Document Semantic Similarity

TIS Project

Alberto Pirovano Francesco Picciotti

Politecnico di Milano

23rd May 2017

Outline

Introduzione

Le tecniche adottate attualmente per trovare la **similitudine semantica tra testi** si basano su tre approcci:

1. NLP Tradizionale
2. Vector Space Model
3. Deep Learning based

Outline

NLP Tradizionale

Questo approccio si basa sull'utilizzo delle tradizionali tecniche di **Natural Language Processing** e si costituisce dei seguenti step:

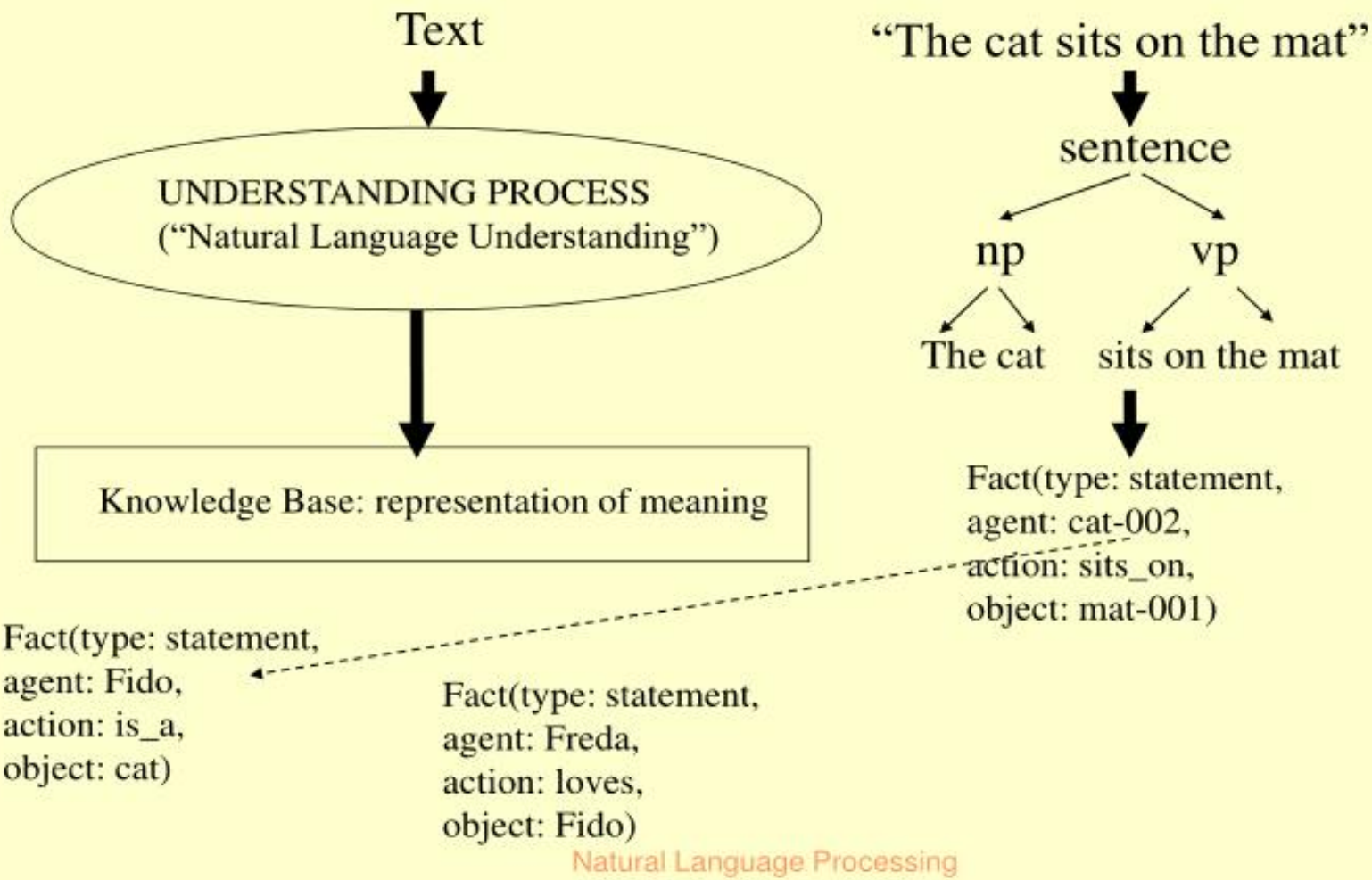
- Cleaning dei dati
- Pos-Tagging
- Stemming o Lemmatisation
- Parsing
- Ontologia

Tuttavia, dato che il nostro lavoro è molto **sensibile** e **dipendente** dalla qualità dei tool utilizzati, abbiamo trovato alcune consistenti **criticità** riguardanti:

- **L'affidabilità** del Pos-Tagger italiano di TreeTagger
- **Reperire** una Ontologia e un parsing tool per la lingua italiana

NLP Tradizionale

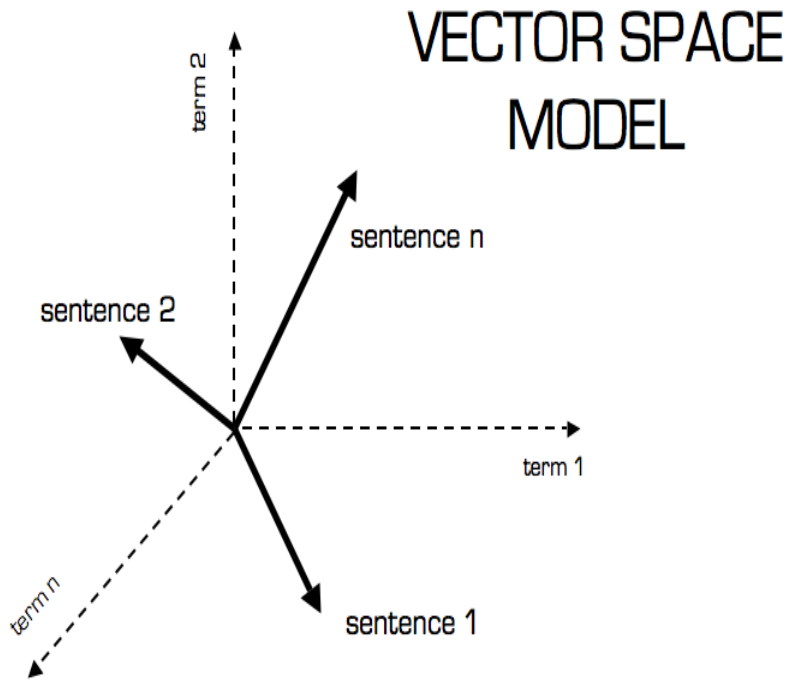
NLP: the process



Outline

Vector Space Model - Explanation

Differentemente dal precedente, questo approccio ha le sue basi nello sviluppo di una **rappresentazione geometrica e vettoriale** per i documenti testuali. **Documenti e query** sono rappresentati da vettori con un numero di elementi pari al numero di termini presenti nel vocabolario. Tipicamente i termini sono le **parole distinte** presenti nell'insieme di documenti. A valle di questa rappresentazione vengono spesso utilizzate le **operazioni vettoriali** per confrontare due **documenti**.



Vector Space Model - Encoding

Nel vector space model proposto da **Salton, Wong and Yang** i vettori sono composti da **weights**, ognuna associata ad un termine del dizionario e calcolata tramite **tf-idf**.
Considerando un **documento** d_j, questo viene rappresentato tramite un **vettore** d_j:

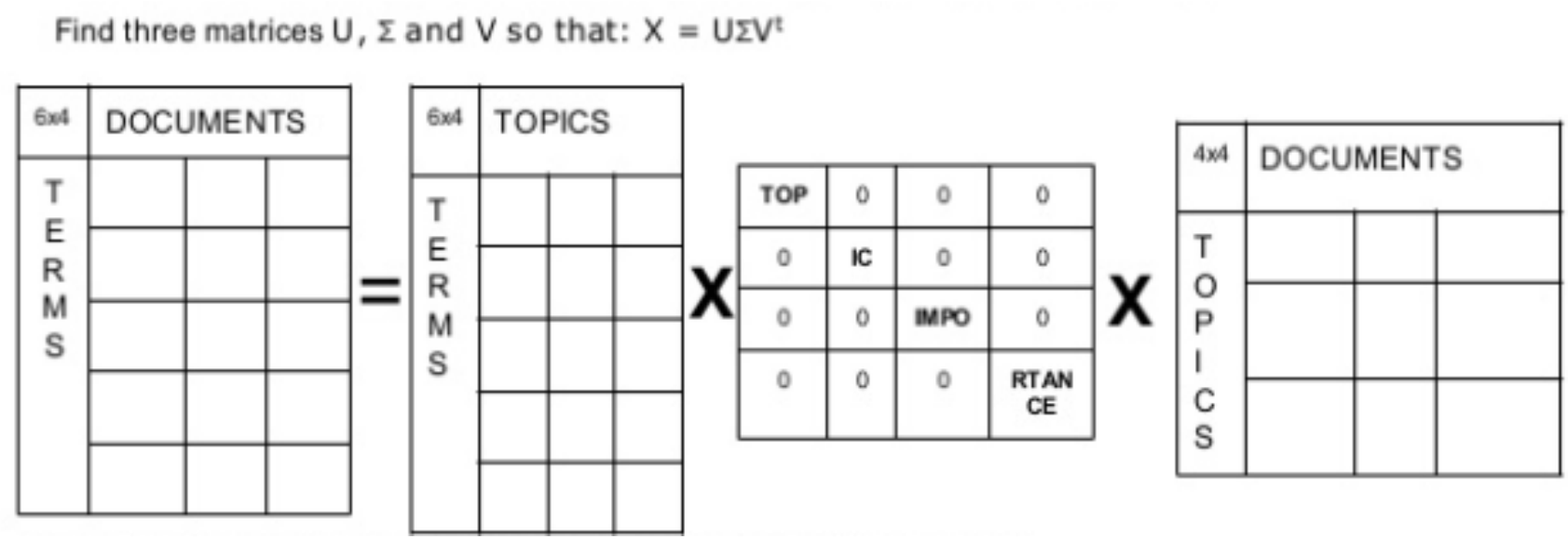
- $d_j = (w_{1,j}, w_{2,j}, ..., w_{t,j})$ **dove**: $w_{i,j} = tf_{t,d} * \log \frac{|D|}{|\{d' \in D | t \in d'\}|}$

Vector Space Model - LSA

Latent Semantic Analysis, anche detta LSI in information retrieval, è una tecnica di **Topic Modelling** che si colloca a valle del **document encoding** con **tfidf**.
È una tecnica di **feature extraction** che permette di migliorare significativamente la qualità di un lavoro di **clustering**.
Questa procedura viene usata per generare una categorizzazione di un set di documenti in un set di topic o anche per osservare le parole che descrivono un certo topic.
Si basa sulla creazione di una **Document-Term Matrix** nella quale le **righe** rappresentano le parole del **Bag Of Words** e ha una **colonna** per **documento** nel corpus.
Il cuore di questa procedura sta nella **riduzione della dimensionalità** di questa matrice tramite **SVD**.

