Università degli studi di Modena e Reggio Emilia

Dipartimento di Ingegneria “Enzo Ferrari”
Corso di studi in Ingegneria Informatica

Valutazione di tecniche di Machine Learning per l’Entity Resolution

Relatore:
Sonia Bergamaschi

Candidato:
Stefano Gavioli

Anno Accademico 2017/2018

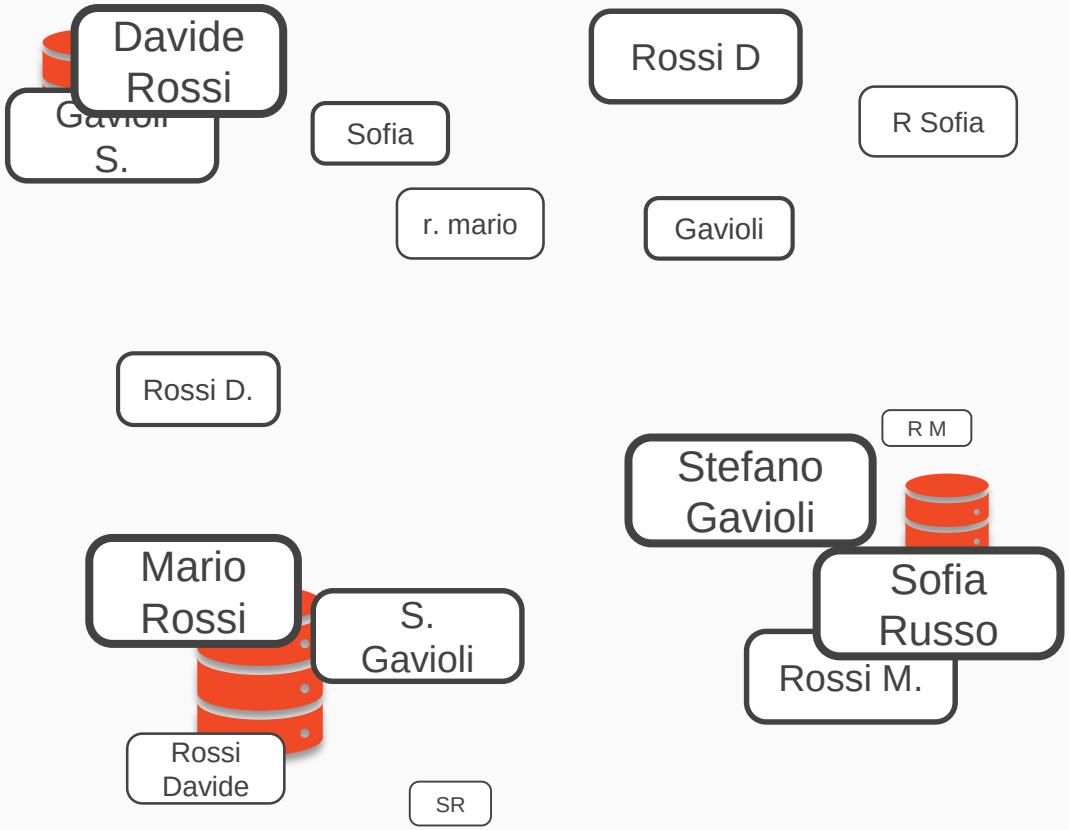
Outline

- Entity Resolution
- Algoritmi di Machine Learning
- Test

Entity Resolution

Attività che identifica due istanze di dati facenti riferimento alla stessa entità reale

Dati strutturati



Machine Learning

Possiamo far imparare ad una macchina il concetto di similarità tra record

Un programma è detto che *impara* da dell'**esperienza E** rispetto ad una qualche classe di **attività T** e **misure di performance P**, se la performance nelle attività T, in relazione alle misurazioni rispetto a P, migliora con l'esperienza E

Attività T: Identificare coppie di record associate alla stessa entità reale

Esperienza E: Coppie già identificate (supervised learning)

Misure P:

- **Precision** $= \frac{truepositives}{truepositive + falsepositives}$
- **Recall** $= \frac{truepositives}{truepositive + falsenegatives}$
- **F_1 score** $= 2 \cdot \frac{precision \cdot recall}{precision + recall}$

