

RAPPRESENTAZIONE SEMANTICA E WORD EMBEDDINGS

Eleonora Litta

DALLA FILOLOGIA ALLA STATISTICA COMPUTAZIONALE

Obiettivo della lezione: Comprendere come i computer rappresentano il significato di parole, grafi e sensi per scopi di NLP (*Natural Language Processing*).

Il Grande Salto: Passaggio dalla conoscenza linguistica **esplicita** (annotazione "questo è un NOME") a quella **implicita** (astrazione statistica basata sul contesto).

Tradizionalmente, cerchiamo il significato sul dizionario (conoscenza esplicita). La linguistica computazionale adotta invece l'**ipotesi distribuzionale**.

L'IPOTESI DISTRIBUZIONALE

"Conoscerai una parola dalla compagnia che frequenta" (Firth, 1957)

Principio Fondamentale: Parole che compaiono in contesti simili tendono ad avere significati simili.

Riferimento Filosofico: L. Wittgenstein (*Ricerche Filosofiche*): "Il significato di una parola è il suo uso nel linguaggio".

Esempio Pratico: Se cerchiamo la parola '*amicitia*' su <https://embeddings.lila-erc.eu/>, il computer non ci restituisce una definizione del dizionario, ma i suoi 'vicini' statistici: *benevolentia*, *coniunctio*, *familiaritas*. Questo dimostra che il modello ha 'imparato' il senso di amicizia osservando quanto spesso compaia vicino a termini di affetto e benevolenza nei testi classici

IL PROBLEMA DELLA RAPPRESENTAZIONE DIGITALE

Perché non bastano le lettere (ASCII)?

Variabilità: Per un computer, "rosa" e "rosae" sono solo sequenze diverse di bit. Una parola di 4 lettere occupa 32 bit, una di 5 ne occupa 40; la dimensione variabile complica il confronto tra termini.

Soluzione => One-hot Representation: Ogni parola è un vettore in cui tutte le dimensioni sono zero tranne quella corrispondente al suo indice nel vocabolario (es. 100 parole = array di dimensione 100).

Limite: Questa rappresentazione, oltre ad occupare uno spazio immenso, non cattura alcuna somiglianza semantica; per il computer, "mensa" e "tabula" risulterebbero distanti quanto "mensa" e "gladius".

IL MODELLO DELLO SPAZIO VETTORIALE (VSM)

La metafora spaziale del significato

Concetto: Il significato di una parola è un luogo in uno spazio geometrico multidimensionale.

Distanza Semantica: La vicinanza fisica tra due punti (vettori) nello spazio corrisponde alla loro somiglianza di significato.

Per rappresentare così il significato, dobbiamo trasformare le parole in un insieme di coordinate. Possiamo farlo in 2 modi:

- coordinate di parole rispetto a **parole**
- coordinate di parole rispetto a **testi**

Coordinate: Le "mappe" vengono costruite automaticamente analizzando le co-occorrenze delle parole in enormi corpora di testo.



METODI BASATI SUL CONTEGGIO (COUNT-BASED)

Matrice Termine-Documento: Le righe sono parole, le colonne sono documenti (es. le Cantiche della Commedia).
Immagine da Jezek e Sprugnoli 2023.

Peso Informativo: Non basta contare le frequenze grezze; parole comuni come "il" o "et" sono poco informative.

Tf-Idf e PMI: Funzioni statistiche che pesano l'importanza di un termine rispetto a quanto esso caratterizza un documento specifico rispetto agli altri.

	INFERNO	PURGATORIO	PARADISO
gloria	2	6	15
corpo	14	16	15
luce	4	12	56
selva	13	5	0
amore	59	20	32
occhi	212	86	77
mondo	143	23	68

Tab. 5.1. Esempio di matrice termine-documento per alcuni termini scelti nelle tre Cantiche della *Divina Commedia*

L'ERA DI WORD2VEC (2013)

Modelli Predittivi vs Modelli di Conteggio

Il Cambiamento: Invece di contare, addestriamo una rete neurale a prevedere una parola partendo dal contesto.

Due Architetture:

- **CBOW (continuous bag-of-words):** Predice la parola attuale usando il contesto circostante.
- **Skip-gram:** Usa una parola target per prevedere le parole del contesto.

