Contexte du projet

Preprocessing

- simplification

- tokenisation

- lemmatisation

Régression logistique

Bag of words

```
(1600000, 500)


[[0 0 0 ... 0 1 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 0 1 1]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
```

Régression logistique

- entraînement rapide

- score F-2 = 0.769

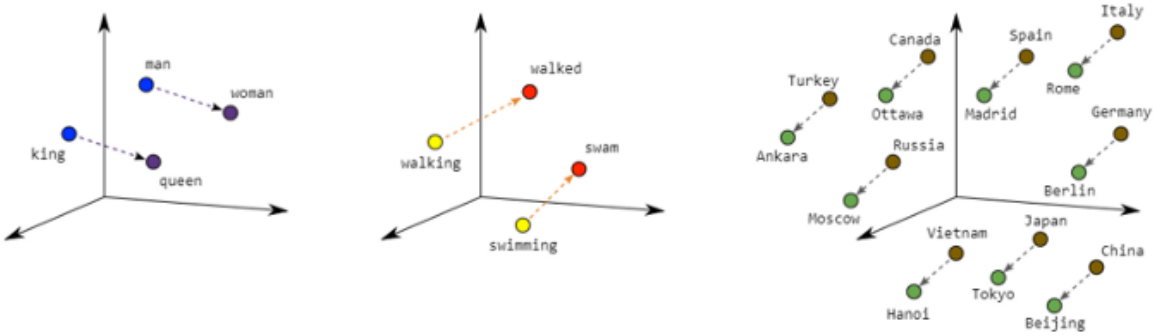
Réseau de neurones

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 67, 300)           299998800