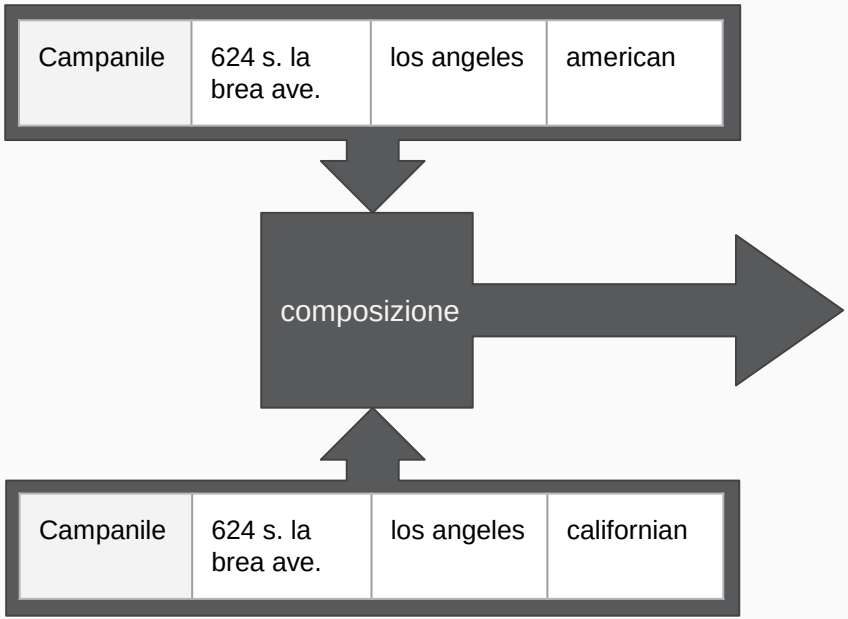
DeepER formula

I valori degli attributi dei record sono composti da un numero j di parole variabile da attributo ad attributo.

Composizione:

- Media
- LSTM

Esempio



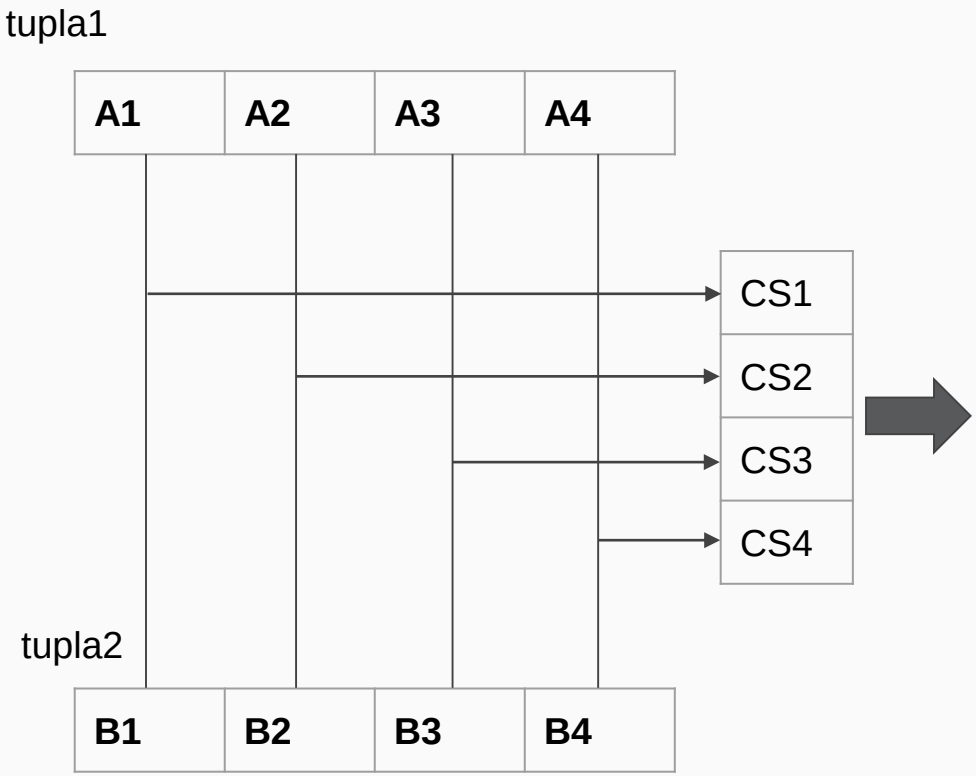
Media

Rappresentazione numerica di un valore di attributo di una tupla data dalla media delle rappresentazioni numeriche delle parole che lo compongono.

