

# Document Semantic Similarity

## TIS Project

Alberto Pirovano   Francesco Picciotti

Politecnico di Milano

8th May 2017



## 1 State of art

- NLP tradizionale
- Vector Space Model
- Deep Learning

## 2 Data Preparation

- Preprocessing
- Cleaning del testo

## 3 Word2Vec

## 4 Doc2Vec



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



- 1 State of art
  - NLP tradizionale
  - Vector Space Model
  - Deep Learning
- 2 Data Preparation
  - Preprocessing
  - Cleaning del testo
- 3 Word2Vec
- 4 Doc2Vec



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

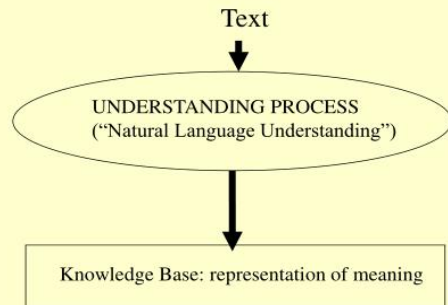
- Cleaning dei dati
- Pos-Tagging
- Stemming o Lemmatisation
- Parsing
- Ontologia

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

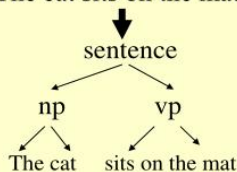
- **L'affidabilità** del Pos-Tagger italiano di TreeTagger
- **Reperire** una Ontologia e un parsing tool per la lingua italiana



## NLP: the process



"The cat sits on the mat"



↓

Fact(type: statement,  
agent: cat-002,  
action: sits\_on,  
object: mat-001)

Fact(type: statement,  
agent: Fido,  
action: is\_a,  
object: cat)

Fact(type: statement,  
agent: Freda,  
action: loves,  
object: Fido)

Natural Language Processing

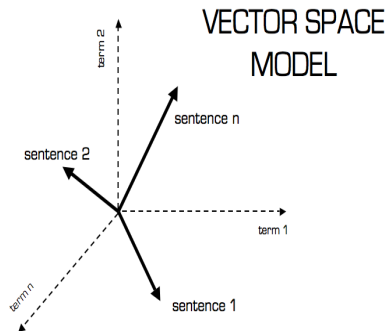


- 1 State of art
  - NLP tradizionale
  - **Vector Space Model**
  - Deep Learning
- 2 Data Preparation
  - Preprocessing
  - Cleaning del testo
- 3 Word2Vec
- 4 Doc2Vec



# 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**.





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**  $d_j$ :

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



**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 corpora.

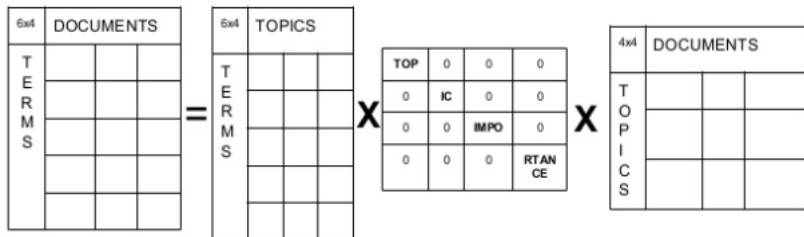
Il cuore di questa procedura sta nella **riduzione della dimensionalità** di questa matrice tramite **SVD**.



# 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.

Find three matrices  $U$ ,  $\Sigma$  and  $V$  so that:  $X = U\Sigma V^T$



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



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:

LSA is essentially low-rank *approximation* of document term-matrix

Word assignment to topics

IT cars

3	4	1	0
4	3	0	1
3	4	4	3
0	1	4	3
2	0	3	3
0	1	3	4

=

	linux	-0.33	-0.53
	modem	-0.32	-0.54
	the	-0.62	-0.10
	clutch	-0.38	0.42
	steering	-0.36	0.25
	petrol	-0.37	0.42

