Document Semantic Similarity

TIS Project

Alberto Pirovano Francesco Picciotti

Politecnico di Milano

28th April 2017

Outline

Introduzione

L'attuale stato dell'arte per similitudine semantica tra documenti si suddivide in:

- NLP Tradizionale
- Vector Space Model
- Deep Learning based

Outline

NLP Tradizionale

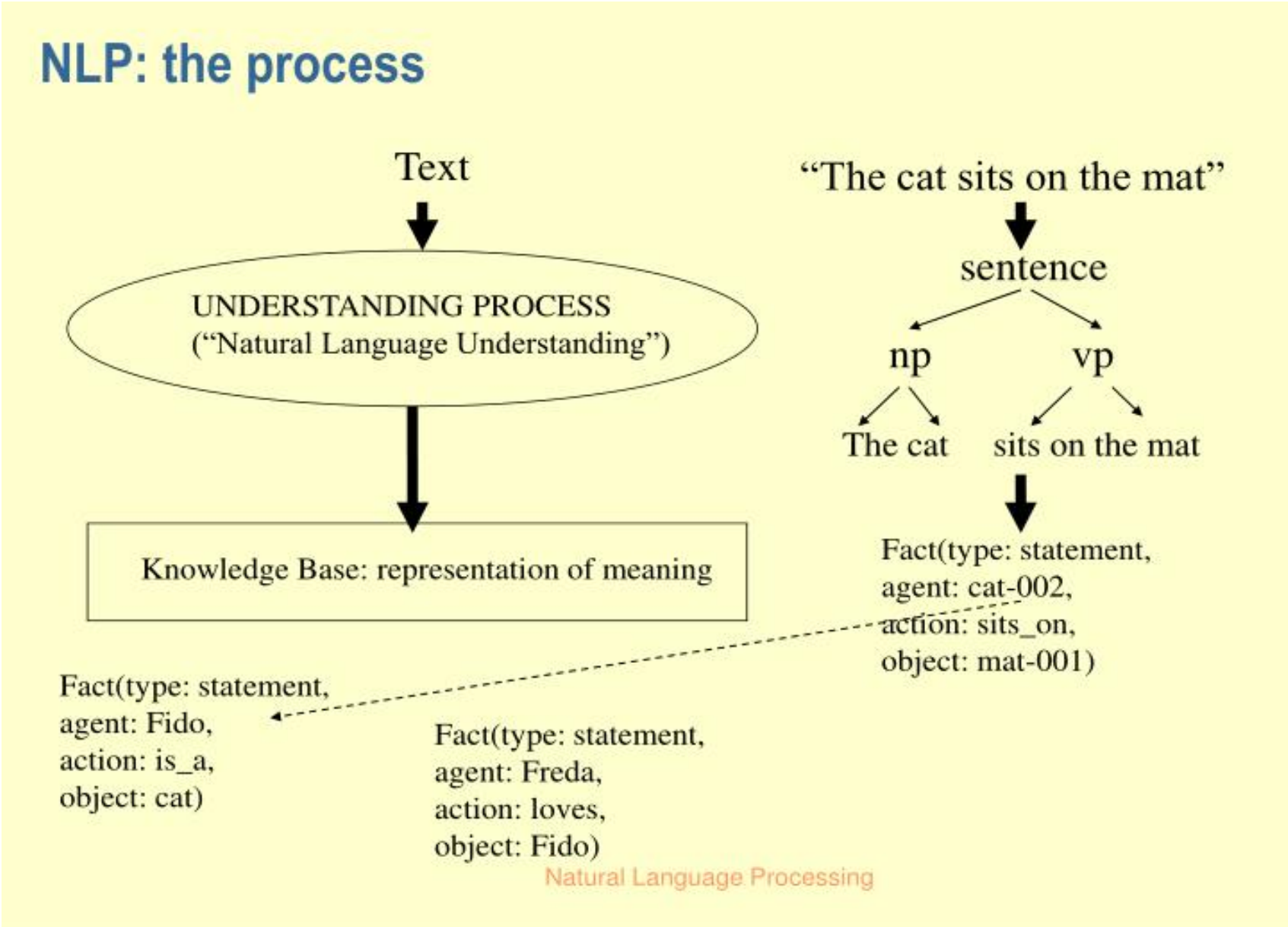
Questo approccio segue la letteratura del Natural Language Processing e consiste dei seguenti passi:

