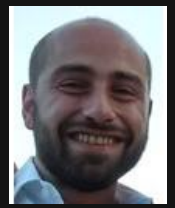
Randstad Artificial Intelligence Challenge



Stefano Fiorucci
Machine learning engineer

randstad

VGEN

Tecnologie impiegate

Baseline: TF-IDF + Naive Bayes

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

```
pipe_0 = Pipeline([
        ('tfidf', TfidfVectorizer()),
        ('clf', MultinomialNB()),
    ])
```

F1 = 0.6638

TF-IDF (con bigrammi) + Naive Bayes

Nella rappresentazione del testo, sono significative le combinazioni di 2 parole consecutive

es.: “object oriented”, “ingegneria informatica”

```
pipe_1 = Pipeline([
      ('tfidf', TfidfVectorizer(ngram_range=(1,2))),
      ('clf', MultinomialNB()),
 ])
```

F1 = 0.7112

+7%

Trattamento parole accorpate

Esaminando il vocabolario di TF-IDF, ci rendiamo conto che esistono alcuni errori nel dataset: parole separate sono state accorpate.

esempi: “programmazioneflessibilità”, “telecomunicazioniconoscenze”, “datawarehouserequisito”, “indeterminatoretribuzione”

Le correggiamo usando un dizionario compilato manualmente.
Per una trattazione più sistematica, algoritmo di Viterbi.

```
pipe_2 = Pipeline([
      ('divide_words', FunctionTransformer(func=divide_words)),
      ('tfidf', TfidfVectorizer(ngram_range=(1,2))),
      ('clf', MultinomialNB()),
 ])
```

F1 = 0.7115

+0.04%

Troncamento Job offer lunghe

Nel dataset, sono presenti annunci
che contengono testo superfluo e non rilevante

i Almeno 3 anni di esperienza in analoga mansione Ottima conoscenza del linguaggio C#, HTML 5, Sql Server Jquery, Json, Ottima conoscenza di .Net framework 3.
5 e successivi Capacità di lavorare in Team Buona capacità di adattamento e flessibilità nel ruolo e nella mansione Capacità di problem solving Orientamento a
ll'obiettivo e al risultato Gradita conoscenza di: Javascript e Bootstrap. Programmazione microcontrollori in C++ Completano il profilo spirito di iniziativa
e autonomia nello svolgimento del progetto assegnato Assunzione a tempo indeterminato direttamente da parte del cliente e RAL al di sopra della media del merc
ato. L'offerta di lavoro si intende rivolta all'uno e all'altro sesso in ottemperanza al D.Lgs. 198/2006.
Contatta l'utente, Grazie per averci aiutato a far rispettare le regole di pubblicazione degli annunci.
x

Segnalazione già inviata!

Grazie per averci aiutato a far rispettare le regole di pubblicazione degli annunci.
x

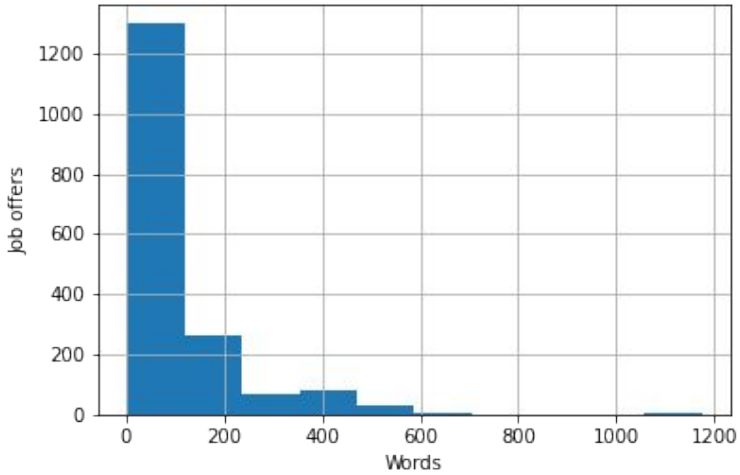
Spiacenti, l'invio della segnalazione non è riuscito

Riprova, oppure contatta l'Assistenza Clienti
x, Ho letto l'informativa sulla Privacy
Do il mio consenso per ricevere email con offerte relative a Kijiji
Do il mio consenso per ricevere email promozionali da parte di terzi, incluse società del gruppo eBay
Conferma di non essere un robot! Cliccando su "Invia" accetti queste condizioni del sito.

