Lemmatization

2

Neural Network

Implementazione della rete neurale per la classificazione

3

Classificazione

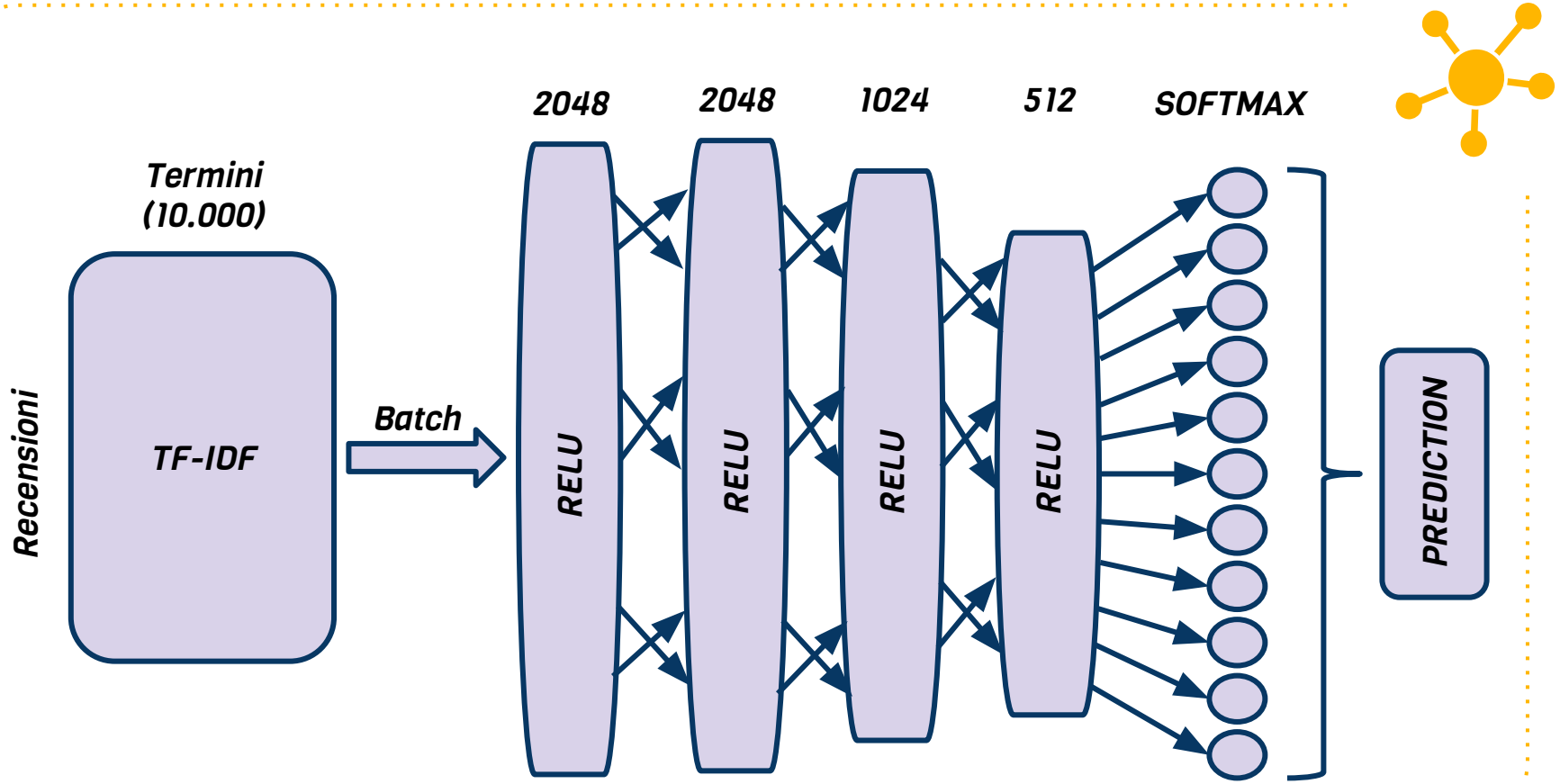
Allenamento del modello di classificazione e validazione delle performance sul test set

4

Valutazione Risultati

Valutazione dei risultati ottenuti attraverso opportune metriche:

- Accuracy
- F1 score


2048
2048
1024
512
SOFTMAX
Termini
(10.000)
Recensioni
TF-IDF
Batch
RELU
RELU
RELU
RELU
PREDICTION

4

Risultati e Valutazioni

Presentazione dei risultati ottenuti

Risultati del training

	Training set	Validation set
Epoca 1	**0.9134**	**0.9302**
Epoca 2	**0.9529**	**0.9352**
Epoca 3	**0.9817**	**0.9322**

Accuracy

0,9817
Training set

0,9322
Validation set

0,9314
Test set

F1 score
0,93

Possibili miglioramenti

1. Aumento del numero di reviews analizzate
2. Implementazione di un modello di classificazione più complesso, con un fine tuning più approfondito dei parametri

Necessaria una adeguata capacità computazionale

Grazie!