 lstm (LSTM)                 (None, 67, 50)            70200

 lstm_1 (LSTM)               (None, 67, 20)            5680

 lstm_2 (LSTM)               (None, 10)                1240

 dense (Dense)               (None, 2)                 22

=================================================================
Total params: 300,075,942
Trainable params: 77,142
Non-trainable params: 299,998,800
```

Réseau de neurones

Plongement de mots

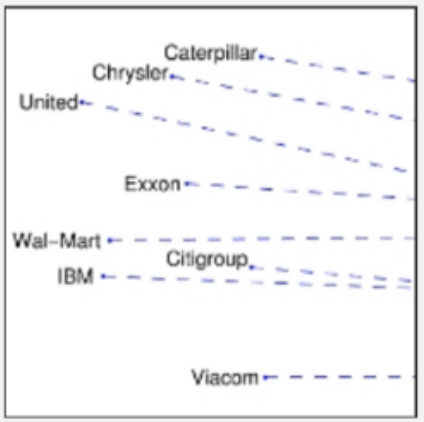
Word2Vec



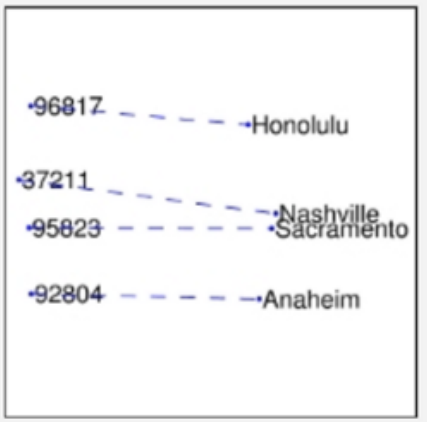
Male-Female Verb Tense Country-Capital

GloVe

man - woman



company - ceo

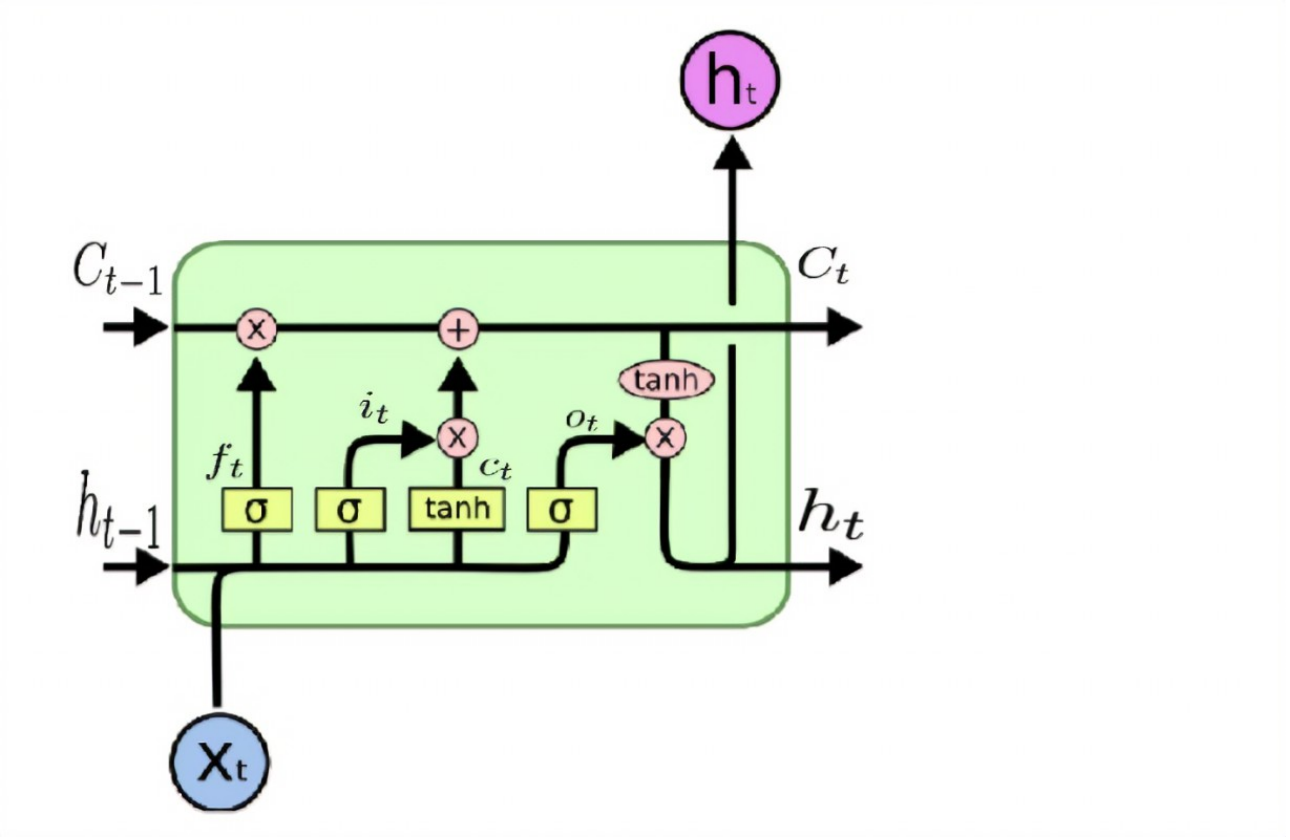


city - zip code

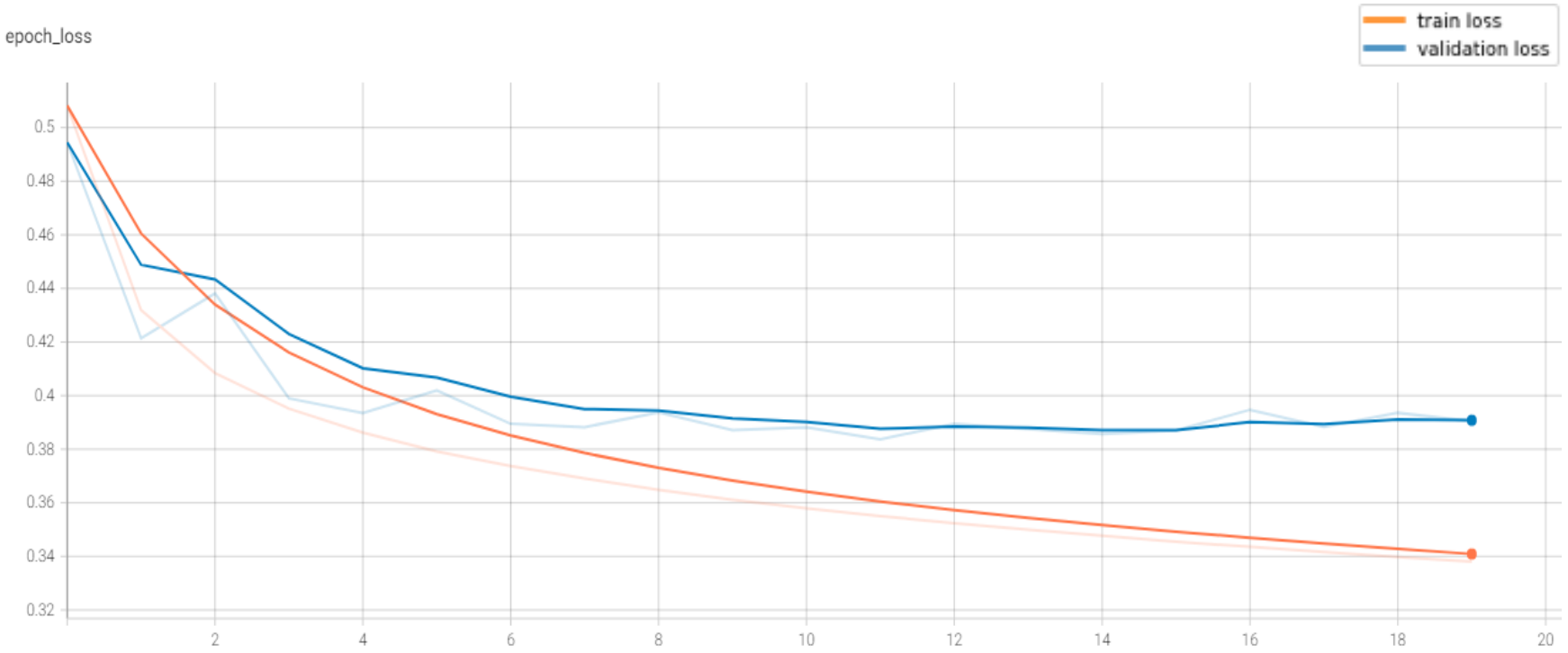
comparative - superlative

Réseau de neurones

LSTM



Réseau de neurones

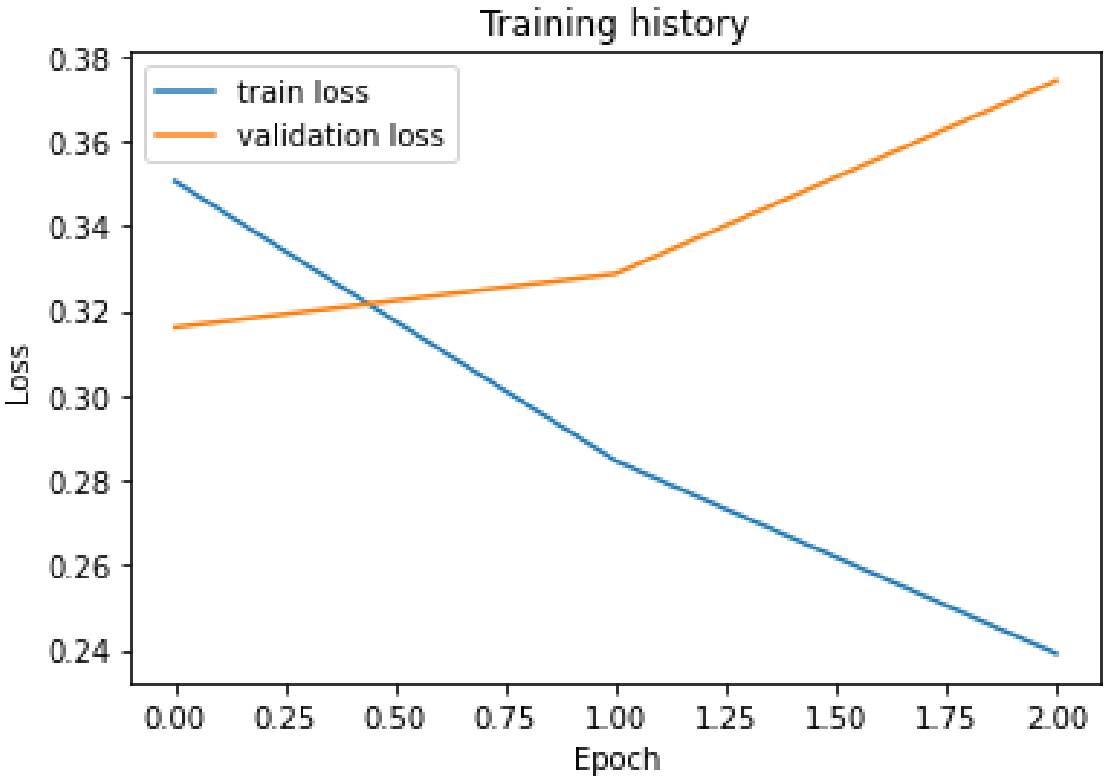


score F-2 = 0.829

DistilBERT