- Cleaning dei dati
- Pos-Tagging
- Stemming o Lemmatisation
- Parsing
- Ontologia

Tuttavia nel caso del nostro scopo presenta delle criticità, ovvero:

- Affidabilità del Pos-Tagger italiano di TreeTagger
- Reperire una Ontologia e un parsing toll nella lingua italiana

NLP Tradizionale

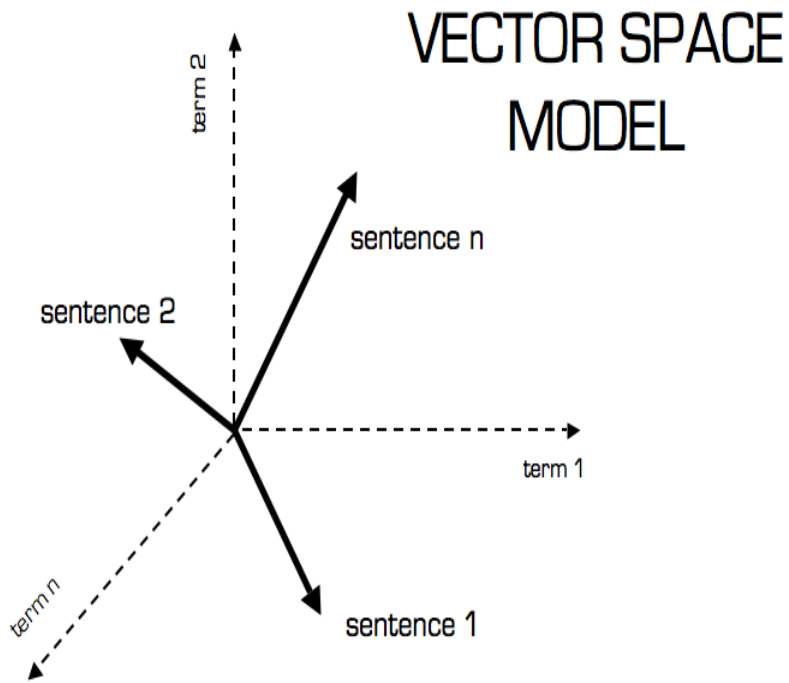


Outline

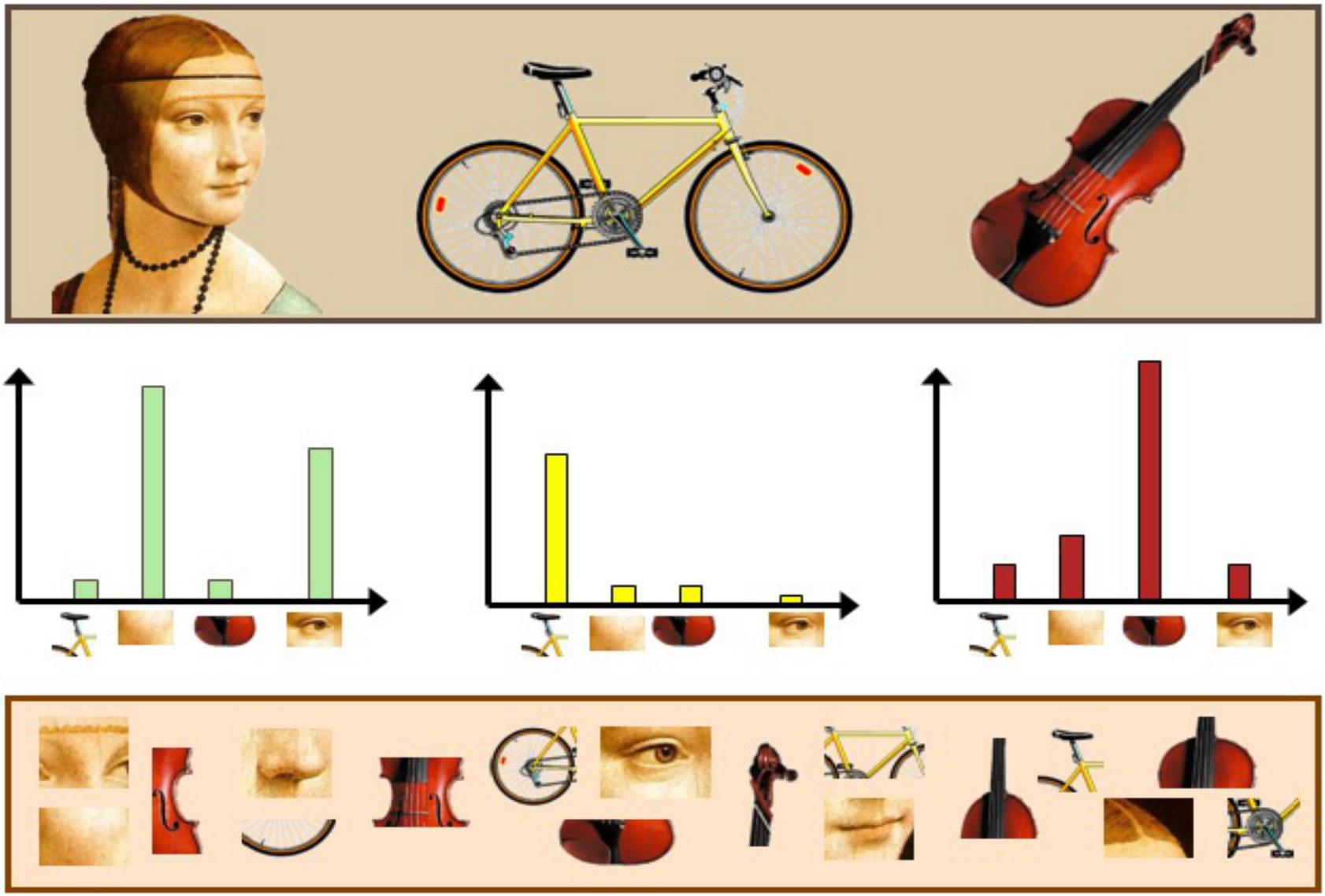
Vector Space Model

È basato sulla rappresentazione vettoriale del documento in un piano dove le diverse