Vector Space Model - LSA - SVD

- **U:** ogni riga rappresenta una parola, mentre ogni colonna rappresenta una **dimensione nel latent space**; le colonne sono **ortogonali** l'una all'altra e sono **ordinate** in base alla varianza dei dati lungo di esse.



È importante sottolineare che usando solo **k** delle **m** dimensioni latenti si ottiene una **approssimazione** di **X**.

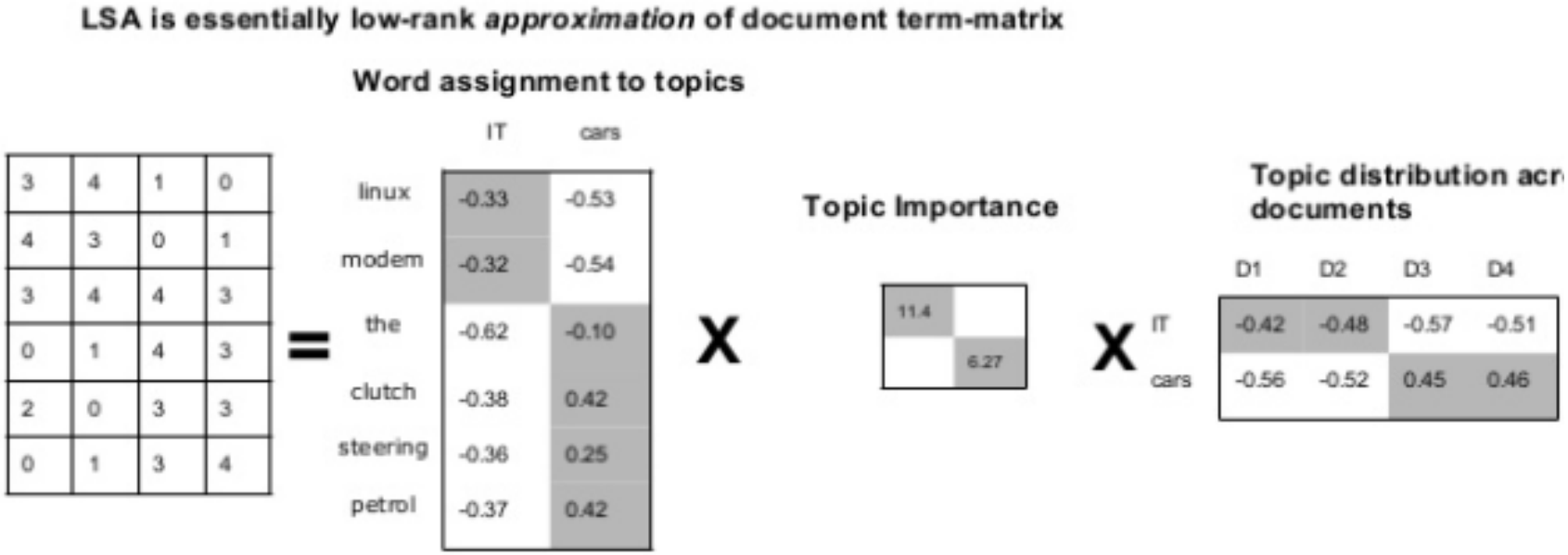
Vector Space Model - LSA

Questa procedura ci permette di:

- **Estrarre** quanti topic desideriamo da un set di documenti.
- **Conoscere** la rilevanza di un certo topic dopo averlo estratto, in questo modo siamo in grado di fermare il processo di estrazione quando i topic cominciano a diventare poco significativi.
- **Categorizzare** documenti in topic
- **Descrivere** topics con le parole del **Bag Of Words**.

Vector Space Model - LSA - Example

In questo esempio possiamo osservare la riduzione di dimensionalità:



Il processo di LSA permette di costruire le 3 matrici che vediamo sopra, ognuna con una sua utilità:

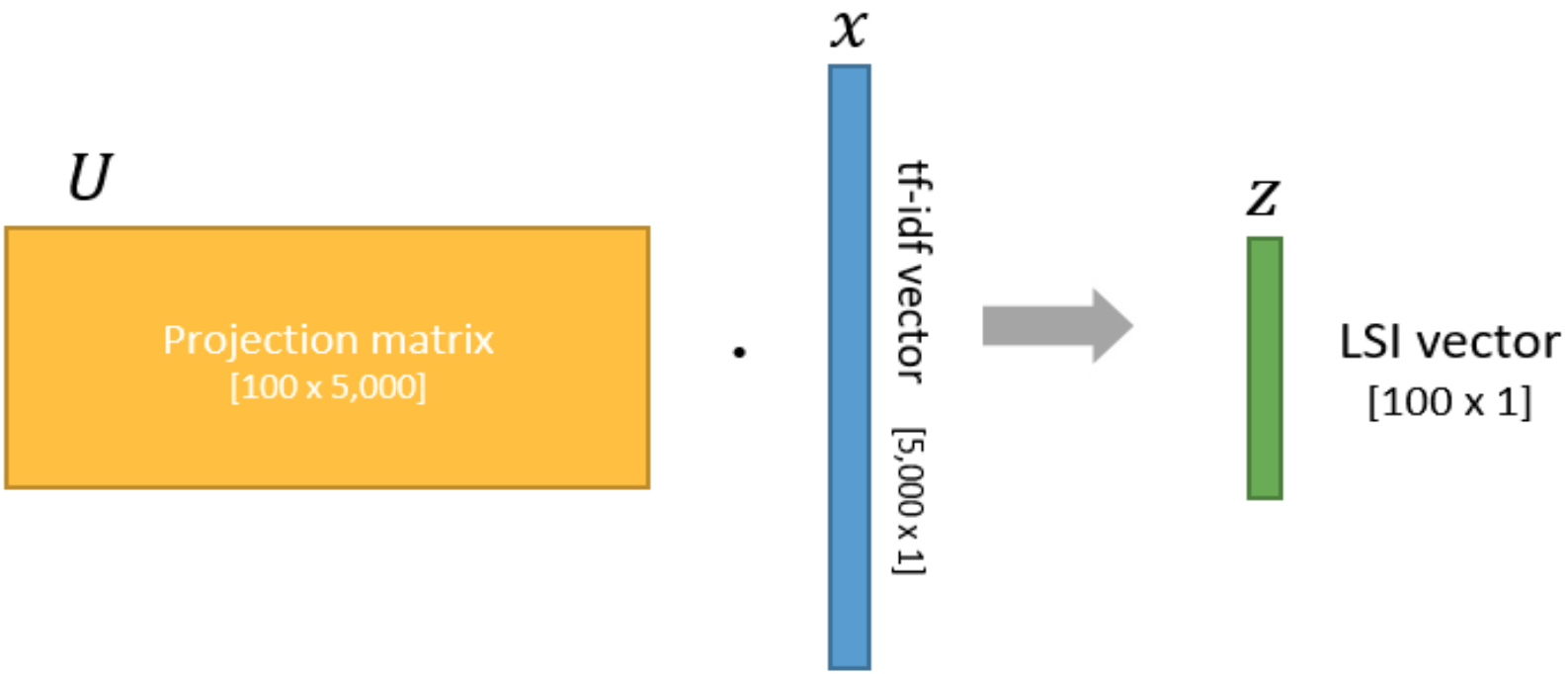
1. **Word assignment to topics**
2. **Topic importance**
3. **Topic distribution across documents**

Vector Space Model - LSA - Example

Ipotizzando un numero di topics pari a 100 e la cardinalità del **Bag of Words** pari a 5000, per ottenere la rappresentazione dei documenti in termini di topics, cioè un vettore **z**, dobbiamo:

1. **Vettorizzare** un documento in **x**.
2. **Proiettare** il **tfidf** vector **x** sul topic space.

