Quora Question Pairs

Can you identify question pairs that have the same intent?

Naïve Bayes Classifier k 0.64885

Snowball stemmer 'having' → 'have'

Feature engineering [question1, question2] ↦ [similar_words, difference]

Similar words

Absolute difference of length

Train a Naïve Bayes Classifier

Validation accuracy: 0.65

Most Informative Feature

```
similar_words = 1   ;   0 : 1 = 1805.6 : 1.0
```

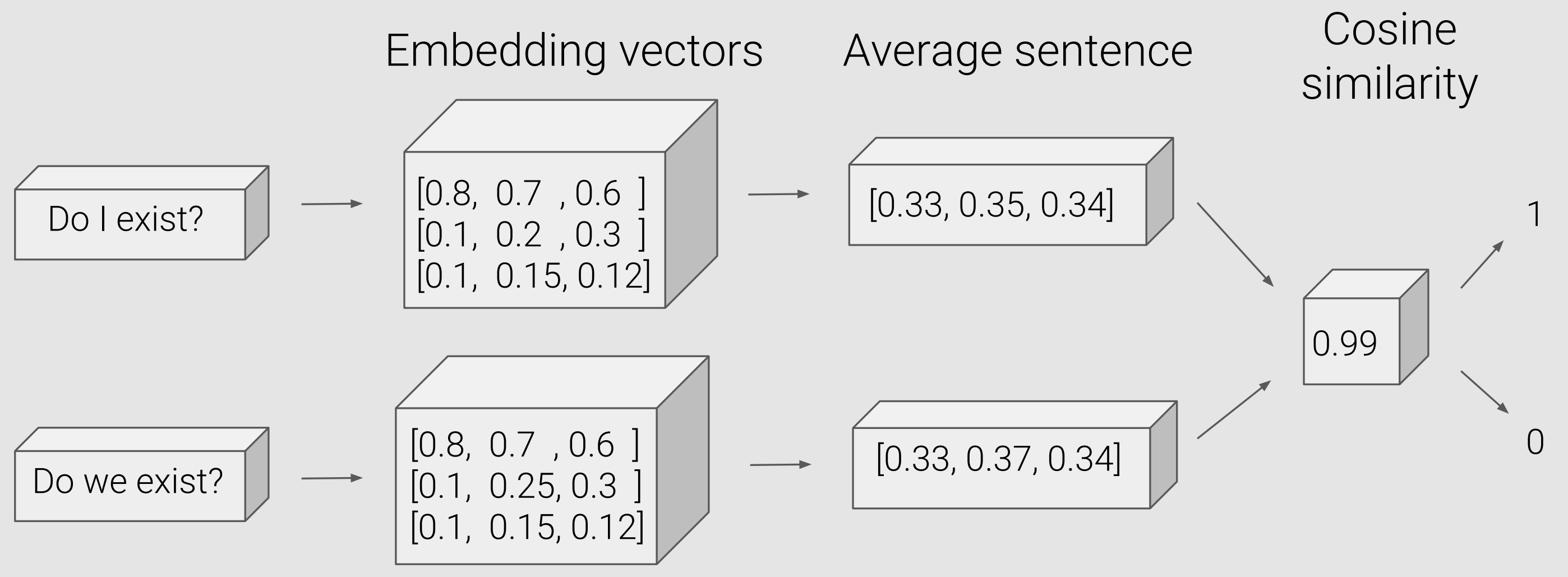
Word2Vec k 0.69281

[king] - [man] + [woman] ≈ [queen]