X

Topic importance

11.4	
	6.27

X

Topic distribution across documents

	D1	D2	D3	D4
IT	-0.42	-0.48	-0.57	-0.51
cars	-0.56	-0.52	0.45	0.46

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

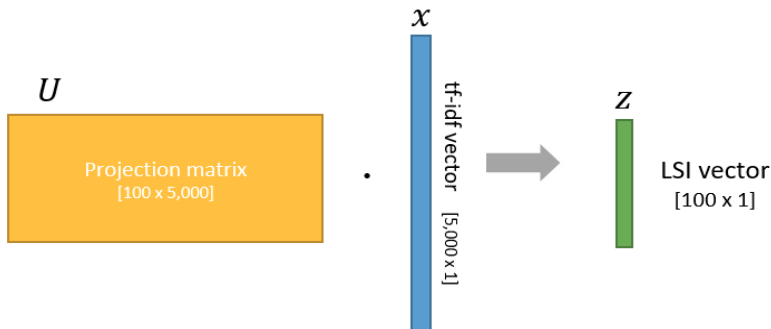


# Vector Space Model - LSA - Example

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  $\mathbf{z}$ , dobbiamo:

- 1 **Vettorizzare** un documento in  $\mathbf{x}$ .
- 2 **Proiettare** il **tfidf** vector  $\mathbf{x}$  sul topic space.

Se definiamo la matrice  $U = (\text{Wordassignment to topics})^T$  possiamo visualizzare la **proiezione** in questo modo:



# Vector Space Model - Algorithm

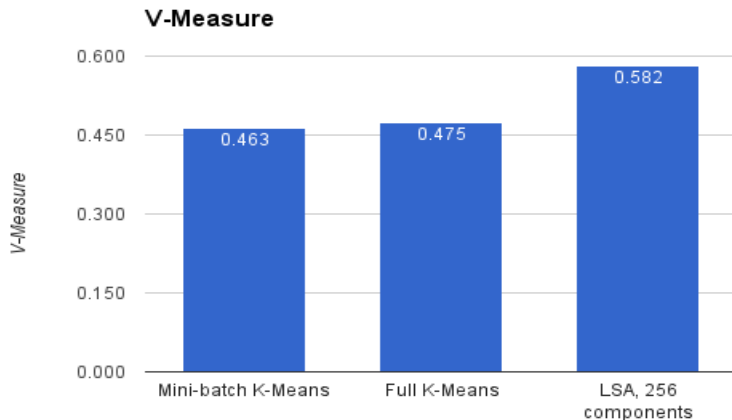
Riassumendo, possiamo individuare i seguenti **passaggi**:

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

- 1 State of art
  - NLP tradizionale
  - Vector Space Model
  - Deep Learning
- 2 Data Preparation
  - Preprocessing
  - Cleaning del testo
- 3 Word2Vec
- 4 Doc2Vec



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:

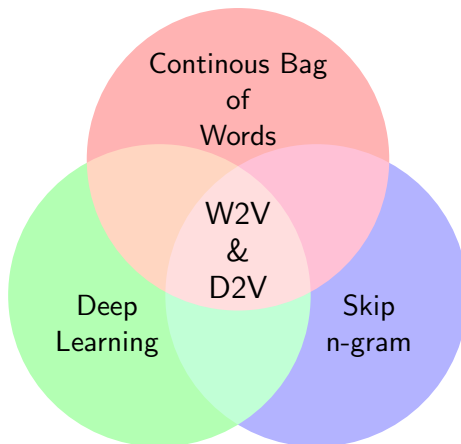
- **Continuous Bag Of Words model**
- **Skip Grammar model**

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 (corpora), 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 dipendente da un preprocessing
- Combina il metodo *Geometrico* con quello *NLP Tradizionale*
- **Non sfrutta una ontologia, ma la crea**
- **Language independent**
- **Tiene in considerazione anche l'ordine delle parole, Context-aware**

## Cons:

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



- 1 State of art
  - NLP tradizionale
  - Vector Space Model
  - Deep Learning
- 2 Data Preparation
  - Preprocessing
  - Cleaning del testo
- 3 Word2Vec
- 4 Doc2Vec



*"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, cioè ci sono 867 articoli duplicati.