Se definiamo la matrice $U = (Wordassignmenttotopics)^T$ possiamo visualizzare la **proiezione** in questo modo:



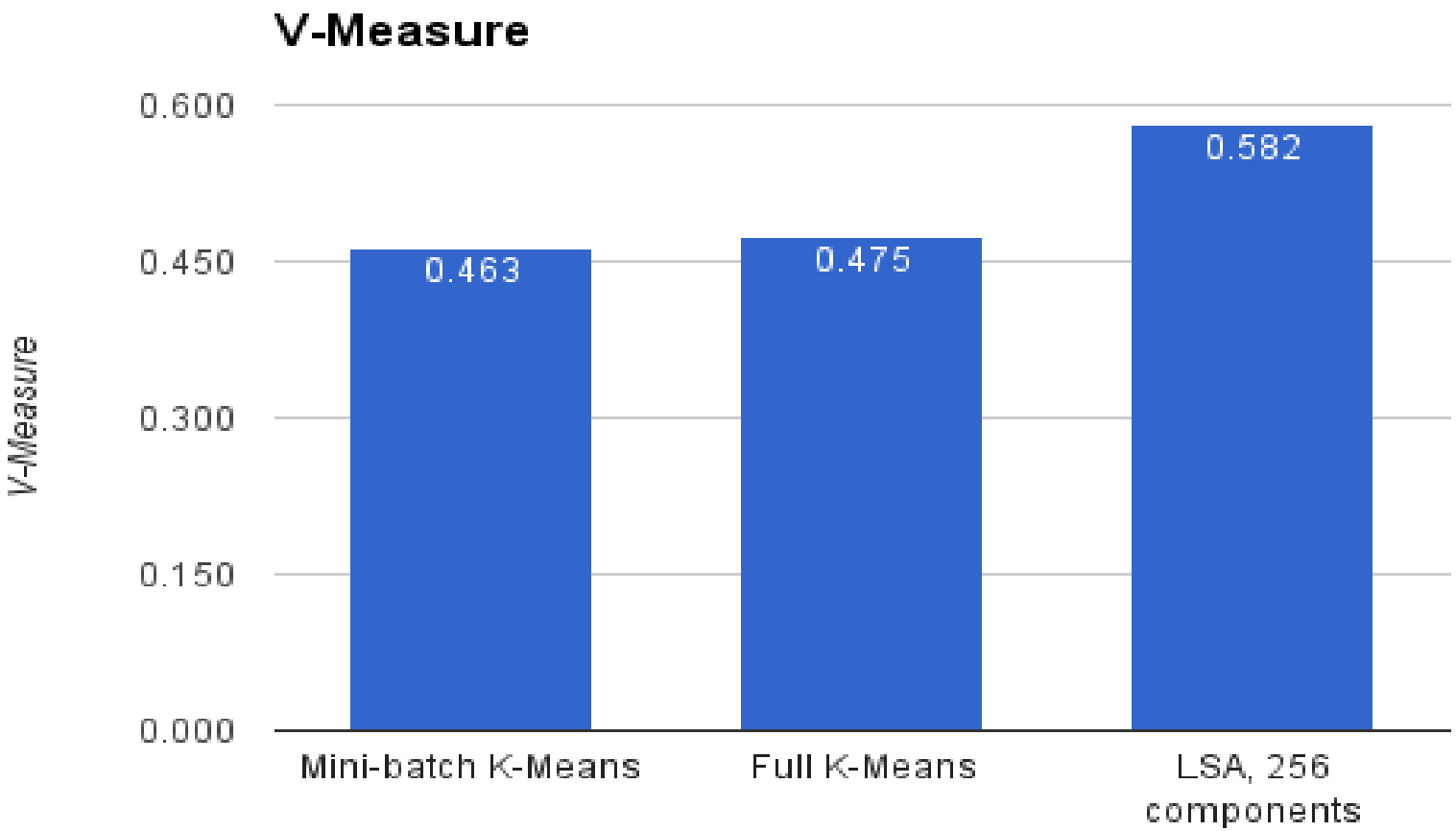
Vector Space Model - Algorithm

Riassumendo, possiamo individuare i seguenti **passaggi**:

1. Cleaning dei dati (Stemming o Lemmatisation)
2. Vectorization using TF/IDF (**x** nella figura nella slide precedente)
3. LSA (optional) (**z** nella figura precedente)
4. Clustering using similarity measure (Cosine, Pearson, ...) tra **x** vectors o **z** vectors

Vector Space Model - LSA - Performance

Come possiamo vedere nella figura, utilizzare le classiche tecniche di **clustering** senza **LSA** può portare una riduzione rilevante delle **performances**:



Vector Space Model (con LSA) - Pros and Cons

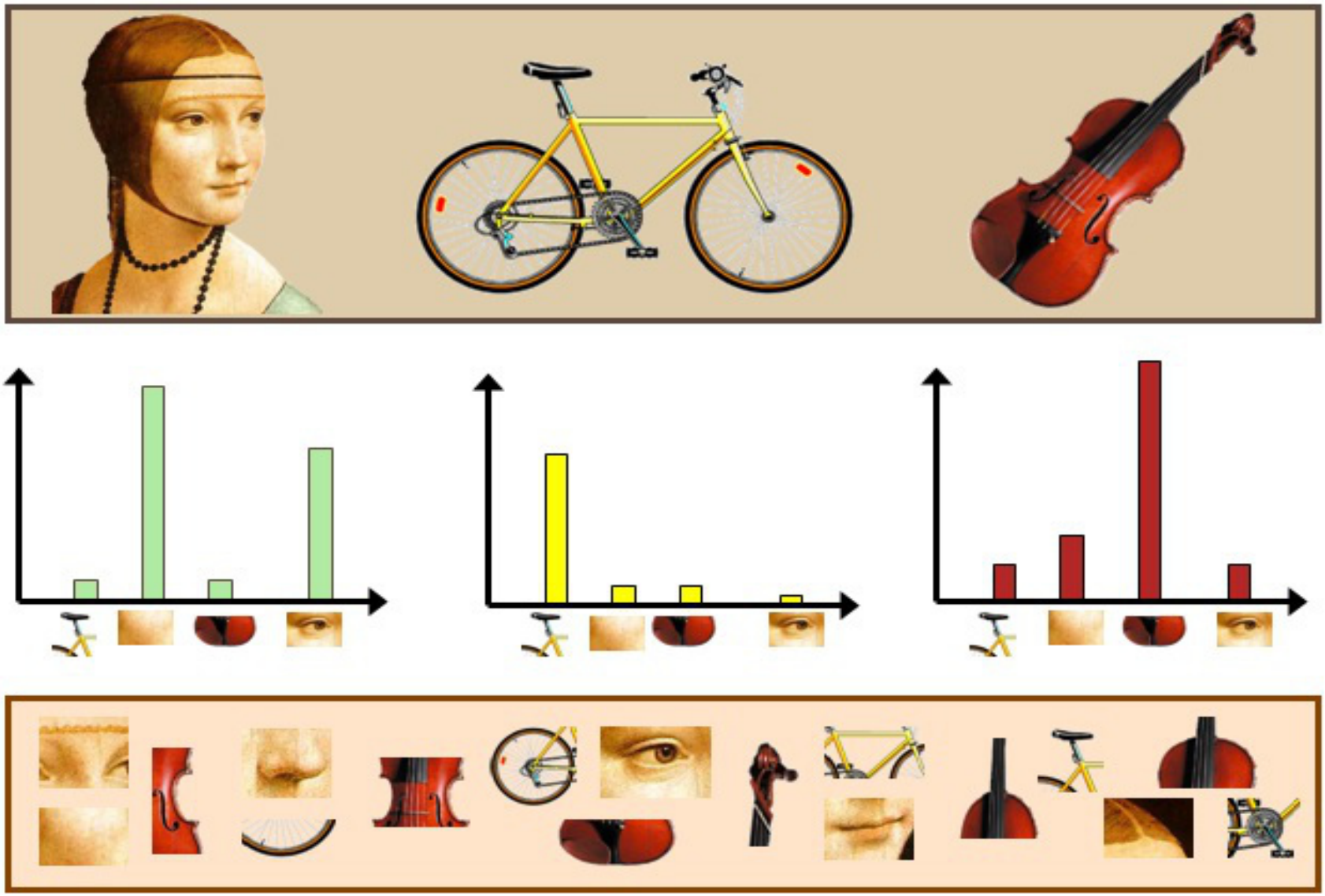
Cons:

- Non adatto a trattare **lunghi documenti**, infatti a causa della alta **dimensionalità** il valore **tfidf** dei componenti si riduce, riducendo di conseguenza il **dot-product**.(LSA fixes)
- **Fortemente sensibile a Falsi Negativi**, infatti documenti nello stesso **contesto** ma con diversa terminologia non saranno considerati simili.(LSA fixes)
- Perde **l'ordine delle parole**.
- **Pre-processing** dipendente.
- **Le dimensioni generate possono essere difficili da interpretare**, sensate matematicamente ma non dal punto di vista del natural language

Pros:

- Modello semplice basato sull'**algebra**.
- Le **weights** non sono binarie.
- Permette di calcolare un grado di **similarità continuo**.

Vector Space Model - Limiti



Outline

Deep learning based

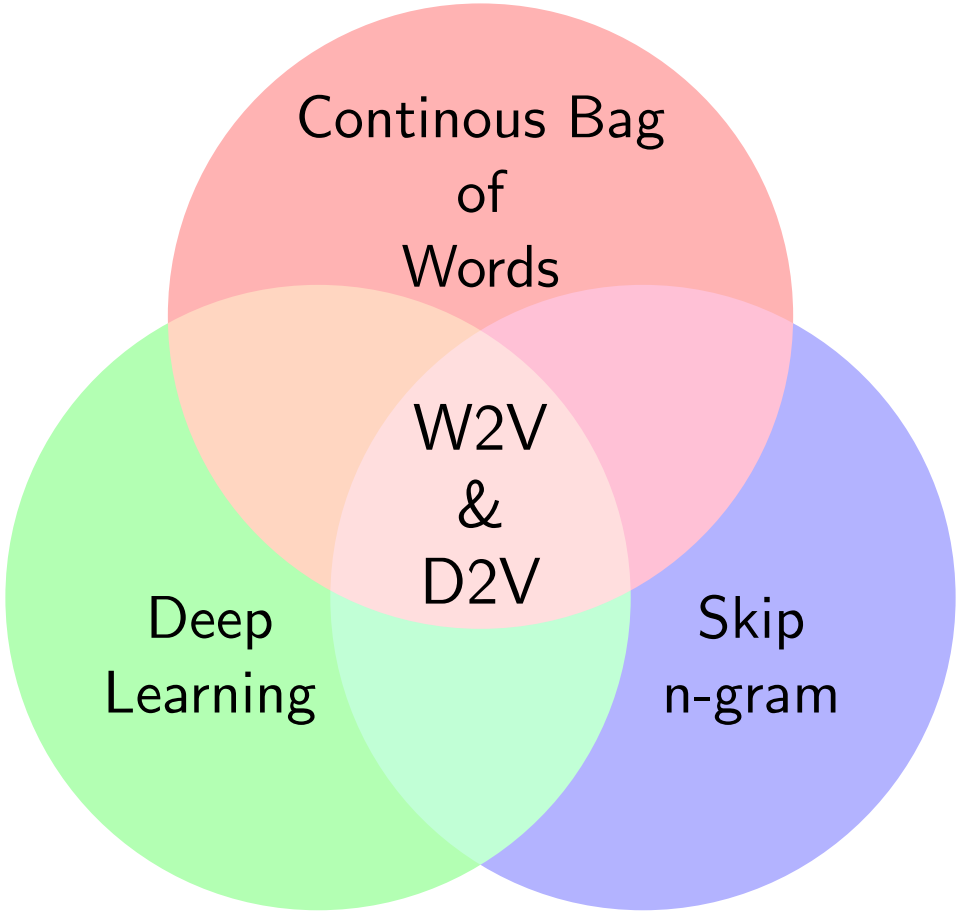
Il terzo approccio che proponiamo è molto diverso dai precedenti, infatti si basa sull'utilizzo di **neural networks**. La fondamentale differenza è che, mentre il secondo **definiva un algoritmo** per determinare le rappresentazioni vettoriali delle parole (tfidf), questo **definisce una neural network** con la task di imparare quell'algoritmo dai dati. Questa **NN** può essere **costruita** in due diversi modi, ognuno dei quali descrive **come** imparare la **word-representation** per ogni parola:

- **Continous Bag Of Words model**
- **Skip Grammar model**

Dato che il processo di apprendimento è **unsupervised**, questi modelli permettono di **determinare la task** tramite la definizione di un target per ogni input.