Hyperparameters

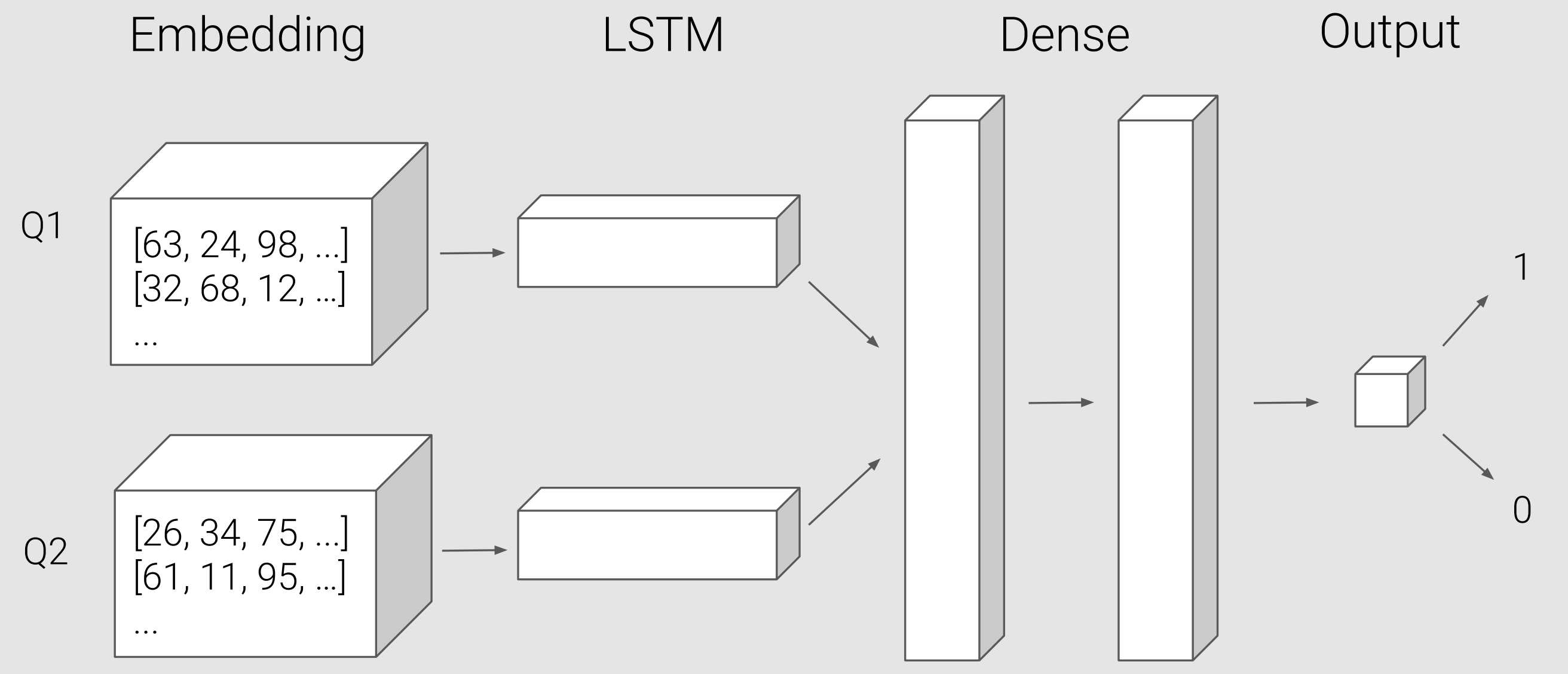
Context - window size = 5

Start from empty model No transfer learning



Deep Learning Model v1.0 k 0.78967

Raw	[What, is, a, Horcrux, ?]	+	**Global Vectors (GloVe)**
Lemma	[what, be, a, horcrux, ?]		Source: Common Crawl 840B tokens 2.2M vocab 300d vectors
Tags	[WP, VBZ, DT, NNP, .]		



Cleaning and Feature Engineering

Removing the shortest and the longest questions

Adding new features

Basic Features | Fuzzy Features

Pre-trained Word2Vec | Spacy | Mistakes (TextBlob)