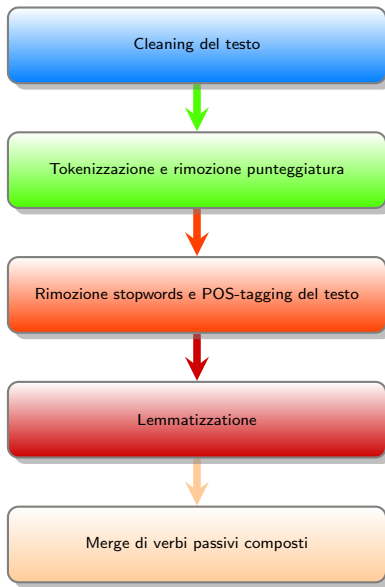




- 1 State of art
  - NLP tradizionale
  - Vector Space Model
  - Deep Learning
- 2 Data Preparation
  - Preprocessing
  - Cleaning del testo
- 3 Word2Vec
- 4 Doc2Vec



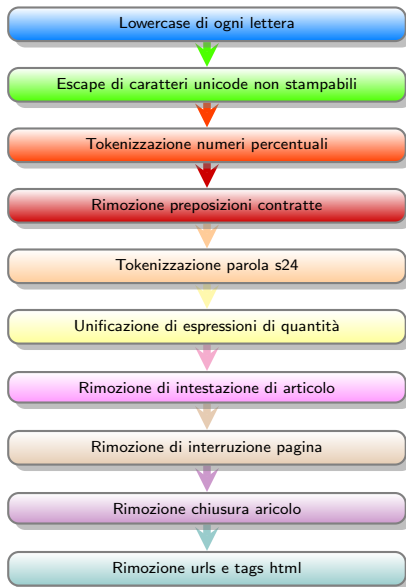
# Pipeline completa



- 1 State of art
  - NLP tradizionale
  - Vector Space Model
  - Deep Learning
- 2 Data Preparation
  - Preprocessing
  - Cleaning del testo
- 3 Word2Vec
- 4 Doc2Vec



# Cleaning pipeline



- 1 State of art
  - NLP tradizionale
  - Vector Space Model
  - Deep Learning
- 2 Data Preparation
  - Preprocessing
  - Cleaning del testo
- 3 Word2Vec
- 4 Doc2Vec



**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 - 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 **Continuous 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 ).



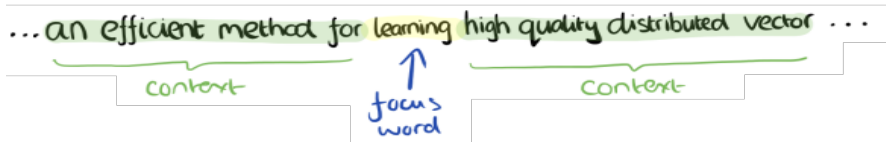
Gli **input-target pairs** sono rappresentati tramite **One Hot Encoding**, mentre l'**output** della NN sarà un vettore di probabilità di dimensione pari al numero di parole del **vocabolario**.

Di conseguenza se abbiamo un vocabolario di **dimensionalità  $V$**  avremo dei **vectors** costituiti da  $V$  elementi di cui un 1 e gli altri 0.





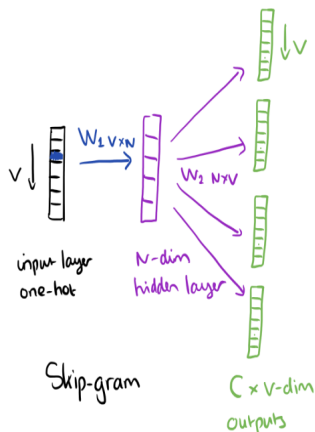
In entrambi i casi è necessario definire un **context** per la parola in input, che identifichiamo con un parametro chiamato  $c = \text{window size}$ . Tramite questa *window size*, una volta fissata una **input word/focus word**, chiamiamo **context** l'insieme delle  $c$  parole prima della input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





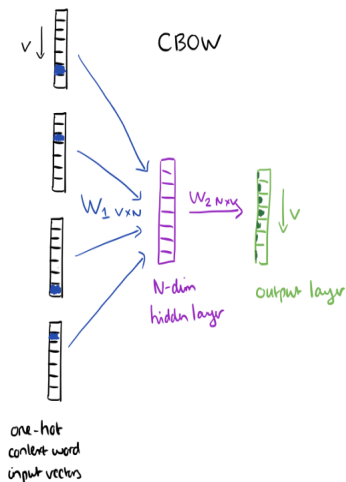
Data una sequenza di **training words**  $w_1, w_2, \dots, w_M$ , con  $M = \text{trainingsetsize}$ , il **training objective** di questo modello è trovare il parametro  $\theta$  che massimizza la **Log-Likelihood**:

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

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**.