Deep Learning based

Questa tecnica quindi si basa su 3 ambiti:



Deep Learning based: Word2Vec & Doc2Vec

Questo metodo è stato implementato da **Google** sotto il nome di **Word2Vec** e **Doc2Vec**. Il primo permette di determinare le **relazioni semantiche** che un particolare **corpus di testi** assegna ad un **Bag Of Words** di parole. **Doc2Vec** invece è una tecnica che si configura come una **estensione di Word2Vec** la quale, preso in ingresso un set di documenti (corpus), genera un **grado di similarità**.

Deep Learning based - Pros and Cons

Dopo una ampia **discussione**, seguita da una approfondita **analisi critica** di questi due approcci, siamo giunti alle seguenti **conclusioni**, che in termini di pro e contro si possono riassumere nel seguente modo:

Pros:

- Molto meno dipente da un preprocessing
- Combina il metodo *Geometrico* con quello *NLP Tradizionale*
- **Non sfrutta una ontologia, ma la crea**
- **Language independent**
- **Context-Aware**, **tiene in considerazione anche l'ordine delle parole**

Cons:

- Tecnica **unsupervised**
- Necessità di un esperto per **validare** la similarità
- Può risultare in **GIGO** system (Garbage In Garbage Out)

Outline

Dataset

"Preprocessing is 80% of NLP work"

Lev Konstantinovskiy

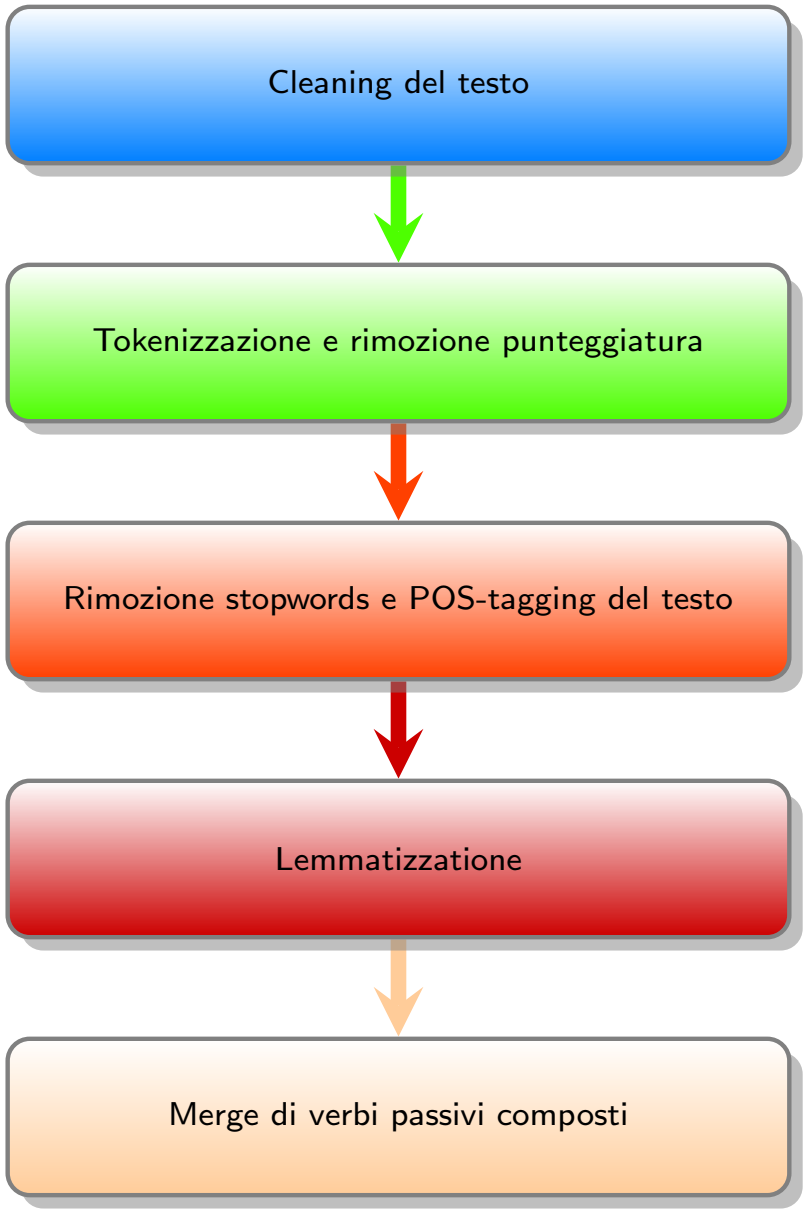
Il **dataset** si suddivide in due corpora:

- 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.