Similarità, *attribute wise*:

Cosine Similarity

$$= cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$



Word embedding

Come rappresentare una parola numericamente (vettorialmente)

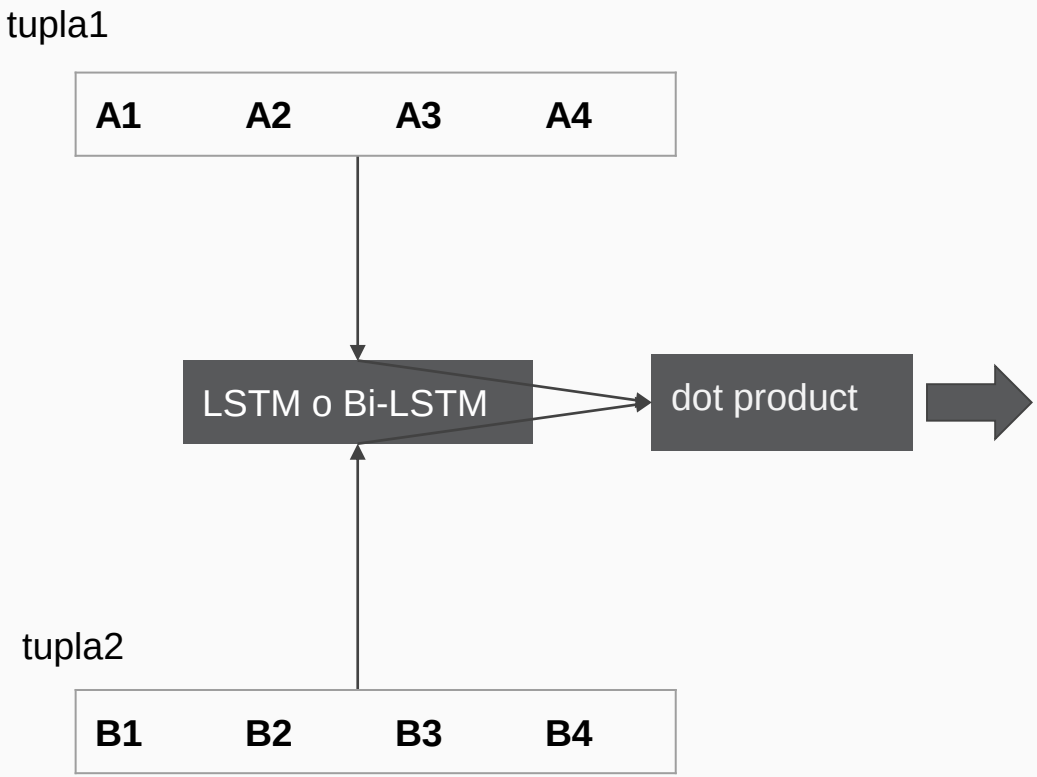
Dizionario GloVe

$$\begin{bmatrix} -0.337\ldots \\ -0.216\ldots \\ -0.006\ldots \\ \vdots \\ -0.522\ldots \\ 0.180\ldots \\ 0.642\ldots \end{bmatrix} \in \mathbb{R}^{300}$$

