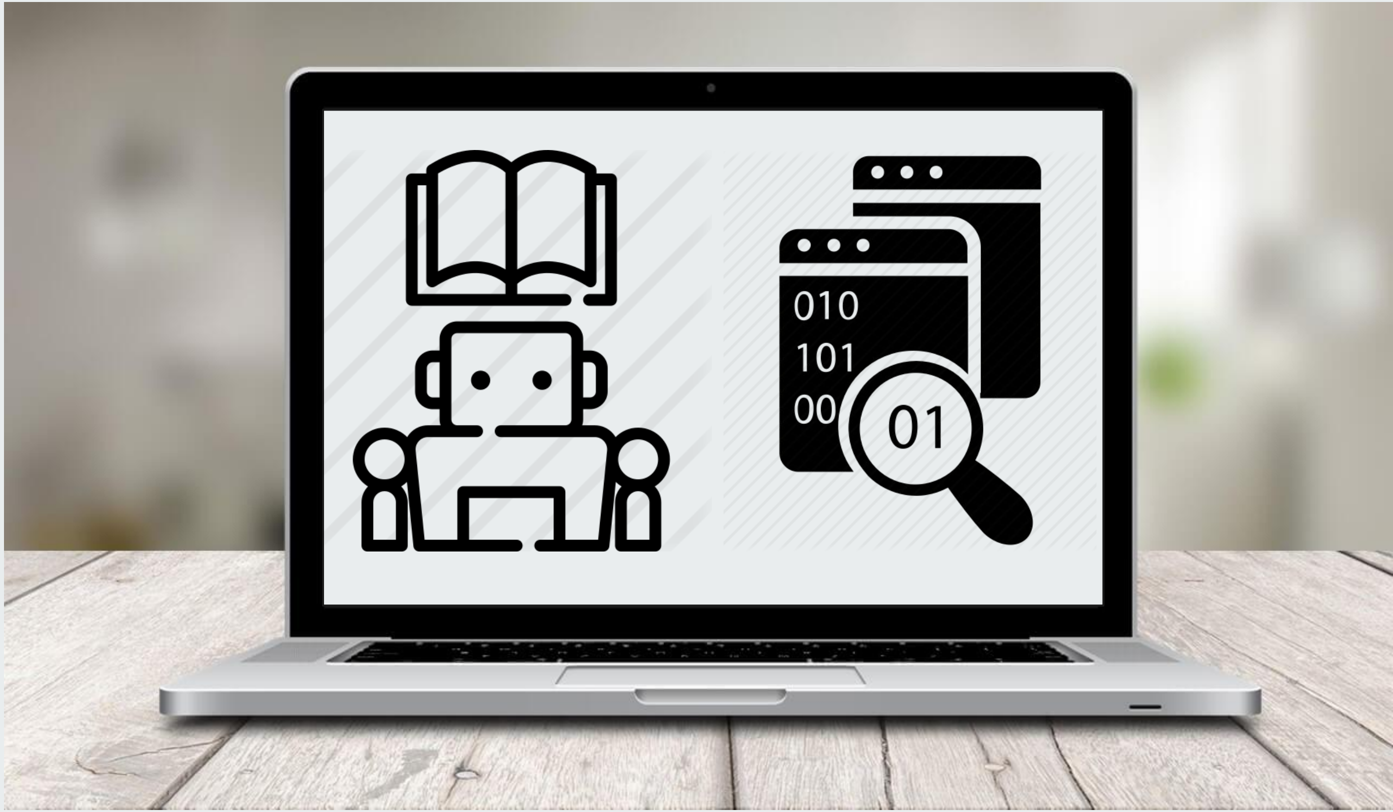
Alberto Marengo

Text to SQL



Using ML to generate queries from Natural Language

Outline

The Question

Data Collection

Data Description

Completed Clean-Up and Modeling

Plan of Action

The Question

Can AI read a Natural Language Query and translate it into a Database Query?

Data Collection

- ***wikiSQL***

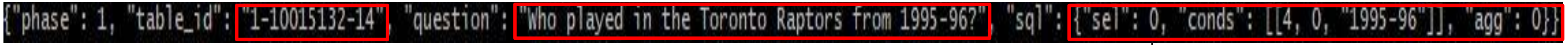
A crowd-sourced dataset of ~56,000 hand-annotated examples of questions and SQL queries distributed across ~24,000 tables from Wikipedia

- ***Spider***

A dataset annotated by 11 Yale students consisting s of ~10,000 questions and ~5,700 unique complex SQL queries on 200 databases. Queries are more complex than wikiSQL