Vettori Densi: I "pesi" appresi dal modello durante l'allenamento diventano i nostri vettori (embeddings). Vettori corti (tra 50 e 1,000 dimensioni), densi e informativi, così chiamati perché sono contenuti ('embedded') in uno spazio semantico.

PROPRIETÀ E ANALOGIE VETTORIALI

Matematica del Significato

Dimensioni Ridotte: Gli embeddings sono vettori densi e corti (solitamente tra 50 e 1000 dimensioni).

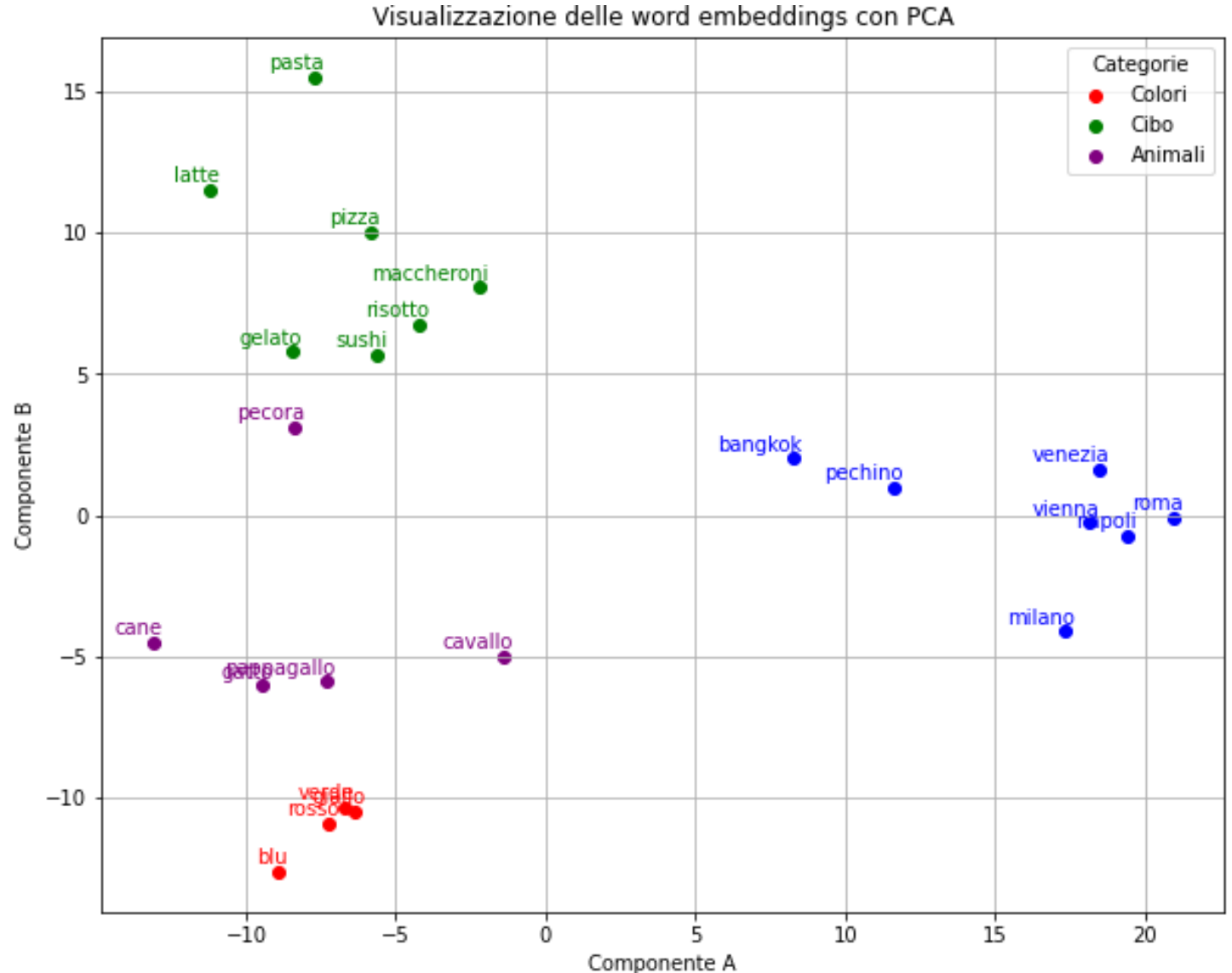
Risoluzione di Analogie: È possibile eseguire operazioni aritmetiche tra concetti:

$$\text{vec}(\text{"rex"}) - \text{vec}(\text{"regina"}) + \text{vec}(\text{"dominus"}) \approx \text{vec}(X)$$