LSTM

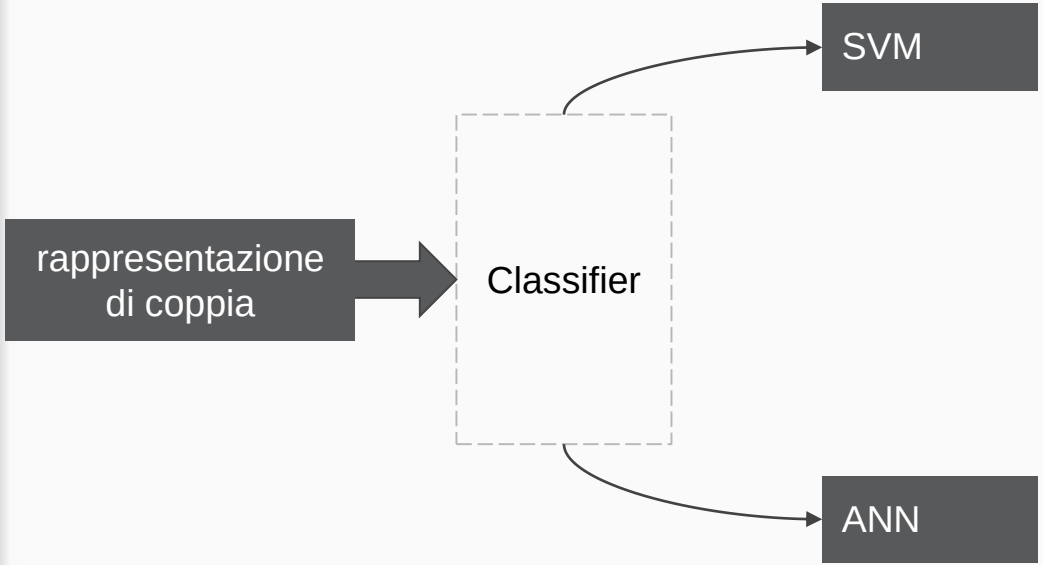
Rappresentazione numerica di una coppia di tuple data dal prodotto elemento per elemento del risultato di una rete ricorrente sul concatenamento dei valori degli attributi.

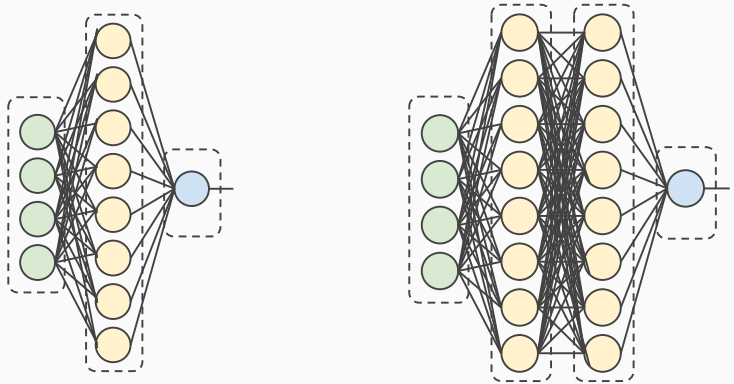
Similarità, *tuple wise*



Classificatore

- Support Vector Machine
- Rete neurale





Datasets

Fodor’s Zagat, DBLP ACM, Amazon Google

Descrizione

Test

Fodor’s Zagat

Insiemi di ristoranti

4 attributi:

- Nome
- Indirizzo
- Città
- Tipo

Dimensione del dataset: 2928 coppie

Nome	Indirizzo	Città	Tipo
arnie morton’s of chicago	435 s. la ciniega blv.	los angeles	american
lespinasse	2 e. 55th st.	new york	american
lespinasse (new york city)	2 e. 55th st.	new york city	asian
gianni’s	5 fulton st.	new york	seafood

DBLP ACM

Insiemi di riferimenti bibliografici

4 attributi:

- Titolo
- Autore
- Origine
- Anno

Dimensione del dataset: 21402 coppie

Titolo	The WASA2 object-oriented workflow management system	DOMINO: databases fOr MovINg Objects tracking	DOMINO: databases fOr MovINg Objects tracking
Autore	Gottfried VOssen, Mathias Weske	Ouri Wolfson, Prasad Sistla, Bo Xu, Jutai Zhou, Sam Chamberlain	utai Zhou, Ouri Wolfson, Sam Chamberlain, A.Prasad Sistla, Bo Xu
Origine	International Conference on Management of Data	International Conference on Management of Data	SIGMOD Conference
Anno	1999	1999	1999

Amazon Google

Insiemi di prodotti tratti dai relativi siti di e-commerce

4 attributi:

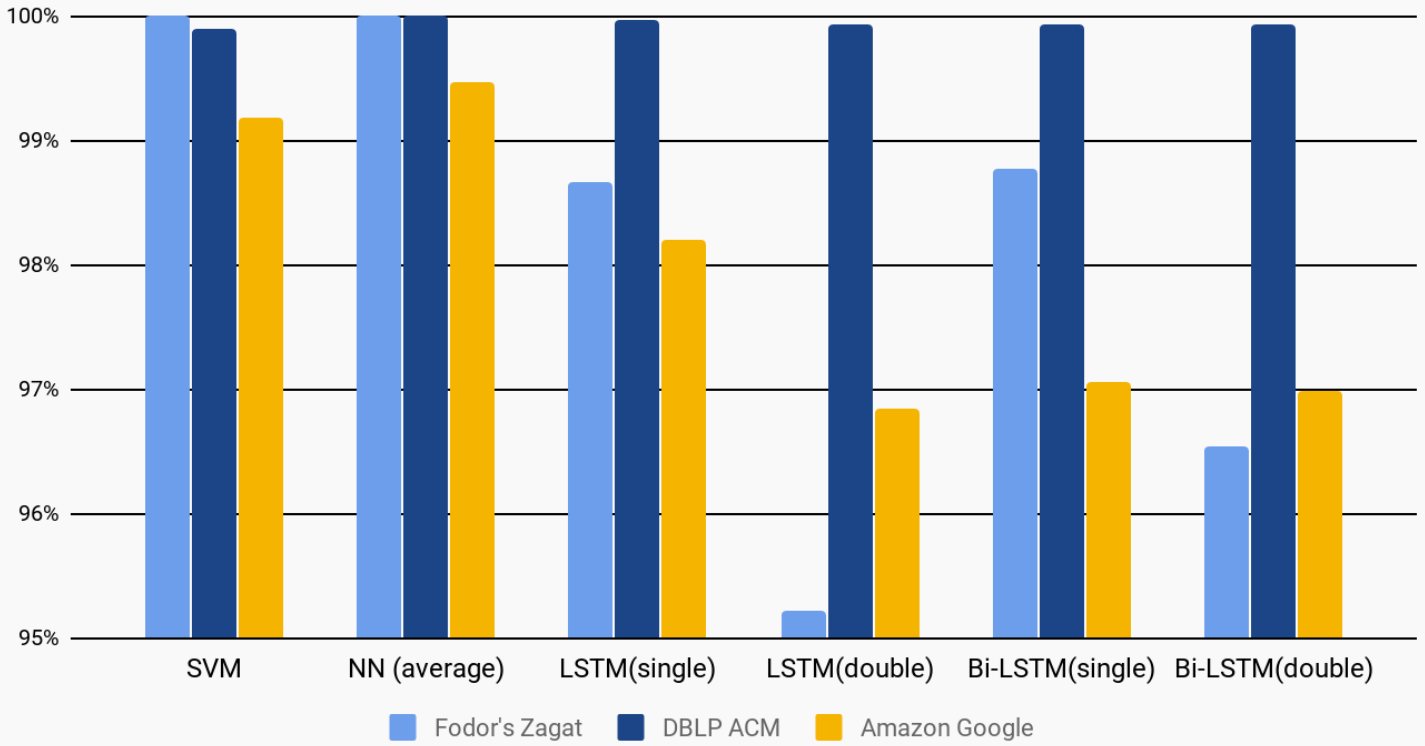
- Titolo
- Descrizione
- Produttore
- Prezzo