Data una sequenza di **training words**  $w_1, w_2, \dots, w_M$ , con  $M = \text{trainingsetsize}$ , il **training objective** di questo modello è trovare il parametro  $\theta$  che massimizza la **Log-Likelihood**:

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

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

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



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).

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

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 ( $v_{wAverage}$ ).



Considerando i due tipi di approccio, possiamo quindi esplicitare le **probabilità scalari** utilizzate nella funzione di costo, con  $w_m = w_{Input}$ :

- **Skip Grammar:** 
$$p(w_{m+j} | w_m) = \frac{\exp(v'_{w_{m+j}} \cdot v_{w_{Input}}^T)}{\sum_{i=1}^V \exp(v'_{w_i} \cdot v_{w_{Input}}^T)}$$
- **CBOW:** 
$$p(w_m | w_{m-c}, \dots, w_{m+c}) = \frac{\exp(v'_{w_m} \cdot v_{w_{Average}}^T)}{\sum_{i=1}^V \exp(v'_{w_i} \cdot v_{w_{Average}}^T)}$$

Tecnicamente il **vettore di probabilità** che viene generato dalla rete, dato un **input vector** o un **input average**, è:

- **Skip Grammar:** 
$$y_{t+j} = \frac{\exp(W_2^T \cdot v_{w_{Input}}^T)}{\sum_{i=1}^V \exp(v'_{w_i} \cdot v_{w_{Input}}^T)}$$
- **CBOW:** 
$$y_t = \frac{\exp(W_2^T \cdot v_{w_{Average}}^T)}{\sum_{i=1}^V \exp(v'_{w_i} \cdot v_{w_{Average}}^T)}$$



# Word2Vec - Output layer

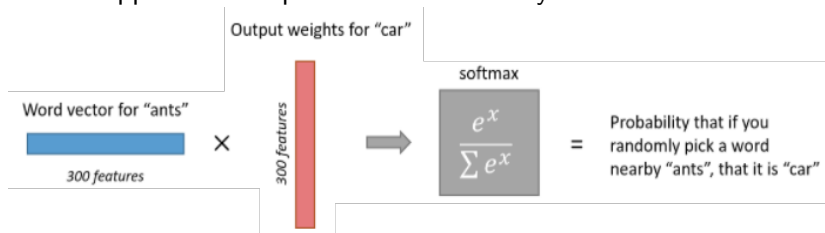
L'**output layer** si compone di due operazioni:

- 1 Calcolare lo score dell'input vector,  $scoreVec = v_{wInput}^T \cdot W_2$  per **Skip Grammar** o  $scoreVec = v_{wAverage}^T \cdot W_2$  per **CBOW**.

- 2 **Multi-class classification** attraverso **softmax**,

$$softmax(scoreVec) = \frac{e^{scoreVec}}{\sum_{i=1}^V e^{scoreVec_i}}$$

Dove  $V = sizevocabulary$ ,  $scoreVec_i$  è l'i-esimo componente di  $scoreVec$  e  $e^{vector}$  rappresenta l'esponenziale element by element.



Il **loss** viene definito come **una media** dei **loss relativi** ad ogni **training sample**, nel nostro caso un training sample è una **parola nel corpus**.

$$L = \frac{1}{M} \cdot \sum_{m=1}^M L_m$$

L'espressione del loss in questi termini serve solo a **visualizzare** il suo valore durante le **iterazioni dell'algoritmo**.

Un **epoch** (iterazione) della procedura di training consiste nell'effettuare per ogni **training sample** la procedura di **forward e back propagation**. Per ogni sample in un **epoch** vengono quindi aggiornate tutte le **weights** del modello.

Ad ogni **epoch** la massimizzazione del **loss** sarà effettuata considerando il loss  $L_m$  relativo a sample che sta attraversando il network.





# Word2Vec - Parameters Learning

Cominciamo considerando il caso di **CBOW** in cui abbiamo una sola parola nel contesto, vale a dire il modello dovrà predire la **focus word** data **una** parola nel **contesto**.