Utilità: Questi modelli permettono di visualizzare raggruppamenti (cluster) di parole simili tramite tecniche come la PCA (*Principal Component Analysis*).

VISUALIZZAZIONI: PCA

Possiamo proiettare i nostri embeddings in uno spazio per visualizzarli. Perché dei vettori a 100 dimensioni come quelli che usiamo siano visualizzabili in 2D o 3D, tuttavia, dobbiamo **ridurre le dimensioni** (da 100 a 2 o 3). Esistono tecniche statistiche per farlo, come ad es. La **principal component analysis** (PCA) usata qui. Queste tecniche permettono di ridurre la complessità di dati multidimensionali, ma comportano una perdita di informazione.

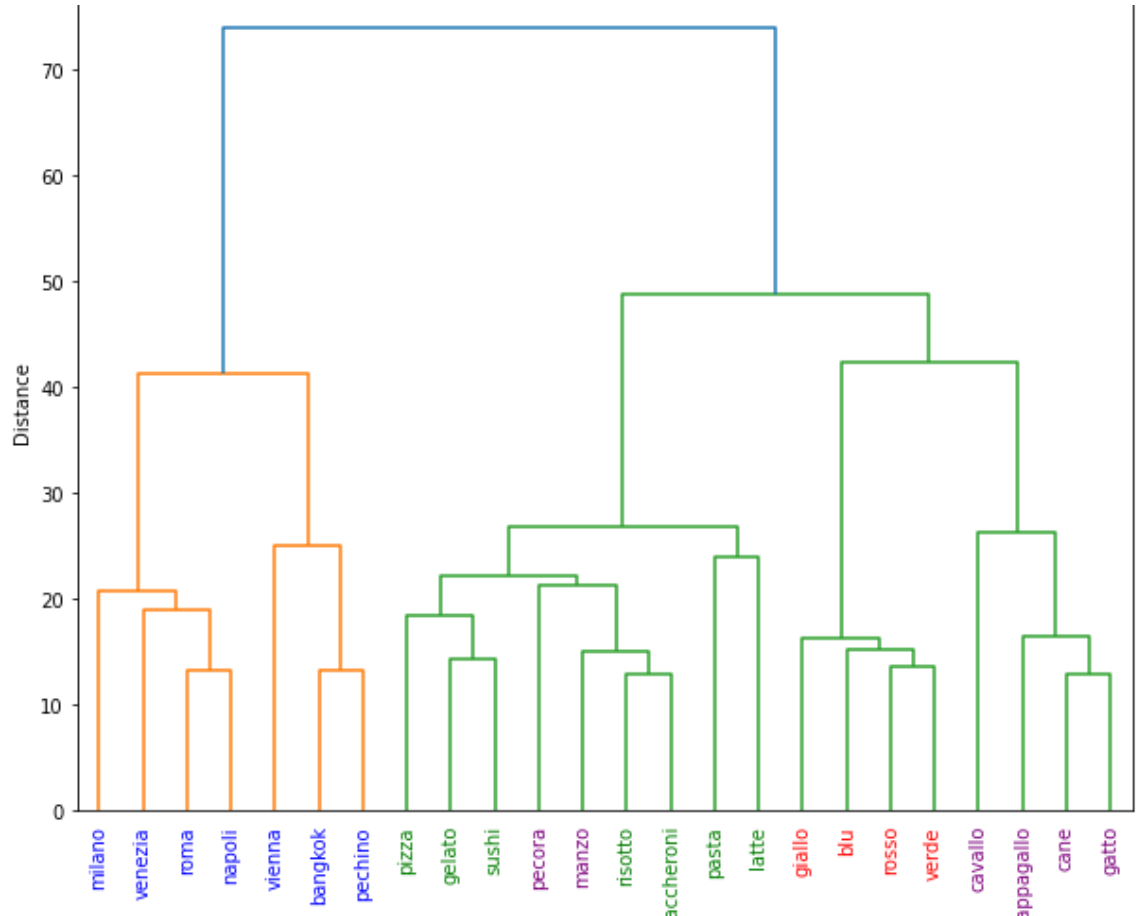


VISUALIZZAZIONI: CLUSTERING

Oppure, per visualizzare gli embedding possiamo usare un **algoritmo di clustering**, ovvero un metodo statistico **non supervisionato** che permette di raggruppare i dati in base alla loro somiglianza. Il **clustering gerarchico** può seguire due strategie:

- **clustering divisivo** (top-down): parte da un unico grande insieme contenente tutti i dati, che viene poi suddiviso progressivamente in gruppi più piccoli;
- **clustering agglomerativo** (bottom-up): parte da ogni dato come un insieme a sé stante, e unisce iterativamente i gruppi più simili, fino a formare un unico insieme complessivo.

I risultati possono essere rappresentati in un **dendrogramma**, un diagramma ad albero che visualizza graficamente il processo di fusione (o divisione) dei gruppi nel tempo, mostrando le relazioni gerarchiche tra i cluster.



LIMITI DEGLI EMBEDDINGS STATICI

La sfida della polisemia

Meaning Conflation Deficiency: Modelli come Word2vec assegnano un unico vettore a ogni parola, ignorando i suoi molteplici significati.

Esempio: La parola *seno* (parte del corpo, insenatura, funzione matematica) riceve la stessa identica rappresentazione statica.

Necessità di Contesto: Per capire se "cell" indica una cellula biologica o una cella carceraria, serve guardare le parole vicine (es. "membrane").