dimensioni sono il Bag of Word di tutti i documenti. Questo approccio richiede dei seguenti steps:

- Cleaning dei dati
- Stemming o Lemmatisation
- Encoding del documento in vettore
- TF/IDF
  + LSA (Latent Semantic Analysis)
- Similarity (Cosine, Pearson, ...)



...limiti?



Geometrico: limiti

Per riassumere:

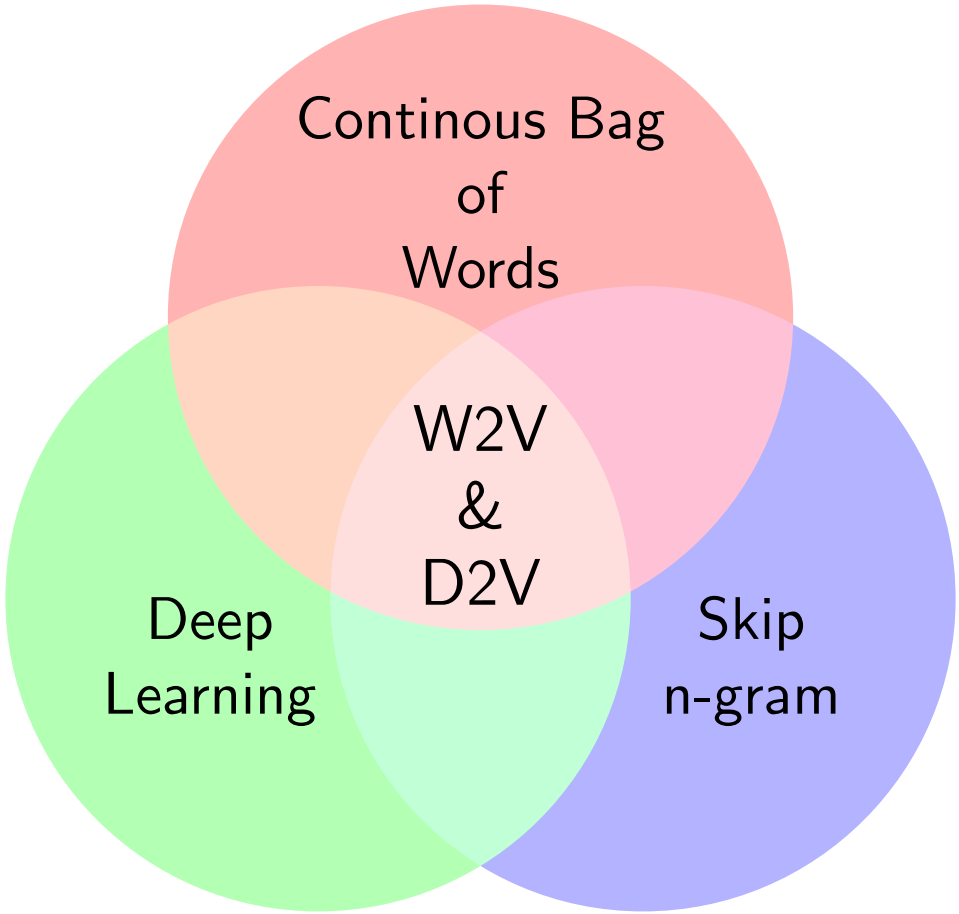
- Molto dipendente dal preprocessing del corpus
- Grande Bag Of World $\rightarrow$ bisogno di LSA, ma non è cos banale. Perché?