Dimensione del dataset: 17667 coppie

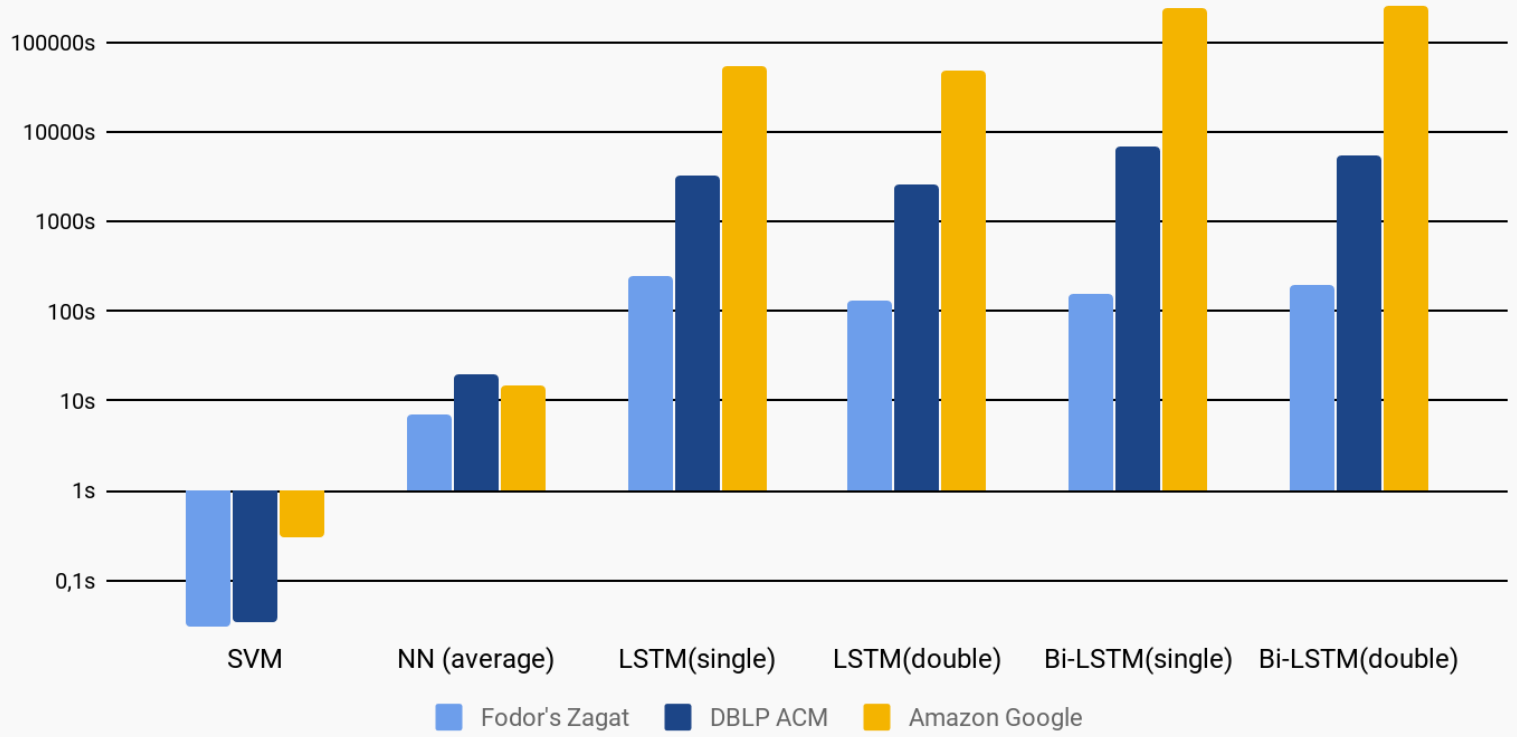
Titolo	acad upgrade dragon naturallyspeaking pro solution 9.0 (a289a-fd7-9.0)	edius pro 4
Descrizione	-marketing information: dragon naturallys … corel wordperfect	whether you are working … broadcast at any time.
Produttore	nuance academic	canopus/grass valley
Prezzo	399.54	585.99

Risultati

F_1 score


100%
99%
98%
97%
96%
95%
SVM
NN (average)
LSTM(single)
LSTM(double)
Bi-LSTM(single)
Bi-LSTM(double)
Fodor's Zagat
DBLP ACM
Amazon Google