Prendiamo il termine *virtus*. Possiamo vedere come il suo vettore sia vicino sia a termini militari (*fortitudo*) sia a termini morali (*constantia*). Questo evidenzia il limite degli embedding statici: il rischio di 'schiacciare' due sfumature diverse in un unico punto.

EMBEDDINGS CONTESTUALIZZATI E BERT

Rappresentazioni dinamiche per ogni uso

BERT (2019): Modello basato sull'architettura *Transformer* che genera embeddings sensibili al contesto.

Bidirezionalità: A differenza dei modelli precedenti, BERT guarda contemporaneamente sia a sinistra che a destra di una parola.

Masked Language Modeling: Durante l'addestramento, il modello deve indovinare parole "mascherate" (nascoste) all'interno di una sequenza.

IL MECCANISMO DI ATTENZIONE (SELF-ATTENTION)

Come il computer "legge" con intelligenza

Focus Selettivo: Consente al modello di "prestare attenzione" alle parti più rilevanti di una sequenza per elaborare un dato elemento.

Dipendenze a Lungo Raggio: Traccia relazioni linguistiche anche tra parole distanti nella frase.

Query, Key, Value: Ogni parola viene trasformata in tre vettori che gestiscono la richiesta, l'offerta e il contenuto informativo dell'attenzione.

GESTIONE DELL'OUT-OF-VOCABULARY (OOV)

Scomposizione in sottoparole (Subword Tokenization)

Scomposizione: Quando incontra parole sconosciute, il sistema le divide in unità più piccole (n-grammi di caratteri).

FastText: Induce l'embedding di una parola mai vista mediando le rappresentazioni dei suoi componenti (es. "memoryless" = "memory" + "less").

Vantaggio per le Lingue Classiche: Utile per gestire la ricca flessione morfologica e i termini rari.

Il visualizzatore di LiLa risolve il problema delle 'parole mai viste' (OOV) collegando gli embedding alla lemmatizzazione.

ANALISI DIACRONICA DEL SIGNIFICATO

Gli **embeddings** non sono solo istantanee statiche del linguaggio, ma possono fungere da vera e propria "finestra" sulla semantica storica.

Invece di creare un unico modello per tutto il latino, i ricercatori addestrano modelli separati su corpora di periodi diversi.

=> visualizzazione di *virtus* in LiLa Lemma Embeddings in due corpora distinti.

<https://embeddings.lila-erc.eu/>