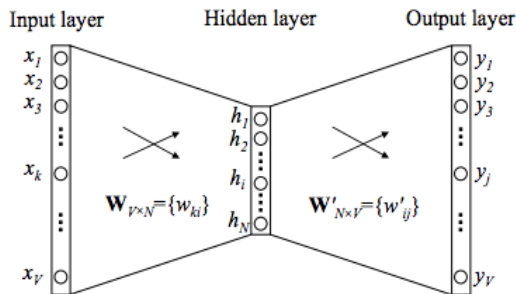


Figure 1: A simple CBOW model with only one word in the context

In questa figura  $\mathbf{W} = \mathbf{W}_1$  e  $\mathbf{W}' = \mathbf{W}_2$ .



Considerando il **training sample  $m$**  che attraversa il network e ipotizzando  $w_{m-1}$  l'unica parola nel contesto di  $w_m$ ;

$$\begin{aligned} L_m &= \log p(w_m | w_{context}) = \log \left( \frac{\exp(v'_{w_m} \cdot v_{w_{context}}^T)}{\sum_{i=1}^V \exp(v'_{w_i} \cdot v_{w_{context}}^T)} \right) = \\ &= \log \left( \frac{\exp(v'_{w_m} \cdot h)}{\sum_{i=1}^V \exp(v'_{w_i} \cdot h)} \right) = \log \left( \frac{\exp(u_m)}{\sum_{i=1}^V \exp(u_i)} \right) = \\ &= \log(\exp(u_m)) - \log \left( \sum_{i=1}^V \exp(u_i) \right) = \\ &= u_m - \log \left( \sum_{i=1}^V \exp(u_i) \right) = -E \end{aligned}$$

Quindi l'errore commesso per **UN SOLO TRAINING SAMPLE** è:

$$E = -\log p(w_m | w_{m-1})$$



Chiamiamo:

- $u_i$  con  $i \in 1, 2, \dots, V$  lo score calcolato per la parola  $i$ .
- $h_i$  con  $i \in 1, 2, \dots, N$  l'output del  $i$ -th neurone dell'hidden layer.
- $y_i$  con  $i \in 1, 2, \dots, V$  l'output del  $i$ -th neurone dell'output layer.

Considerando un **OHE** vector in input  $x_{context}$  di dimensioni  $[V, 1]$  con il  $k$ -th elemento  $x_k \neq 0$ , definiamo:

- $h = \sigma(W^T \cdot x_{context}) = W^T \cdot x_{context} = v_{w_{context}}^T$ , vale a dire che la funzione di attivazione per il primo layer è la funzione identità (**lookup table**).
- $h_i = \sum_{k=1}^V x_k \cdot w_{ki}$
- $u_j = v_{w_j}'^T \cdot h$
- $y_j = p(w_j | w_{context}) = \frac{\exp(u_j)}{\sum_{i=1}^V \exp(u_i)}$ , la probabilità a posteriori che  $w_j \in V$  sia la parola campionata nel contesto di  $w_{context}$ .



# Word2Vec - Parameters Learning - $W'$ ( $W_2$ )

$$\begin{aligned}\frac{\partial E}{\partial u_j} &= \frac{\partial \left( \log \left( \sum_{i=1}^V \exp(u_i) \right) - u_m \right)}{\partial u_j} = \\&= \frac{\partial \left( \log \left( \sum_{i=1}^V \exp(u_i) \right) \right)}{\partial u_j} - \frac{\partial (u_m)}{\partial u_j} = \\&= \frac{\exp(u_j)}{\sum_{i=1}^V \exp(u_i)} - t_j = \\&= y_j - t_j = e_j\end{aligned}$$

$$\text{dove } t_j = \begin{cases} 1 & \text{if } j = m \\ 0 & \text{if } j \neq m \end{cases}$$

$$\begin{aligned}\frac{\partial u_j}{\partial w'_{ij}} &= \frac{\partial \left( v_{w_j}'^T \cdot h \right)}{\partial w'_{ij}} = h_i \\ \frac{\partial E}{\partial w'_{ij}} &= \frac{\partial E}{\partial u_j} \cdot \frac{\partial u_j}{\partial w'_{ij}} = e_j \cdot h_i\end{aligned}$$