dmlc XGBoost k 0.81027

Input

Extracted features

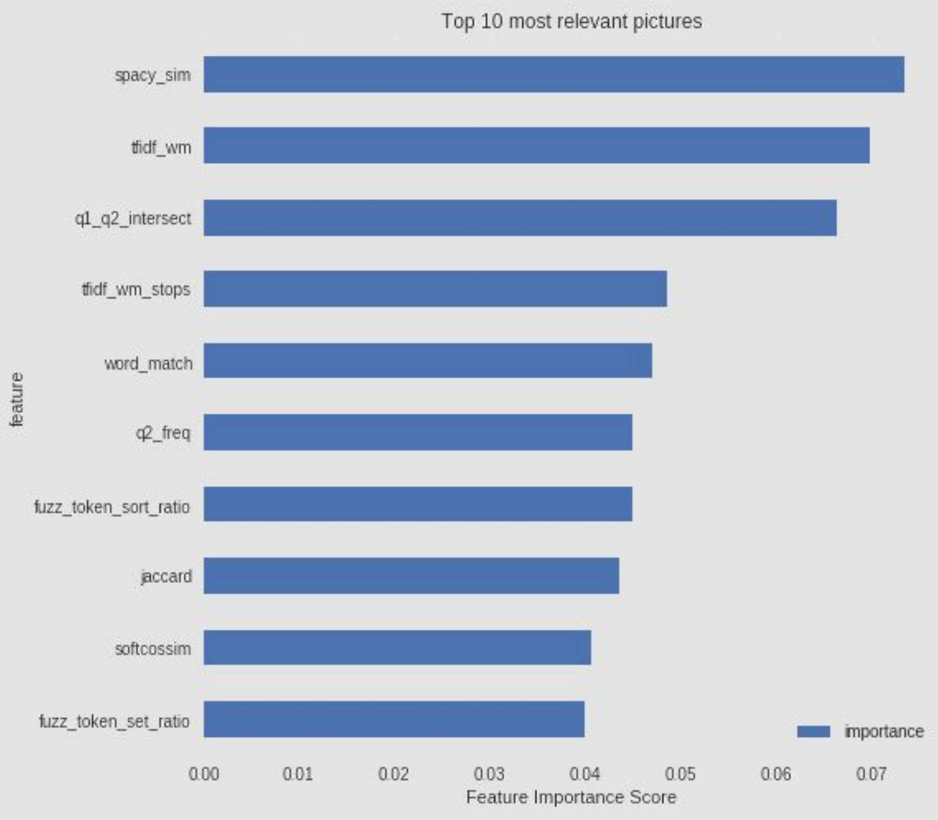
is_duplicate label

Hyperparameters tuning

Cross-validation: 10 folds

Grid search

Randomized search



Logistic Regression k 0.81049

L2 penalty

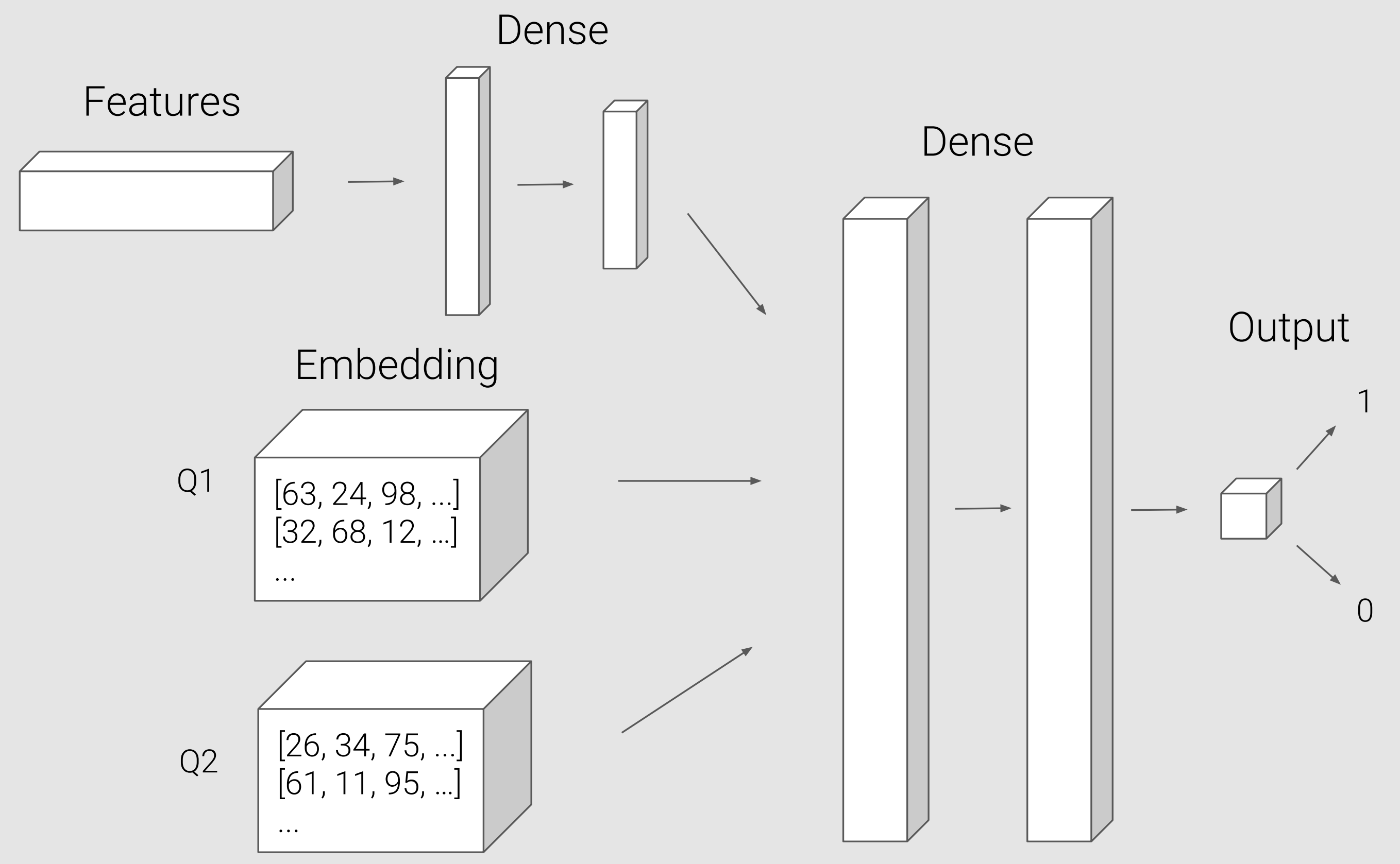
Hyperparameters tuning

C ∈ [0.001,1000] in a logarithmic step

Grid search

Choose C with smallest diff. training-validation accuracy

Deep Learning Model v2.0 k 0.81929



Ensemble Learning k 0.85912

Bagging

Less variance

Fights overfitting

7 aggregated models (k > 0.79)

4x Deep Learning models

1x Logistic Regression model

2x XGBoost models

"The majority cannot be wrong"

"Two heads are better than one"

Each model will have the same voting power:

[0, 1, 0, 0, 0, 0, 0] → 0

[1, 1, 1, 1, 1, 0, 0] → 1

kaggle

Albert **Folch** | Xavier **Moreno**