Data Description

wikiSQL Dataset - row example

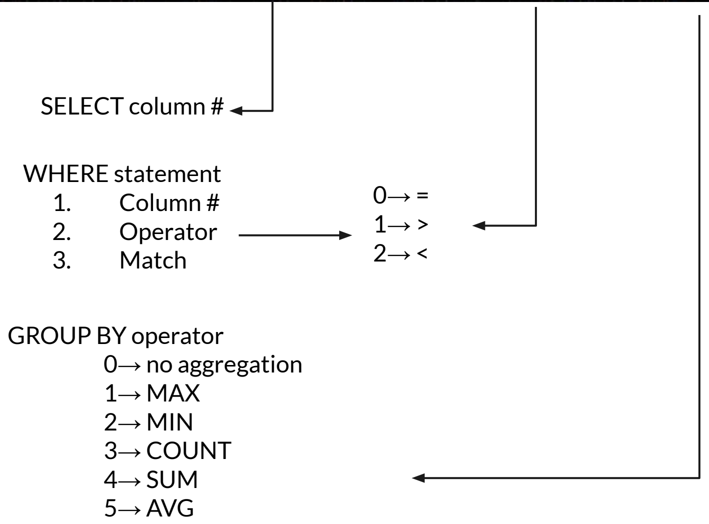


{"phase": 1, "table_id": "1-10015132-14", "question": "Who played in the Toronto Raptors from 1995-96?", "sql": {"sel": 0, "conds": [[4, 0, "1995-96"]], "agg": 0}}

Question:
Who played in the Toronto Raptors from 1995-1996?

SQL query:

SELECT column_0
FROM 1-10015132-14
WHERE column_4 = “1995-96”;



SELECT column #

WHERE statement

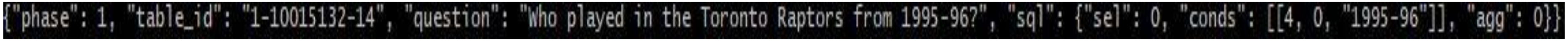
1. Column #
2. Operator
3. Match

0→ =
1→ >
2→ <

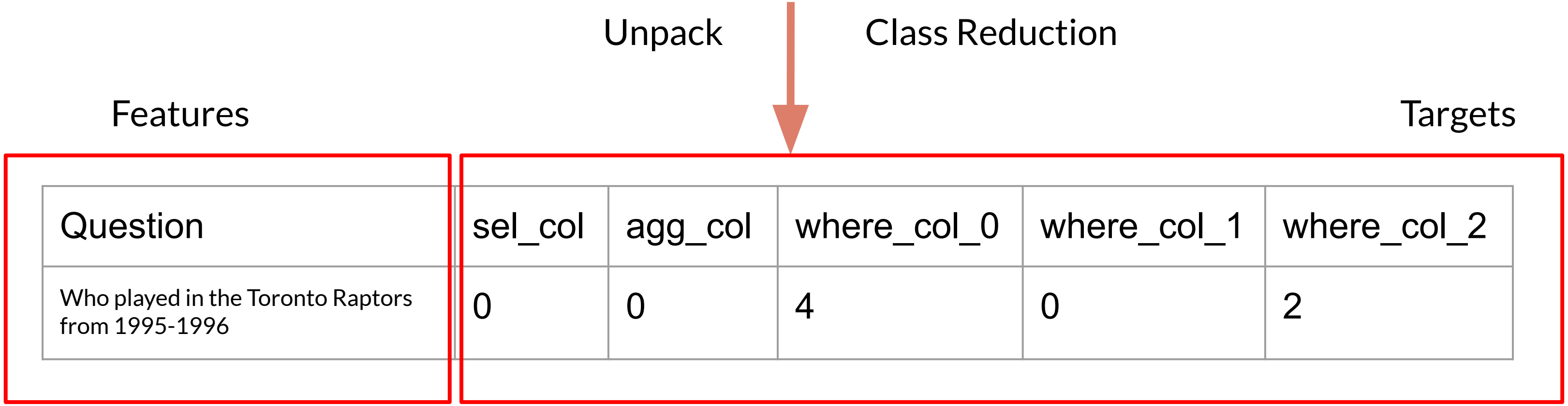
GROUP BY operator

0→ no aggregation
1→ MAX
2→ MIN
3→ COUNT
4→ SUM
5→ AVG

Data Clean-Up



```
{"phase": 1, "table_id": "1-10015132-14", "question": "Who played in the Toronto Raptors from 1995-96?", "sql": {"sel": 0, "conds": [[4, 0, "1995-96"]], "agg": 0}}
```



