

Dataset sulla storia climatica nei quotidiani





Lo Scopo

- Studiare la **storia climatica di una città** analizzando i dati riportati all'interno dei quotidiani storici.
- Studiare **eventi climatici estremi** come alluvioni e frane eventi attraverso le narrazioni contenute nei quotidiani

CINECA



Le Fonti

- Digitalizzazioni di due importanti quotidiano storici italiani:
 - “**Il Corriere della Sera**” negli anni 1876 – 1944, **70 anni e 25.886 uscite**;
 - “**Il Popolo d’Italia**” negli anni 1914-43, per un totale di **30 anni e 9.456 uscite**.

CINECA



L'elaborazione

- A partire dalle digitalizzazioni, si è analizzato il **layout del quotidiano** identificando:
 - Titoli
 - Testi degli articoli
 - Fotografie
 - Grafici e disegni
 - Annunci
- Trascrizione tramite **OCR**
- Identificazione tramite vocabolario controllato di **parole inerenti al «clima»**

CINECA



Layout Analysis

- Lo strumento **open** allo stato dell'arte identificato, LayoutParser, seppur performante, poco si adatta all'analisi di quotidiani storici italiani.
- Si è sviluppata una **piattaforma** per aiutare gli esperti di dominio nella creazione di un dataset annotato, base che si è utilizzata per il **fine-tuning** dello strumento.

Entrambi gli strumenti sono a disposizione fra i servizi richiedibili all'interno di DPaaS:

- *Piattaforma:* **LWP Data Preparation**
- *Modello specializzato:* **lp_carlino40**

Trascrizione

- Tutti i testi sono stati trascritti tramite uno strumento **open** di OCR

➤ OCR: **TESSERACT**

The CINECA logo is rendered in a bold, white, sans-serif typeface. It is positioned in the bottom-left corner of the slide, partially overlapping the dark, stormy background image. The letters are clean and modern, with a slight shadow effect that makes them stand out against the darker background.



Il Vocabolario

- Le parole inerenti al clima sono state identificate a partire dall'elaborazione di uno specifico vocabolario.
- Vocabolario creato a partire da:
 - *Glossario metereologico ARPAL*
 - *Il Nuovo Soggettario di Firenze*

CINECA

* Per la lista completa dei vocaboli vedasi
Appendice A del manuale operativo



Dataset

- Il dataset è composto da tutti gli articoli contenenti i **vocaboli identificati come appartenenti al vocabolario** *
- Sono presenti sia **le immagini** che tutti i **dettagli all'articolo originale contenente le parole ricercate**: nome del quotidiano, data e numero di pagina, posizione in pixel) del testo
- Ogni articolo è corredato dalla sua trascrizione.

* la ricerca è stata eseguita per lemma, senza tener conto del significato anche figurato che possono avere i vocaboli se inseriti in determinate espressioni – es: pioggia di applausi.



Il Dataset in numeri

- Analizzate interamente due importanti testate nazionali:
 - “Il Popolo d’Italia” (1914-43)
 - “Il Corriere della Sera” (1876 – 1944)
- 35.342 uscite processate
- Più di 1ML di articoli contenenti le parole del vocabolario controllato

CINECA