irrilevante

Troncamento Job offer lunghe

Analizziamo la distribuzione del numero di parole nel dataset




```
pipe_3 = Pipeline([...,
      ('trunc_input', FunctionTransformer(func=truncate_input,
kw_args={'n_words':600})),
      ... ])
```

F1 = 0.7135

+0.28%

Riduzione numero features

Numero features (dimensione vocabolario TF-IDF): 49623

Possiamo escludere le features meno significative, agendo su due fronti:

- controllando max_df di TF-IDF
- usando un'analisi statistica univariata per selezionare le features più rilevanti

```
pipe_4 = Pipeline([...,
      ('tfidf', TfidfVectorizer(ngram_range=(1,2), max_df=0.7)),
      ('select_features', SelectPercentile(score_func=f_classif,
percentile=98)),
...])
```

F1 = 0.7146

+0.15%

TF-IDF (con bigrammi) + classificatore SGD

Vogliamo provare altri algoritmi di classificazione, adatti al testo.

Scegliamo una SVM lineare, addestrata con discesa stocastica del gradiente (SGD)

```
pipe_5 = Pipeline([
        ...
      ('clf', SGDClassifier()),
 ])
```

F1 = 0.8366

+17% !!!

Tuning degli iperparametri: grid search

```
hyperparameters = {
        'trunc_input__kw_args': [{'n_words':600}, {'n_words':700}],
        'tfidf__max_df':[0.6, 0.7, 0.8, 0.9],
        'select_features__percentile': [x/100 for x in random.sample(range(9000,10000),20)],
        'clf__max_iter':list(range(10,21,1))
        }
```

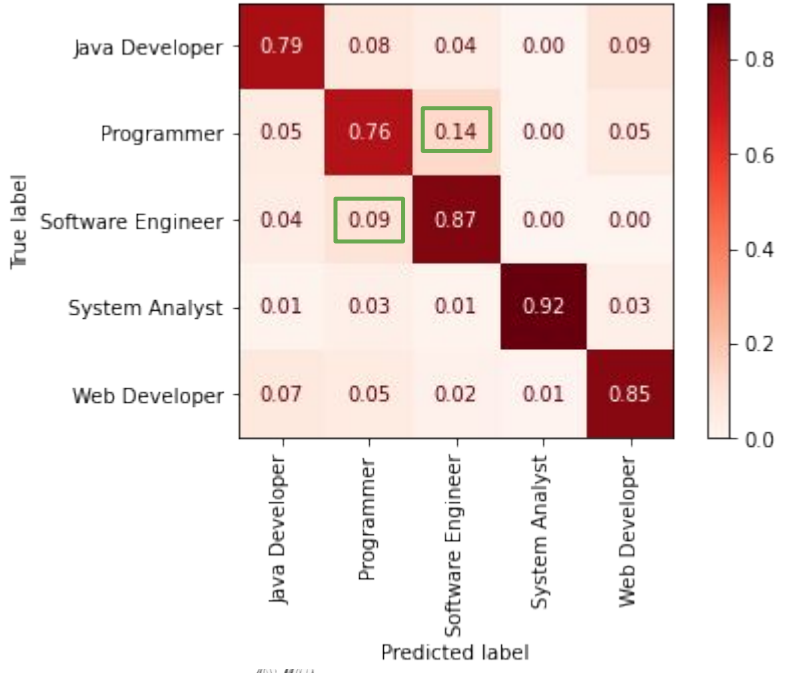
	precision	recall	f1-score	support
Java Developer	0.85882	0.80220	0.82955	91
Programmer	0.78351	0.79167	0.78756	96
Software Engineer	0.77381	0.86667	0.81761	75
System Analyst	0.98592	0.94595	0.96552	74
Web Developer	0.86275	0.85437	0.85854	103
accuracy			0.84738	439
macro avg	0.85296	0.85217	0.85175	439
weighted avg	0.85017	0.84738	0.84805	439

BEST MODEL

F1 = 0.8481

Potenziali miglioramenti

★ Revisione classi



★ Pulizia dataset

I'm looking for a young profile, even a recent graduate, [...]

Was Sie erwartet: Sie treiben neue Technologien voran und sind [...]

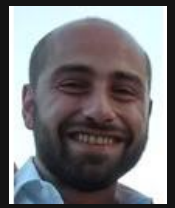
★ Incremento dataset → modelli più sofisticati

fastText

Transformers

Google BERT

Randstad Artificial Intelligence Challenge



Stefano Fiorucci
Machine learning engineer

randstad

VGEN