Unpack

Class Reduction

Features

Targets

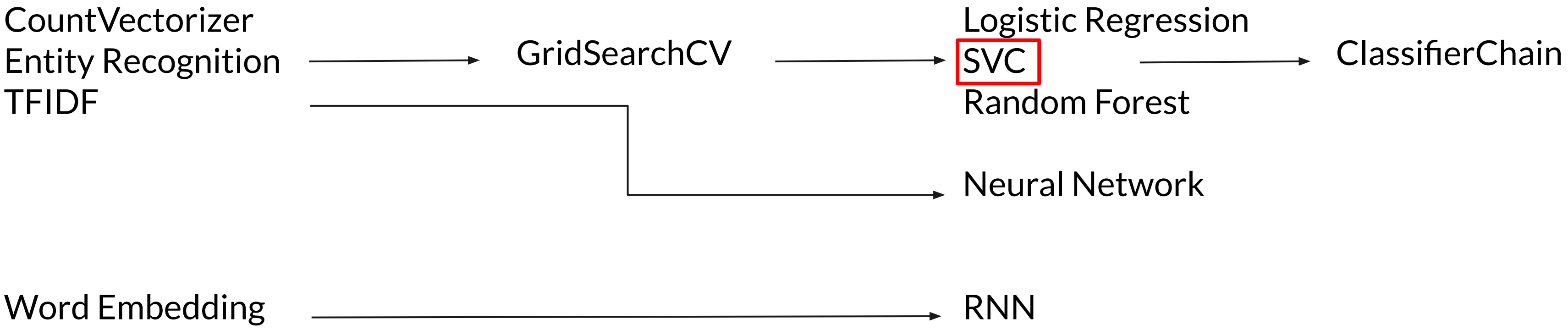
Question	sel_col	agg_col	where_col_0	where_col_1	where_col_2
Who played in the Toronto Raptors from 1995-1996	0	0	4	0	2

Multi-class multi-output classification problem

Data Modeling

Text Vectorization

Machine Learning



CountVectorizer
Entity Recognition
TFIDF

GridSearchCV

Logistic Regression
SVC
Random Forest

ClassifierChain

Neural Network

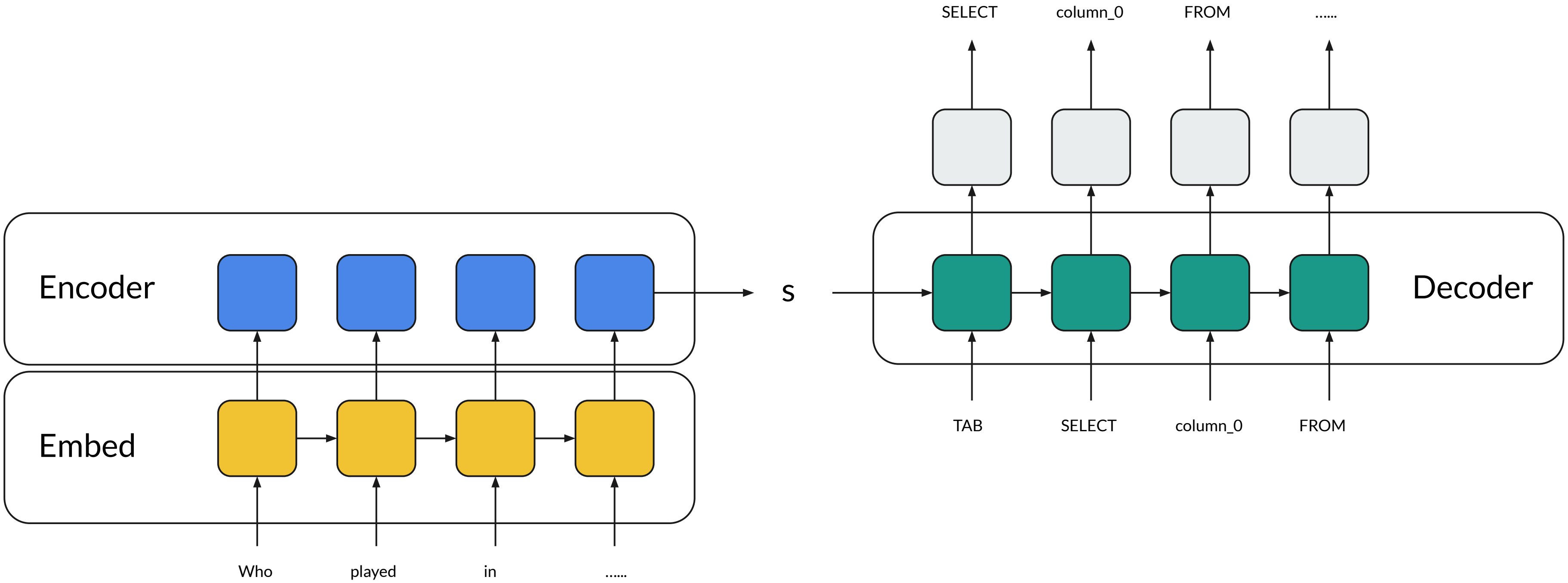
Word Embedding

RNN

Action Plan

- ➔ Understand ClassifierChain accuracies
- ➔ Optimize RNN hyperparameters
- ➔ Build a SeqToSeq model
- ➔ Train the model on the Spider dataset

SeqToSeq Model



Questions?