  "LSA assumes that words that are close in meaning will occur in similar pieces of text"

- Semantica e ambiguità
- Un documento è un insieme non ordinato di parole

Outline

E non solo



Deep Learning: W2V & D2V

Negli ultimi anni il Deep Learning trova numerose applicazioni con ottimi risultati.
La creazione di Word2Vec da Google offre un modello che ha principalmente, i seguenti vantaggi e svantaggi.

Pros:

- Molto meno dipente da un preprocessing
- Context-aware
- Combina il metodo *Geometrico* con quello *NLP Tradizionale*
- Non sfrutta una ontologia, ma la crea

Cons:

- Tecnica unsupervised
- Bisogno di un esperto per validare gradi di similarit
- Può risultare in GIGO system (Garbage In Garbage Out)

Outline

Dataset

"Preprocessing is 80% of NLP work"

Lev Konstantinovskiy

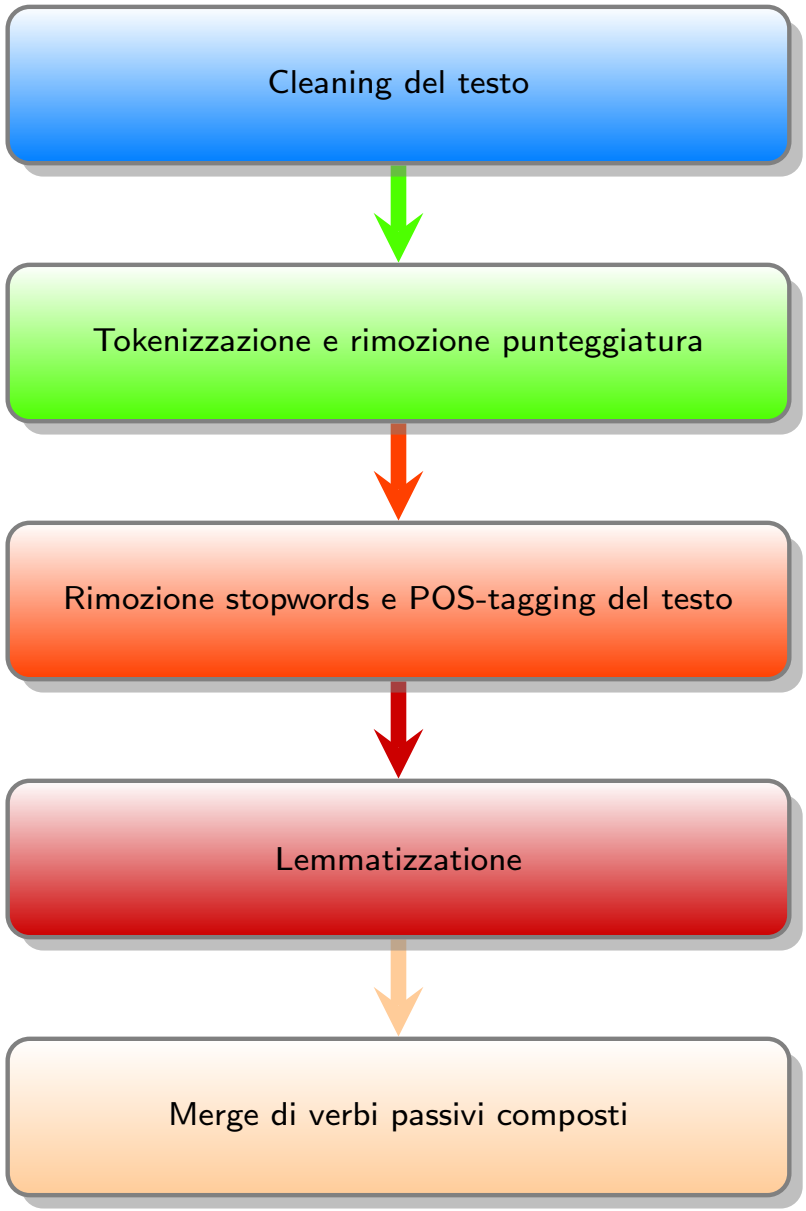
Il dataset fornitoci è composto da due corpus:

- il corpus del Sole 24 Ore con 3265 articoli, di cui 31 non hanno body
- il corpus di Radiocor con 6916 articoli

Il corpus prima del preprocessing contiene quindi 10150 articoli.
Togliendo i duplicati otteniamo 9283 articoli, cioé ci sono 867 articoli duplicati.

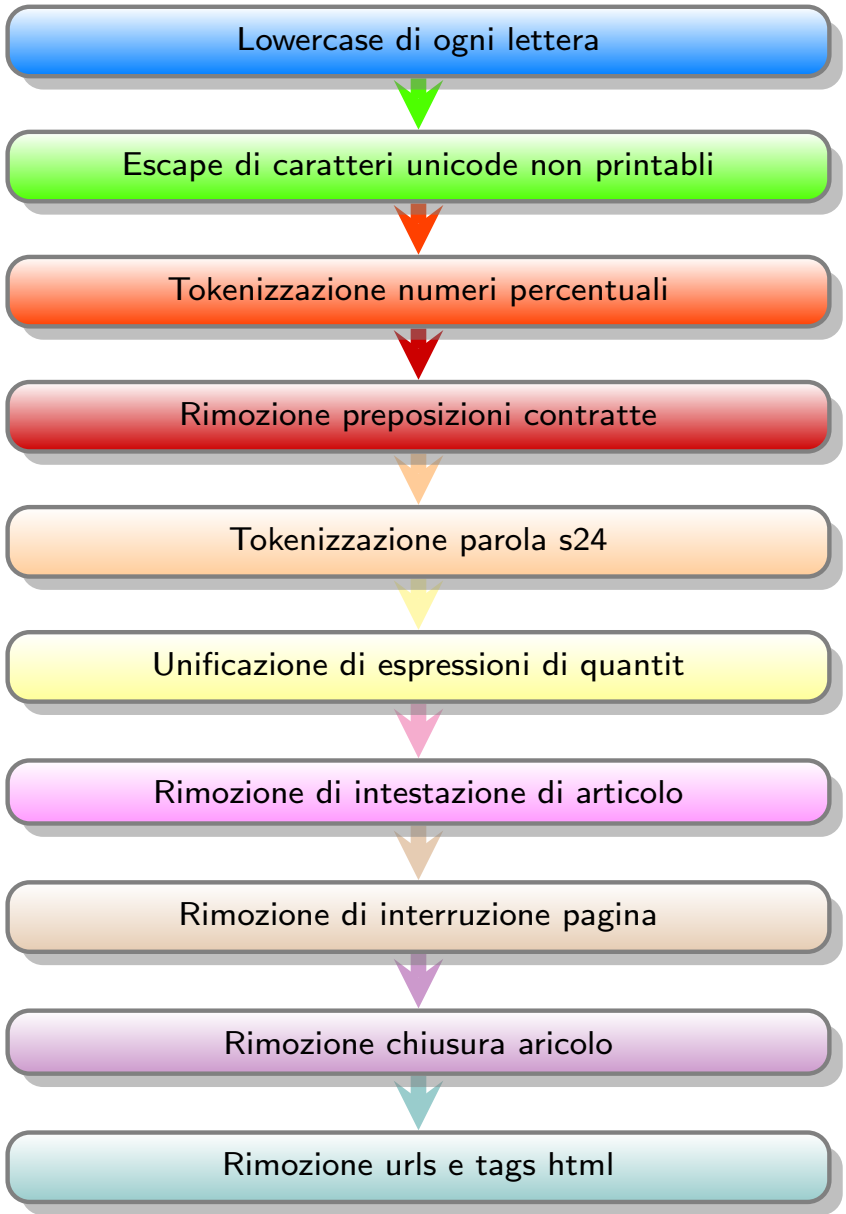
Outline

Pipeline completa



Outline

Cleaning pipeline



Outline

Hello

Qui speghiamo per bene come funziona word2vec

Outline

Hello

Qui speghiamo per bene come funziona word2vec