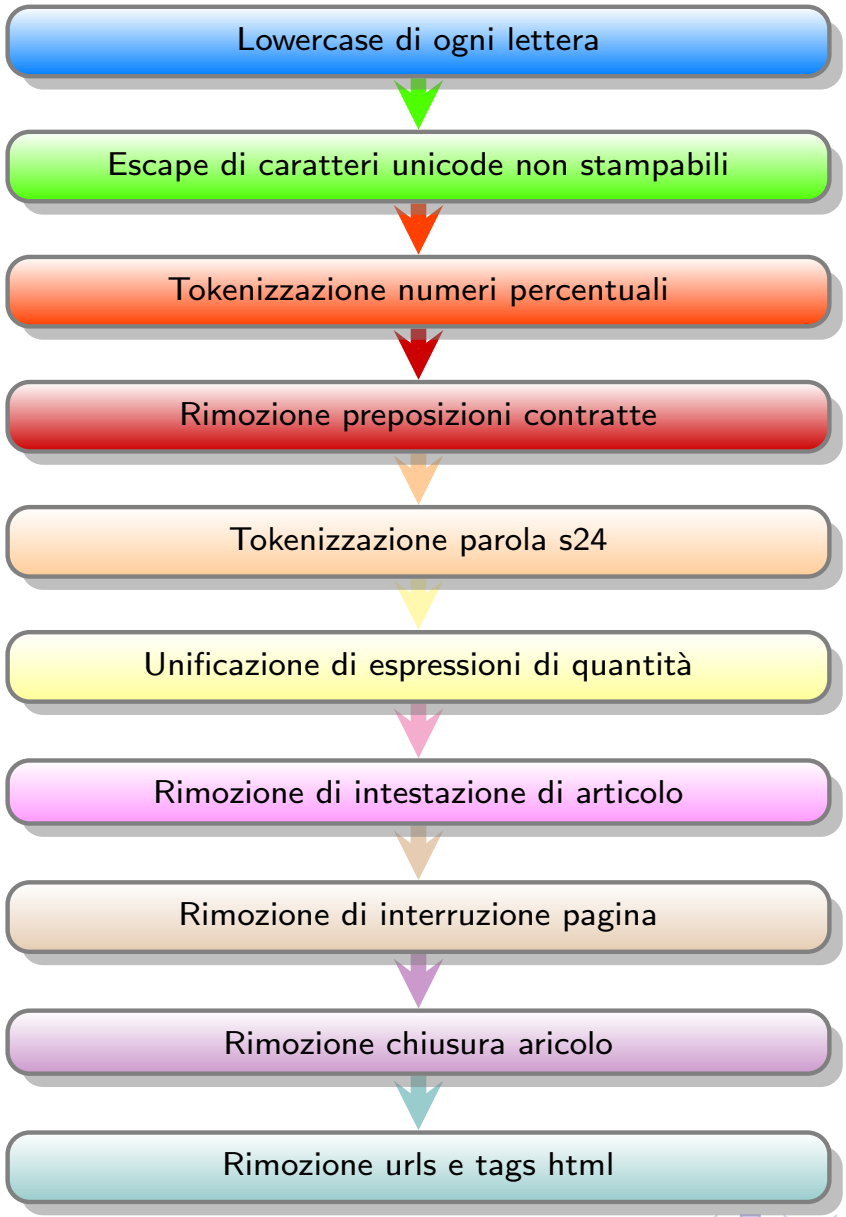
Outline

Pipeline completa



Outline

Cleaning pipeline



Outline

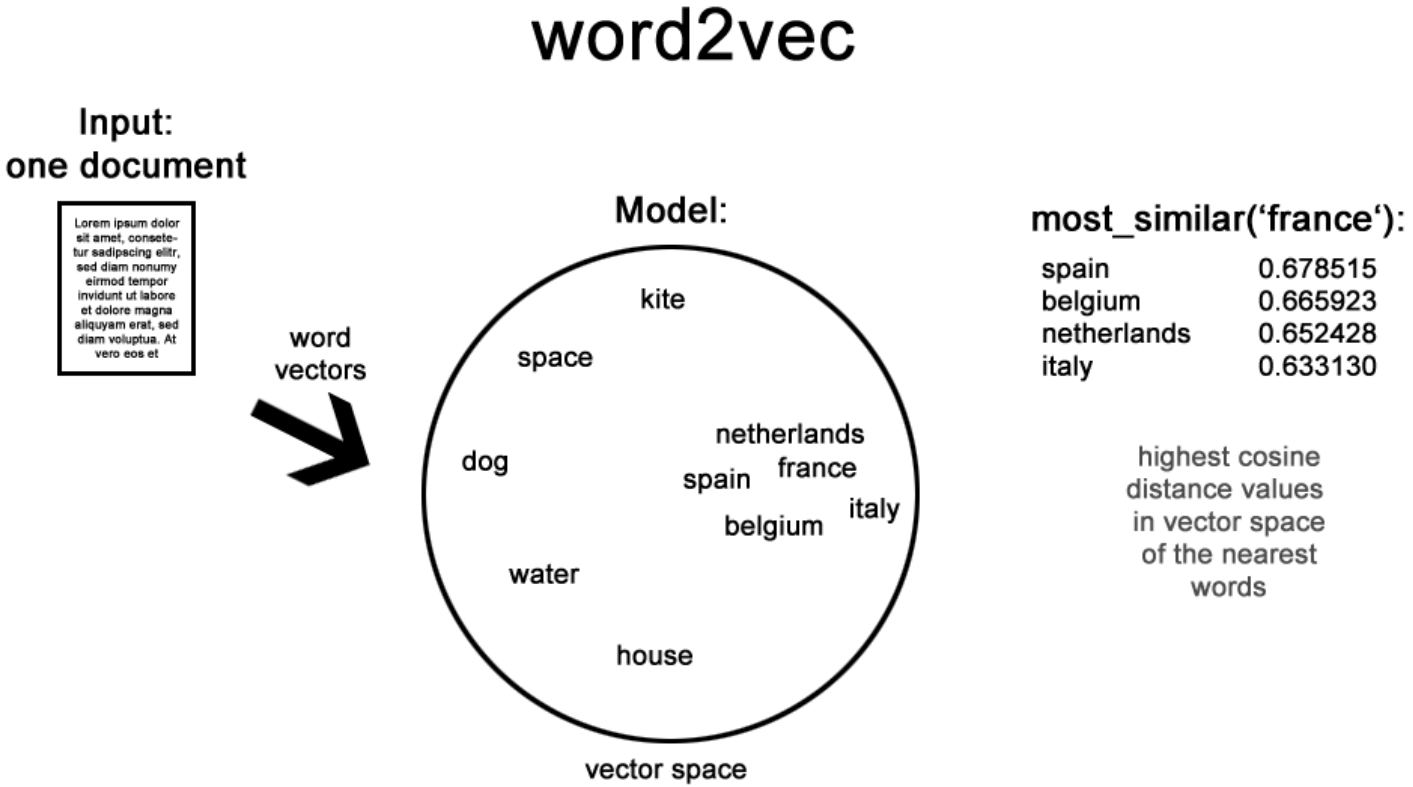
Word2Vec - Task

Word2Vec è un gruppo di modelli che vengono usati per generare **word-embeddings** da **unstructured text**. Per questo scopo viene usata una **two-layers neural network** che viene trainata per ricostruire il **contesto linguistico** delle parole.

- **Input: corpus** di testo
- **Output: vector-space**

Ogni parola **unica nel corpus** verrà rappresentata con un vettore nel **vector-space**. L'algoritmo genererà degli word-embeddings tali per cui parole che condividono lo stesso contesto nel corpus risulteranno vicine nel **vector space**.

Word2Vec - Idea



Word2Vec - From unsupervised to supervised

Questa task può essere definita in due modi:

1. **Skip Grammar** - predire il **contesto** data la parola in input
   - **Input:** focus-word
   - **Target:** context-words
2. **Continous Bag Of Words** - predire la parola in input dato il **contesto**
   - **Input:** context-words
   - **Target:** focus-word

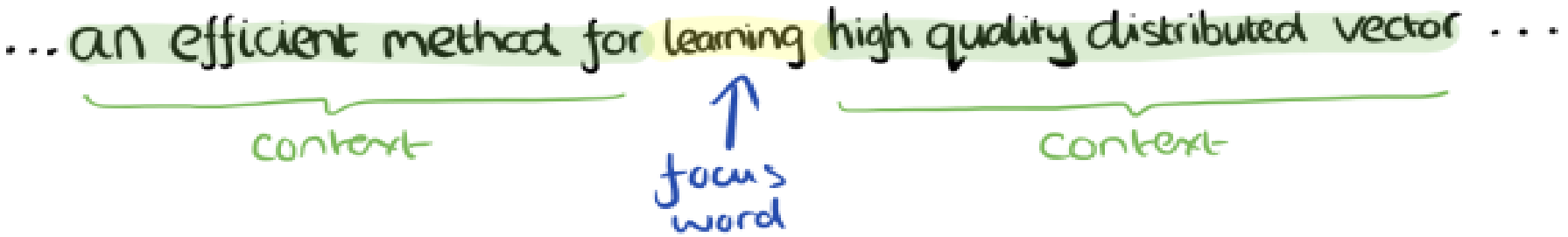
Definendo input e target pairs abbiamo definito un **supervised problem**, che affronteremo usando **logistic classifiers** (logistic regression).

Word2Vec - Encoding

Gli **input-target pairs** sono rappresentati tramite **One Hot Encoding**, mentre l'**output** della NN sarà un vettore di probabilità di dimensione pari al numero di parole del **vocabolario**.
Di conseguenza se abbiamo un vocabolario di **dimensionalità V** avremo dei **vectors** costituiti da V elementi di cui un 1 e gli altri 0.

Word2Vec - Context

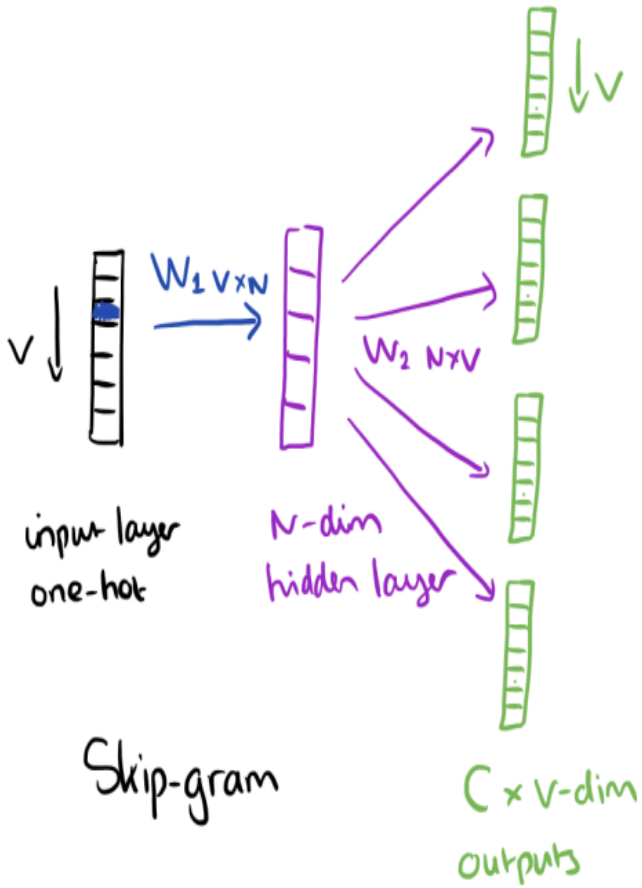
In entrambi i casi è necessario definire un **context** per la parola in input, che identifichiamo con un parametro chiamato $c = windowsize$. Tramite questa *windowsize*, una volta fissata una **input word/focus word**, chiamiamo **context** l'insieme delle c parole prima della input word più le c parole dopo la input word.



Terminologia:

- V = numero di elementi nel vocabolario
- N = numero di weights per word embeddings
- c = window size

Word2Vec - Skip Grammar



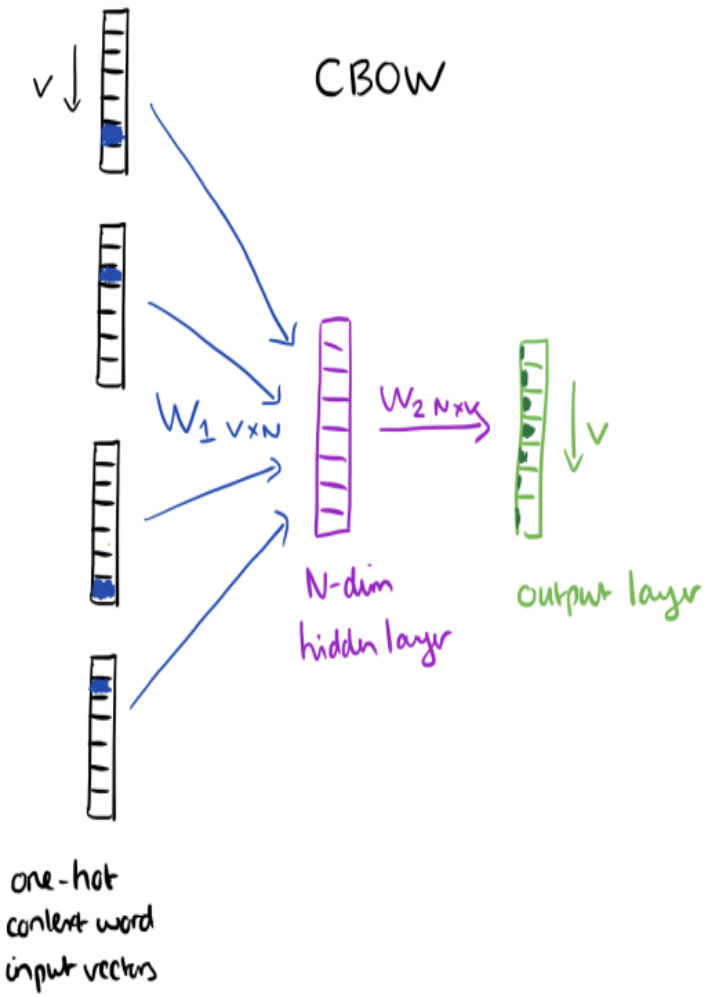
Data una sequenza di **training words** $w_1, w_2, ...w_M$, con $M = trainingsetsize$, il **training objective** di questo modello è trovare il parametro θ che massimizza la **Log-Likelihood**:

$$L = \frac{1}{M} \cdot \sum_{m=1}^{M} \sum_{-c<j<c, j \neq 0} \log p(w_{m+j}|w_m)$$

Ogni classification task, data una **focus word** w_m e una parola del contesto w_{m+j}, calcola **la probabilità** $p(w_{m+j}|w_m)$.

