**Tokenization – Tokenizzazione**

Si intende il processo di segmentazione del testo in unità linguistiche minime (scomporre un testo in tokens: parole, caratteri alfanumerici, punteggiatura, etc.).

Questi tokens vengono successivamente analizzati ed elaborati per ottenere tokens linguisticamente significative e utili.

Esempio:

[ricerca, addetti, alle, vendite, di, auto,…]

**Rimozione di stopwords, fonti, area geografica, parole errate.**

Per ogni job description, tokens che fanno riferimento alla fonte (e.g. monster, indeed), ad un’area geografica (comuni, città) o rappresentano una stopwords (*Le stopwords sono parole che, data la loro elevata frequenza in una lingua, sono di solito ritenute poco significative )* sono rimosse.

**Tecnica di Stemming:**

Stemming: è il processo di riduzione delle parole/tokens alla forma base o radice

Passeggiare => passegg

Passeggiata =>passegg

p a r o l a => p a r o l

p a r o l e => p a r o l

Questo serve per aiutare il modello nella compresione di parole che hanno la stessa radice e quindi il modello le considera come identiche.

**reference =** <http://islab.di.unimi.it/ontoweb/materiale/nlp.pdf>

**Bag of Words processing**

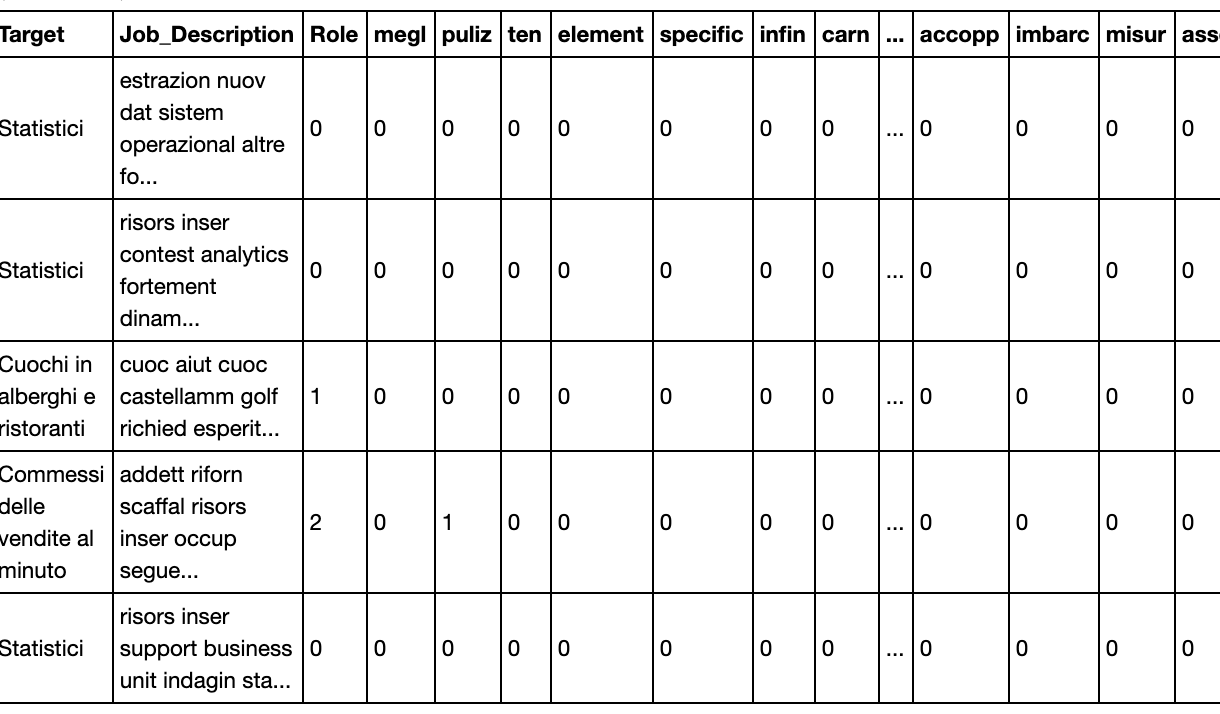
Costruzione di una matrice di dati che ha:

* per colonne l’insieme di parole uniche nel testo
* Per righe le job descriptions
* Come valori le frequenze della parola nella job description

**One Hot Encoding technique:** tecnica non ancora utilizzata ma da utilizzare

Utilizzare come reference una descrizione completa di ogni ruolo lavorativo e rappresentare le parole/tokens significative di ogni descrizione come una variabille one-hot-encoded che viene aggiunta alla matrice formata con l'elaborazione Bag-of-Words.

**Esempio:**

****

**Modello di Machine Learning adottato:**

Naïve Bayes