L'aggiornamento dei parametri quindi risulta:

$$\begin{aligned}w'_{ij}(t+1) &= w'_{ij}(t) - \eta \cdot h_i \cdot e_j & i \in 1, 2, \dots, N \wedge j \in 1, 2, \dots, V \\v'_{w_j}(t+1) &= v'_{w_j}(t) - \eta \cdot h \cdot e_j & j \in 1, 2, \dots, V\end{aligned}$$

**Di conseguenza ad ogni ciclo vengono aggiornati tutti i pesi della matrice  $W_1$ .**



$$\begin{aligned}\frac{\partial E}{\partial h_i} &= \sum_{j=1}^V \left( \frac{\partial E}{\partial u_j} \cdot \frac{\partial u_j}{\partial h_i} \right) = \sum_{j=1}^V \left( e_j \cdot \frac{\partial u_j}{\partial h_i} \right) = \\ &= \sum_{j=1}^V (e_j \cdot w'_{ij}) = e \cdot (W_{2i})^T\end{aligned}$$

Dove  $W_{2i}$  è la  $i$ -esima riga di  $W_2$  ( $[1, V]$ ), che trasposta da un vettore  $[V, 1]$ , il quale, moltiplicato per un vettore  $[1, V]$  da uno scalare.

$$\begin{aligned}\frac{\partial h_i}{\partial w_{ki}} &= x_k \\ \frac{\partial L_m}{\partial w_{ki}} &= \frac{\partial L_m}{\partial h_i} \cdot \frac{\partial h_i}{\partial w_{ki}} = x_k \cdot e \cdot (W_{2i})^T\end{aligned}$$



# Word2Vec - Parameters Learning - $W$ ( $W_1$ )

L'aggiornamento dei parametri quindi risulta:

- $w_{ki}^{(t+1)} = w_{ki}^{(t)} - \eta \cdot x_k \cdot e \cdot (W_{2i})^T$ , con  $k \in 1, 2, \dots, V$  e  $i \in 1, 2, \dots, N$
- $W_1^{(t+1)} = W_1^{(t)} - \eta \cdot x \cdot (e \cdot W_2^T)$

Dove  $x \cdot (e \cdot W_2^T)$  è di dimensionalità  $[V, N]$ .

Considerando poi che  $x$  è un **OHE vector**, solo un coefficiente di  $x$  sarà  $\neq 0$ . Di conseguenza questa operazione di aggiornamento lascia inalterate tutte le **input representation** della matrice  $W_1$  tranne quella corrispondente all'indice  $x_k \neq 0$  di  $x$ .

Quindi l'unico aggiornamento che avviene è:

$$\bullet v_{W_{context}}^{(t+1)} = v_{W_{context}}^{(t)} + \eta \cdot (e \cdot W_2^T)$$

Di conseguenza vediamo una delle particolarità di **word2vec**, vale a dire che **per ogni parola in input l'aggiornamento di  $W_1$  si limita solo al word vector relativo alla input word, in questo caso di**

$v_{W_{context}}$ .



# Word2Vec - Parameters Learning - Batch GD

Invece di considerare un solo sample per volta possiamo anche considerare un **batch**  $X$  di  $M$  samples:

- $X$  = una **OHE** riga per ogni **input sample** in  $M$ ,  $[M, V]$
- $T$  = una **OHE** riga per ogni **target sample** in  $M$ ,  $[M, V]$

La **forward pass** sarà:

- $H = X \cdot W_1$   $[M, N]$
- $U = H \cdot W_2$ ,  $[M, V]$
- $Y = \text{softmax}(U)$ ,  $[M, V]$ , element-wise

La **back pass** sarà:

- $\frac{\partial E}{\partial U} = \frac{1}{M} \cdot (Y - T)$ ,  $[M, V]$
- $\frac{\partial E}{\partial W_2} = H^T \cdot \frac{\partial E}{\partial U}$ ,  $[N, V]$
- $\frac{\partial E}{\partial H} = \frac{\partial E}{\partial U} \cdot W_2^T$ ,  $[M, N]$
- $\frac{\partial E}{\partial W_1} = X^T \cdot \frac{\partial E}{\partial H}$ ,  $[V, N]$





Considerando il modello **CBOW** con multiple parole nel contesto della parola  $w_m$  avremo pochi cambiamenti rispetto al modello appena spiegato.