- Di conseguenza per ogni input vector definiamo circa **2c logistic classifiers**.

Word2Vec - CBOW



Data una sequenza di **training words** $w_1, w_2, ...w_M$, con $M = trainingsetsize$, il **training objective** di questo modello è trovare il parametro θ che massimizza la **Log-Likelihood**:

$$L = \frac{1}{M} \cdot \sum_{m=1}^{M} \log p(w_m | w_{m-c}, ..., w_{m+c})$$

Ogni classification task, date una serie di **context words**, le calcola **la probabilità** $p(w_m | w_{m-c}, ..., w_{m+c})$.

- Di conseguenza per ogni input vector definiamo **1 logistic classifier**.

Word2Vec - Matrices

Le due **weight matrices** $W_1[V \cdot N]$ e $W_2[N \cdot V]$ sono della stessa dimensione e contengono entrambe un **word-embedding vector** per ogni vocabulary word.
La **prima** fornisce la **input vector representation** delle vocabulary words e la **seconda** fornisce la **output vector representation** delle vocabulary words.
In generale data una parola $w \in V$:

- v_w è la sua **input representation**, vale a dire un vettore $[1 \cdot N]$ relativo a w che otteniamo da W_1:
- v'_w è la sua **output representation**, vale a dire un vettore $[1 \cdot N]$ relativo a w che otteniamo da W_2:

Word2Vec - Hidden layer

È interessante notare come, dato un input **OHE**, l'**hidden layer** funga da **lookup table** per selezionare il **word vector** relativo alla input word, v_{wInput}(no activation function).

$$[0 \quad 0 \quad 0 \quad 1 \quad 0] \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ 10 & 12 & 19 \\ 11 & 18 & 25 \end{bmatrix} = [10 \quad 12 \quad 19]$$

Possiamo quindi dire che:

- **Skip Grammar:** L'**output dell'hidden layer** (v_{wInput}) è il **word vector** relativo all'**input word**, in questo caso $v_{wInput} = [10, 12, 19]$.
- **CBOW:** Ogni **OHE** vector delle parole del contesto effettua il lookup e ottiene il word-embedding dalla matrice W_1. Successivamente viene fatta una media degli word embeddings ottenuti in modo tale da ottenere l'output dell'hidden layer ($v_{wAverage}$).

Word2Vec - Fake task

L'aspetto interessante di questa tecnica è che, dopo il training, **la rete non verrà utilizzata per la task su cui è stata addestrata**. Il training serve solo per aggiornare e migliorare i pesi delle matrici W_1 e W_2.

- Quello che ci interessa ottenere dopo la fase di training sono le weights della matrice W_1 che rappresenteranno gli **word vectors** per le parole del vocabolario (row-wise).

Word2Vec - Improvements

Il modello presentato ha dei problemi nel momento in cui si vuole applicare la backpropagation per l'aggiornamento delle weights. Infatti abbiamo due matrici da aggiornare che hanno in media 1M di weights ciascuna. Questo risulta in alcune problematicità, come ad esempio:

- time complexity.
- abbiamo bisogno di molti training data per evitare over-fitting.

Word2Vec - Improvements

Nel secondo paper gli autori di questa tecnica hanno proposto 3 miglioramenti:

- Phrases identification.
- Subsampling common words.
- Modificare il training objective usando la tecnica del negative sampling che porta ogni training example ad aggiornare solo una piccola percentuale di weights.

Questo procedimento ha anche significativamente migliorato le performance del modello.

Word2Vec - Improvements - Negative Sampling

Ci soffermiamo sulla terza tecnica perchè è la più interessante.

- W_1**:** Aggiorno solo il **weight vector** relativo alla input word, ma questo avviene a prescindere dal **"negative subsampling"**.
- W_2**:** Invece di aggiornare tutti gli **weight vectors** della matrice W_2 ad ogni training sample processato, prevede di selezionare un sottoinsieme di "negative words" da aggiornare.
  In questo contesto "negative word" corrisponde ad una parola per cui il target vector presenta uno 0.
  Solitamente il numero di negative selections che vengono fatte sono da 5 a 20.

Word2Vec - Improvements - Negative Sampling

Per selezionare gli **word vectors della matrice** W_2 da aggiornare consideriamo solo quelli riferiti alle parole più frequenti nel testo. La probabilità per una parola di essere campionata è:

- $p(w_i) = \frac{f(w_i)^{\frac{3}{4}}}{\sum_{j=0}^{V} f(w_j)^{\frac{3}{4}}}$

In questo modo possiamo **ridurre drasticamente il numero di weights che vengono aggiornate ad ogni training sample**, indicativamente con 3M di weights e 1800 aggiornamenti con "negative sampling" aggiorniamo lo 0.06% delle weights.

Word2Vec - Giustification

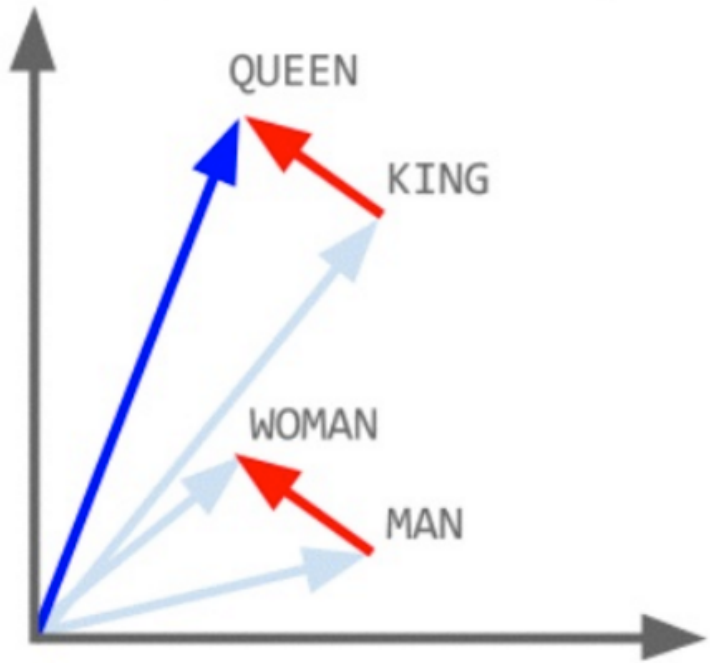
Se due parole vengono usate in contesti simili, allora il modello deve generare risultati molto simili per queste due parole.
Un modo che la NN ha per fare predizioni simili per queste due parole è utilizzare due **word vectors** simili.
La rete così risulta motivata a imparare simili **word vectors** per queste due parole.

Word2Vec - Example

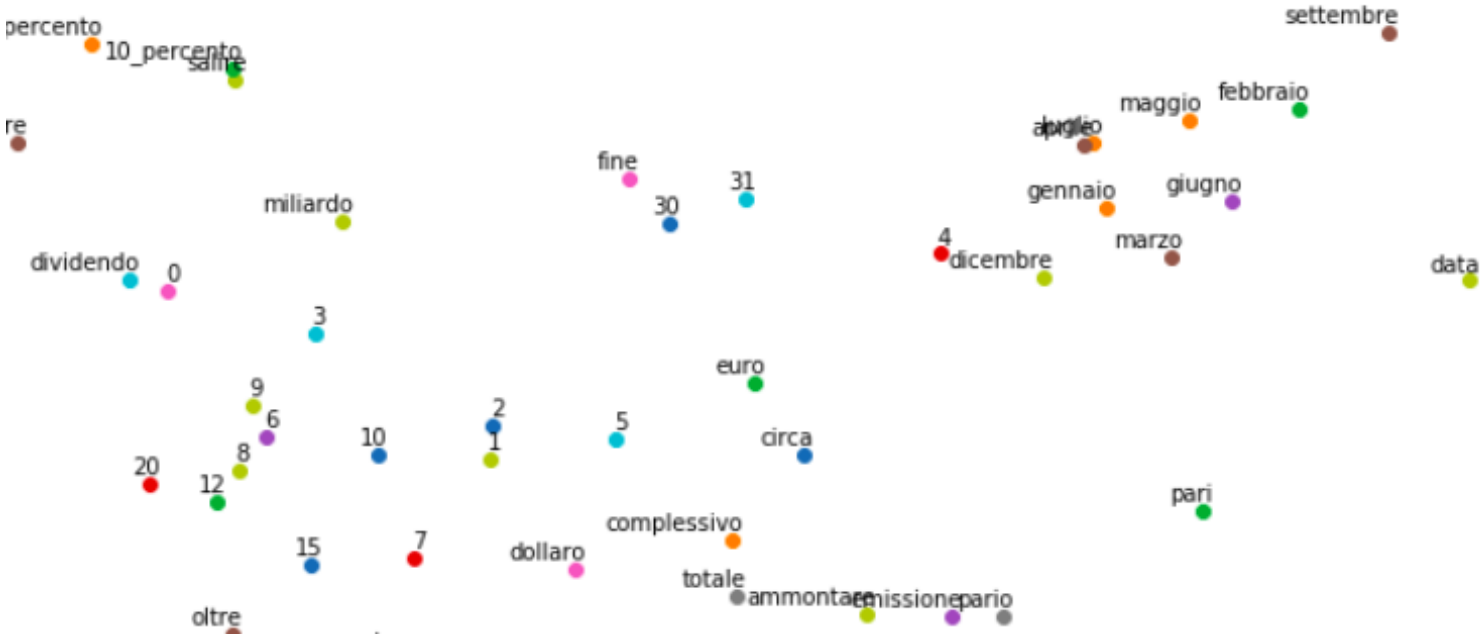
Gli **word embeddings** ottenuti tramite **word2vec** hanno la particolare qualità di catturare le relazioni tra i termini.
Per esempio il risultato della espressione:

$$vector(king) - vector(man) + vector(woman)$$

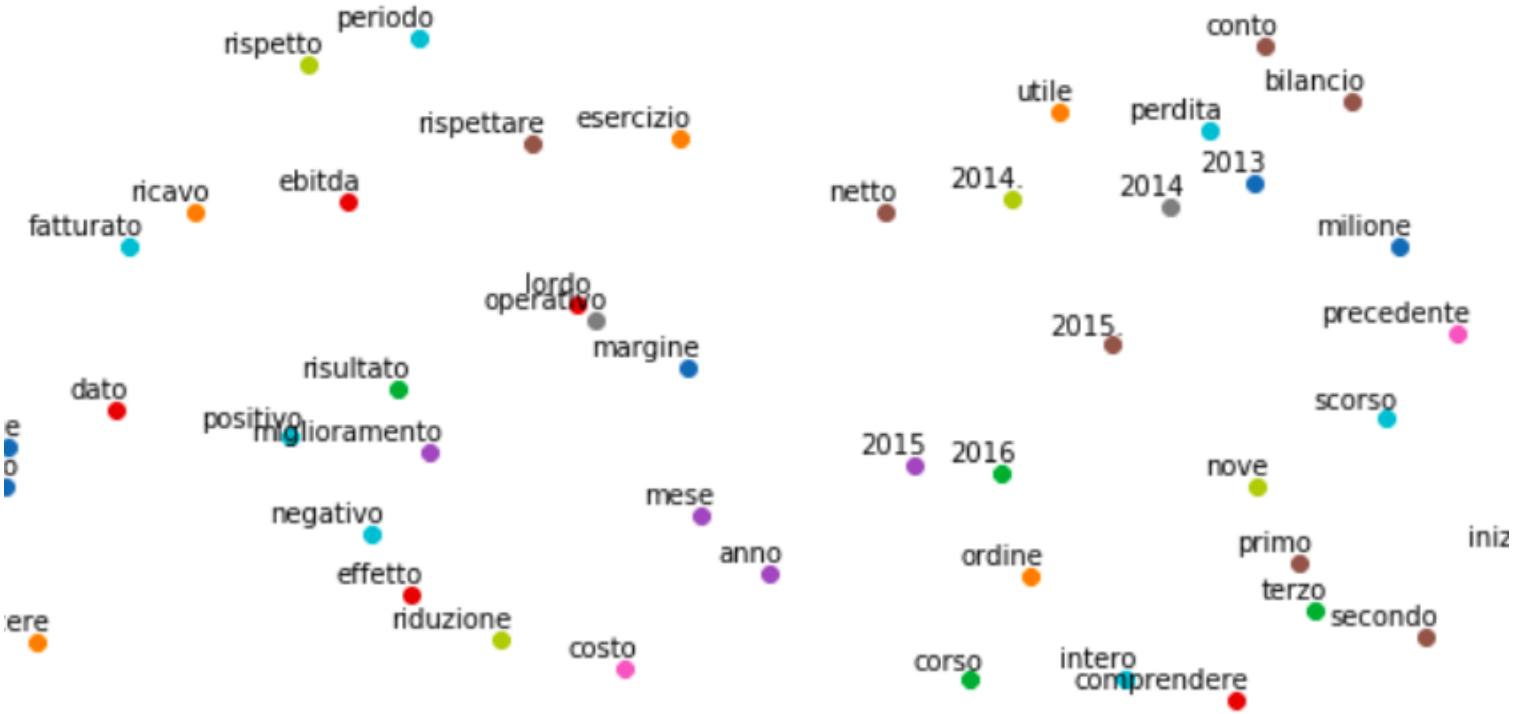
è un vettore **vicino** a $vector(queen)$.



Word2Vec - RadioCor e Sole 24 Ore - Dettaglio 1



Word2Vec - RadioCor e Sole 24 Ore - Dettaglio 2



Outline

Doc2Vec - Task

Doc2Vec è la naturale estensione di **Word2Vec** che, dato un corpus di documenti, permette di calcolarne la **similarità**.
È una **unsupervised task** che si basa su una **two-layers neural network** che viene trainata su un **corpus di documenti**.

- **Input: corpus** di testi di lunghezza arbitraria
- **Output:** document **vector-space**

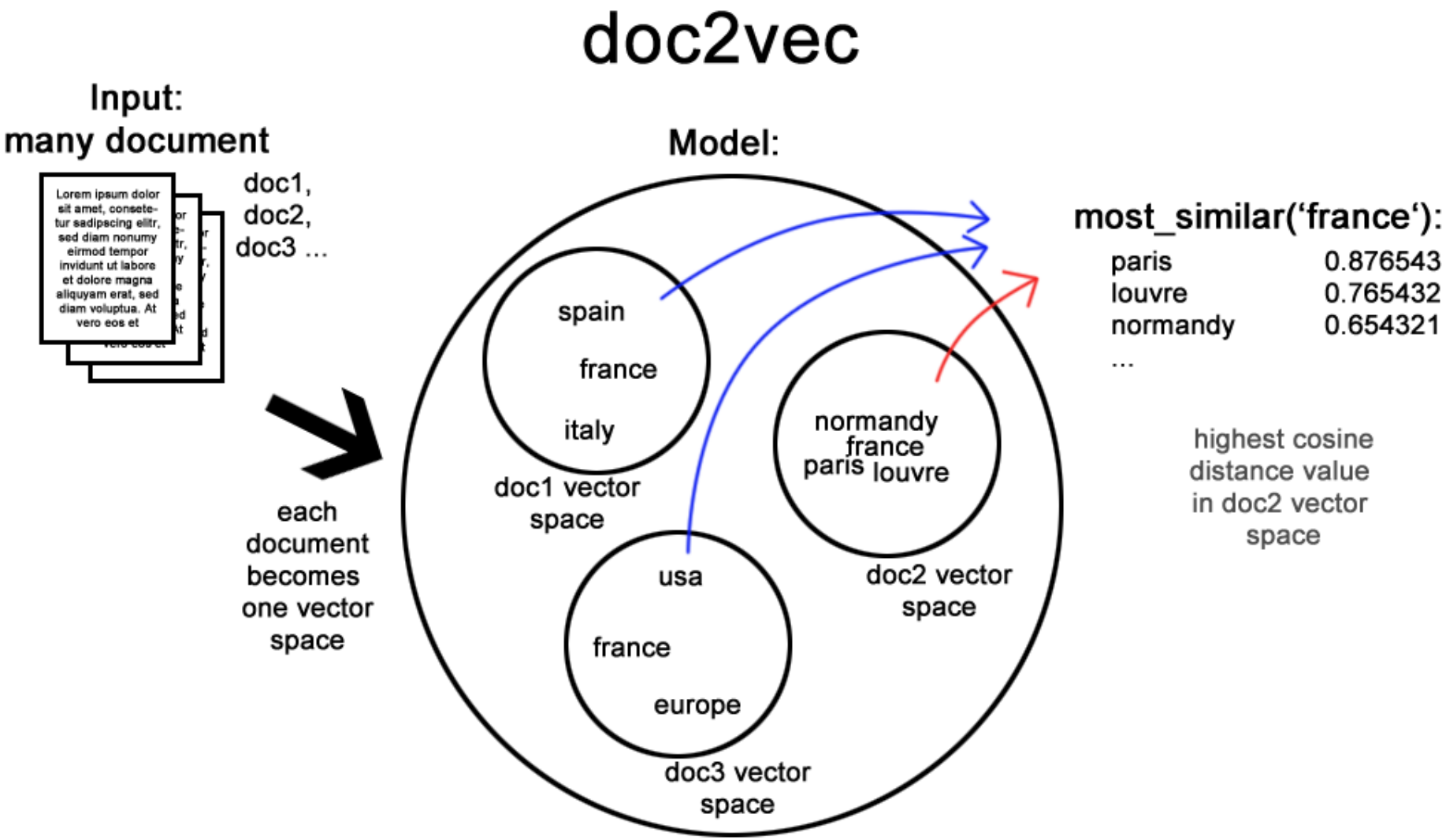
Doc2Vec - Notazione

- Vocabolario **V**: l'insieme di tutte le parole contenute in tutti i documenti del corpus.
- Paragraph Matrix **D**: la matrice contenente gli OHE **Paragraph Vector**s.
- Word Matrix **W**: la matrice contenente gli OHE **Word Vector**s.

Doc2Vec - Paragraph Vector

- Il **Paragraph Vector** può essere visto come una altra parola che agisce da memoria del contesto corrente. È unico per ogni documento ma è condiviso dai contesti dello stesso documento.
- Gli **Word Vector**s sono condivisi tra tutti i documenti.