```
(transformer): Transformer(
  (layer): ModuleList(
    (0): TransformerBlock(
      (attention): MultiHeadSelfAttention(
        (dropout): Dropout(p=0.1, inplace=False)
        (q_lin): Linear(in_features=768, out_features=768, bias=True)
        (k_lin): Linear(in_features=768, out_features=768, bias=True)
        (v_lin): Linear(in_features=768, out_features=768, bias=True)
        (out_lin): Linear(in_features=768, out_features=768, bias=True)
      )
      (sa_layer_norm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
      (ffn): FFN(
        (dropout): Dropout(p=0.1, inplace=False)
        (lin1): Linear(in_features=768, out_features=3072, bias=True)
        (lin2): Linear(in_features=3072, out_features=768, bias=True)
        (activation): GELUActivation()
      )
      (output_layer_norm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
```

DistilBERT



score F-2 = 0.865

Comparaison

	F-2 score	Total train time (min)	1 epoch train time (min)	On GPU
Linear regression	0,770	4	4	NO
Neural network	0,829	40	2	YES
DistilBERT	0,865	825	275	YES

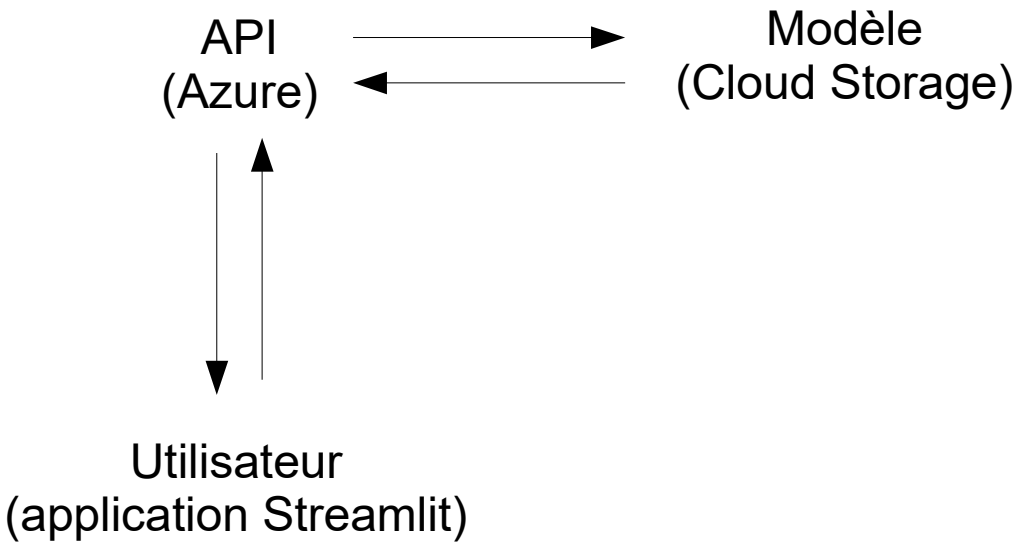
Déploiement cloud

- Azure App Service

- Google Cloud Storage

- Github

Déploiement cloud



Déploiement cloud