- $$h = \frac{1}{C} \cdot W_1^T \cdot (x_{m-C} + \dots + x_{m-2} + x_{m-1} + x_{m+1} + x_{m+2} + \dots + x_{m+C})$$

Di conseguenza il vettore  $h$  diventa uno **pseudo-OHE**, cioè un vettore che ha dimensione del vocabolario e un numero di 1 pari al numero di parole nel contesto.

La procedura si svolge nello stesso modo, tranne per quanto riguarda l'aggiornamento dei pesi della matrice  $W_1$ , infatti ora aggiorno un numero di **input representation** pari al numero della parole nel contesto:

- $$v_{w_{m-c}}^{(t+1)} = v_{w_{m-c}}^{(t)} + \eta \cdot \frac{1}{C} \cdot (e \cdot W_2^T), \text{ con } c \in C, \dots, -C \text{ e } c \neq 0$$

Per quanto riguarda  $W_2$  invece non cambia niente:

- $$v'_{w_j(t+1)} = v'_{w_j(t)} + \eta \cdot h \cdot e_j, \text{ con } j \in 1, 2, \dots, V$$



# Word2Vec - Parameters Learning - Skip Gram - $L_m$

Considerando il **training sample**  $\mathbf{m}$  che attraversa il network e ipotizzando  $w_{m-1}$  l'unica parola nel contesto di  $w_m$

$$\begin{aligned} L_m &= \sum_{-c < j < c, j \neq 0} \log p(w_{m+j} | w_m) = \sum_{-c < j < c, j \neq 0} \frac{\exp(v'_{w_{m+j}} \cdot v_{w_m}^T)}{\sum_{i=1}^V \exp(v'_{w_i} \cdot v_{w_m}^T)} \\ &= \sum_{-c < j < c, j \neq 0} \frac{\exp(u_{m+j})}{\sum_{i=1}^V \exp(u_i)} = \\ &= \sum_{-c < j < c, j \neq 0} \log \exp(u_{m+j}) - \sum_{-c < j < c, j \neq 0} \log \sum_{i=1}^V \exp(u_i) = \\ &= \sum_{-c < j < c, j \neq 0} u_{m+j} - C \cdot \log \sum_{i=1}^V \exp(u_i) = -E \end{aligned}$$

Quindi l'errore commesso per **UN SOLO TRAINING SAMPLE** è:

$$E = - \sum_{-c < j < c, j \neq 0} \log p(w_{m+j} | w_m)$$



# Word2Vec - Parameters Learning - Skip Gram - Forward Propagation

Chiamiamo:

- $u_{c,j'} = u_{j'} = v_{w_{j'}}'^T \cdot h$ , per ogni parola nel contesto di  $v_{w_m}$ , infatti gli output panels condividono la stessa matrice di weights  $W_2$ .
- $h = v_{w_m}^T$



# Word2Vec - Parameters Learning -Skip Gram - $W'$ ( $W_2$ )

$$\begin{aligned}\frac{\partial E}{\partial u_{m+c,j'}} &= \frac{\partial \left( C \cdot \log \sum_{i=1}^V \exp(u_i) - \sum_{-c < j < c, j \neq 0} u_{m+j} \right)}{\partial u_{m+c,j'}} = \\ &= \frac{\partial \left( C \cdot \log \left( \sum_{i=1}^V \exp(u_i) \right) \right)}{\partial u_{m+c,j'}} - \frac{\partial \left( \sum_{-c < j < c, j \neq 0} u_{m+j} \right)}{\partial u_{m+c,j'}} = \\ &= \frac{\partial \left( C \cdot \log \left( \sum_{i=1}^V \exp(u_i) \right) \right)}{\partial u_{m+c,j'}} - t_{m+c,j'} =\end{aligned}$$



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 ).



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.



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.



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.





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.



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.



Gli **word embeddings** ottenuti tramite **word2vec** hanno la particolare qualità di catturare le relazioni tra i termini.

Per esempio il risultato della espressione:

$$\text{vector}(\text{king}) - \text{vector}(\text{man}) + \text{vector}(\text{woman})$$

è un vettore **vicino** a  $\text{vector}(\text{queen})$ .



- 1 State of art
  - NLP tradizionale
  - Vector Space Model
  - Deep Learning
- 2 Data Preparation
  - Preprocessing
  - Cleaning del testo
- 3 Word2Vec
- 4 Doc2Vec



Qui spaghiamo per bene come funziona doc2vec