Doc2Vec - Idea

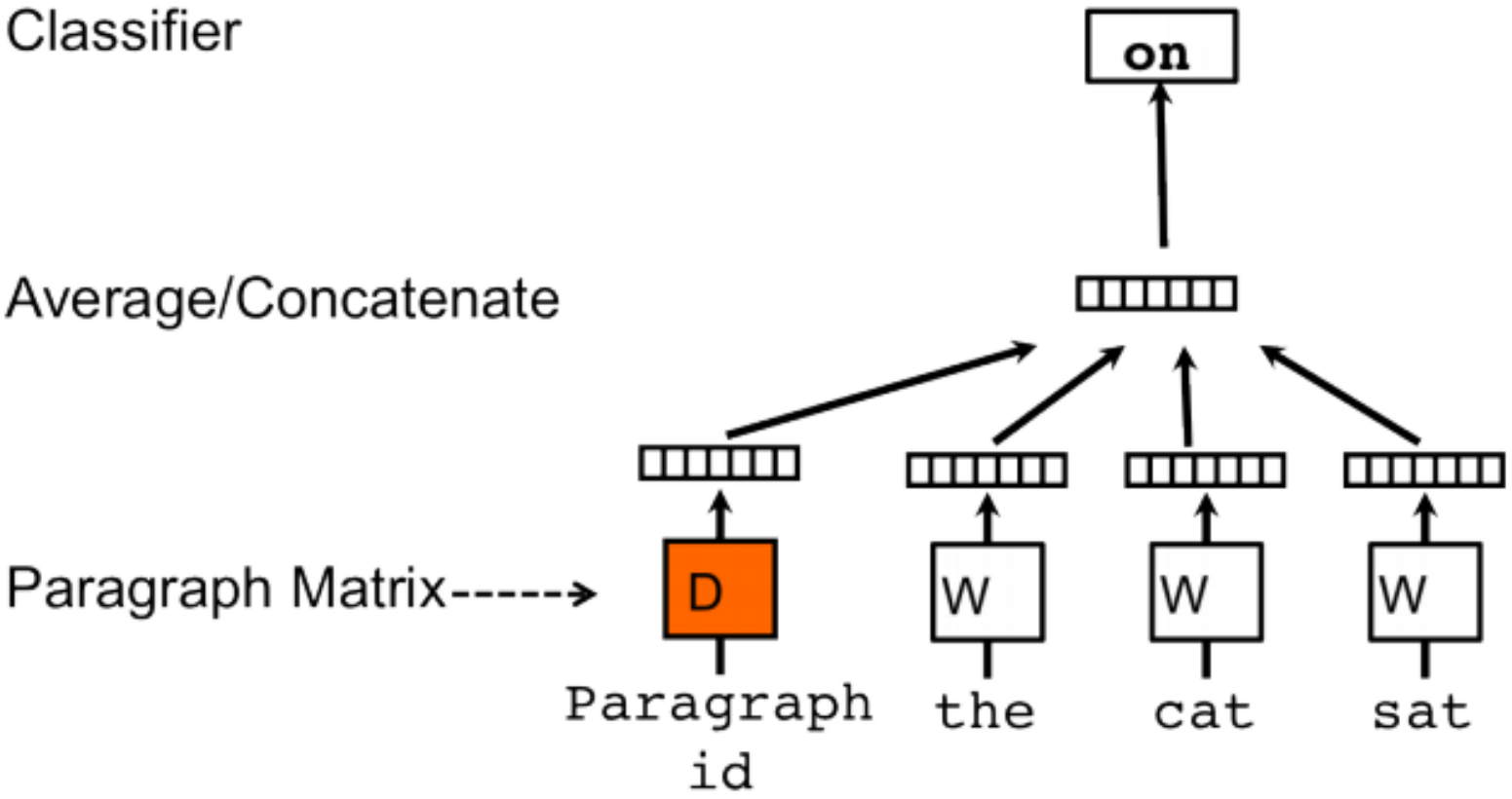


Doc2Vec - (PV-DM) Model

Una implementazione di **Doc2Vec** viene chiamata **Distributed Memory Model of Paragraph Vectors (PV-DM)**.
Si ispira a **Word2Vec CBOW**, quindi si basa su una **context window** che viene fatta scorrere in ogni paragrafo e dalla quale vengono campionati i relativi **Word Vectors** $w_1, w_2, ..., w_c$.
Tuttavia, diversamente dal **CBOW**, la predizione della **next word** w_{c+1} viene fatta considerando **le parole precedenti + Paragraph Vector**.
La struttura è la seguente:

- **Input:** $w_1, w_2, ..., w_c$ parole nel paragrafo k e il **Paragraph Vector** p_k rappresentante lo specifico paragrafo k.
- **Target:** è la prossima parola w_{c+1}

Doc2Vec - (PV-DM)dal corpus Model



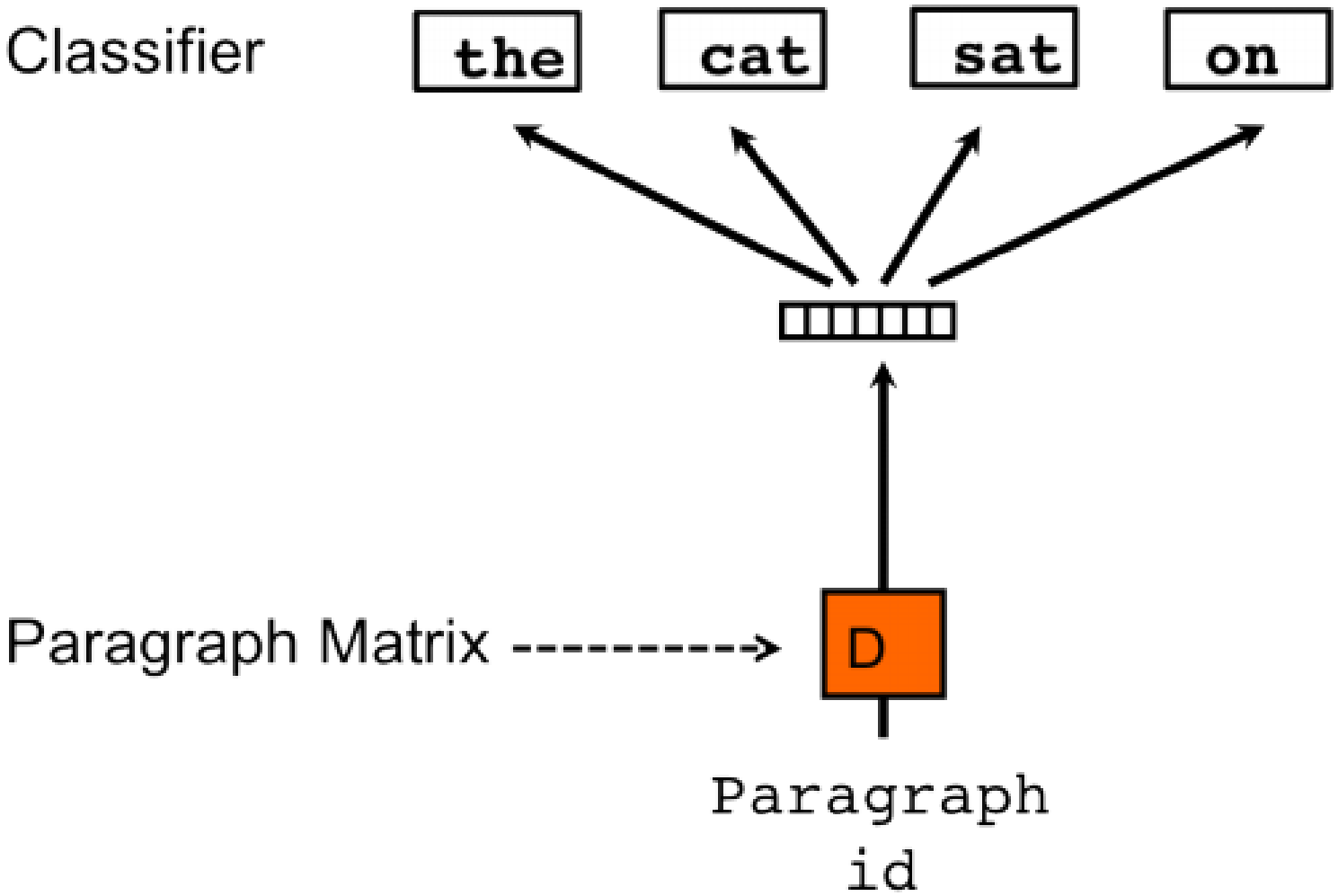
Gli **Word Vectors** e il **Paragraph Vector** vengono concatenati o averaged.

Doc2Vec - (PV-DBOW) Model

Una altra implementazione di **Doc2Vec** viene chiamata **Distributed Bag Of Words version of Paragraph Vector (PV-DBOW)**.
Si ispira a **Word2Vec Skip Grammar** e la struttura è la seguente:

- **Input: Paragraph Vector** p_k rappresentante lo specifico paragrafo k.
- **Target:** $w_1, w_2, ...w_c$, parole del paragrafo, dove $\mathrm{C} = \textit{windowsize}$

Doc2Vec - (PV-DBOW) Model



Doc2Vec - Steps

L'algoritmo si sviluppa in due **main phases**:

1. **Training stage**: ottenere dal corpus gli **Word Vectors W**, le **softmax weight U**, il **bias b** e **Paragraph Vectors D**.
2. **Inference stage**: tramite **Gradient Descend** vengono aggiunti alla matrice **D** i **Paragraph Vectors** relativi a nuovi documenti. Durante questa fase vengono tenuti costanti i parametri **W, U, b** ottenuti nel **Training stage**.

Doc2Vec - Applicazioni

Alla fine della **Training phase** otteniamo un **modello geometrico** per gli **input documents** in grado anche di generare una rappresentazione vettoriale per nuovi documenti non presenti nel training set.
A partire da questo modello si possono applicare **on top** algoritmi per task di **classificazione** e/o **clustering**.

Doc2Vec - Performance

Considerando che non possediamo un grado di similarità tra ogni coppia di documenti, abbiamo deciso di valutare le performance del nostro modello in base alla **self-similarity**.
La accuracy è il **ratio** di documenti del corpus che hanno se stesso tra i 3 più simili, ovvero:

$$accuracy = \frac{\sum_{i=1}^{D} s_i}{D} \quad \text{dove } s_i = \begin{cases} 1 & \text{if } d_i \in \mathrm{model.top3similar(d_i)} \\ 0 & \text{otherwise} \end{cases}$$

$\mathrm{model.top3similar(d_i)}$ = i 3 documenti più simili a d_i

Doc2Vec - Future work

Per trovare i parametri ottimi di **Doc2Vec** è necessario disporre di un ulteriore dataset, preparato da un esperto di dominio, che esponga il grado di similarità **reale** tra le coppie di documenti.
In questo modo si può anche validare la capacità di **generalizzazione** del modello.

References

1. **Efficient Estimation of Word Representations in Vector Space:** presentazione Skip Gram e CBOW models
2. **Distributed Representations of Words and Phrases and their Compositionality:** Word2Vec
3. **Paragraph Vectors**
4. **Document Embedding with Paragraph Vectors:** Doc2Vec