Tempo di addestramento

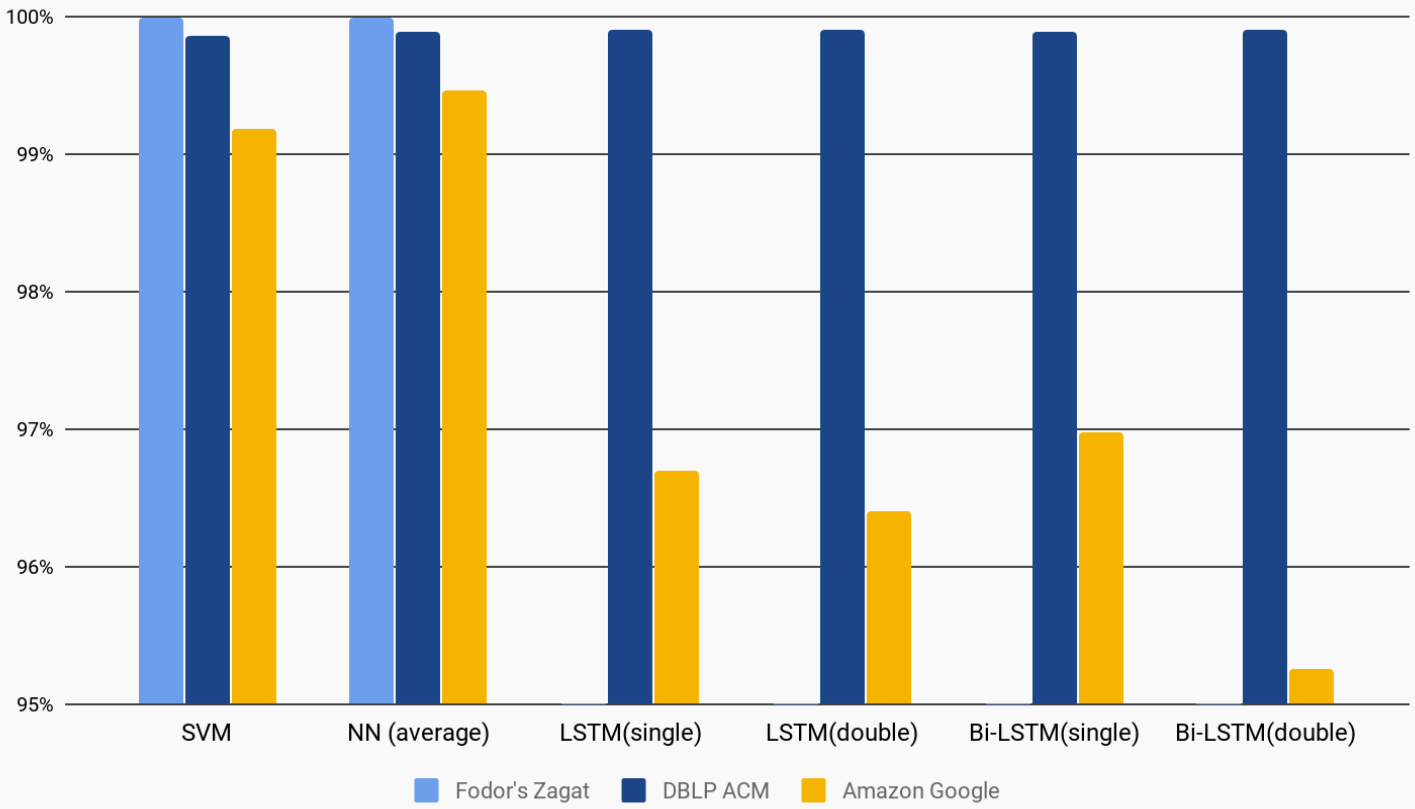


Riduzione dell’insieme di training

Può essere interessante valutare gli algoritmi con meno esempi.

Il training set utilizza il 17% delle tuple che aveva precedentemente disponibili.

F_1score


100%
99%
98%
97%
96%
95%
SVM
NN (average)
LSTM(single)
LSTM(double)
Bi-LSTM(single)
Bi-LSTM(double)
Fodor's Zagat
DBLP ACM
Amazon Google

Conclusioni

Algoritmo ottimo

Algoritmo più veloce, meno efficiente

Composizione: Media

Classificatore: Rete Neurale

→ Miglior F1 score

Composizione: Media

Classificatore: SVM

→ Più veloce,
leggermente meno efficiente

Grazie per l’attenzione