TAPAS – TablePArSing – Technique to Retrieve Information from tabular data using NLP

* TAPAS is essentially a [BERT](https://en.wikipedia.org/wiki/BERT_(language_model)) model-based approach to question answering over tables
* Google has recently open-sourced one of their models called ‘TAPAS’ (for **TA**ble **PA**r**S**ing) wherein you can ask questions about your data in natural language.
* **BERT Stands for Bidirectional Encoder Representations from Transformers**
* SQuAD (Stanford Question Answering Dataset)
* TAPAS is essentially a [BERT](https://en.wikipedia.org/wiki/BERT_(language_model)) model-based approach to question answering over tables.
* is a weak-supervision technique relying on denotations (i.e. literal or primary meaning of the words and not the underlying idea or emotions).
* predicts the denotation by selecting table cells, optionally applies a corresponding aggregation operator to such selection and makes end-to-end predictions.

**BERT**

BERT is a deeply bidirectional, [unsupervised](https://en.wikipedia.org/wiki/Unsupervised_learning) language representation, pre-trained using only a plain [text corpus](https://en.wikipedia.org/wiki/Text_corpus). Context-free models such as [word2vec](https://en.wikipedia.org/wiki/Word2vec) or [GloVe](https://en.wikipedia.org/wiki/GloVe_(machine_learning)) generate a single word embedding representation for each word in the vocabulary, whereas BERT takes into account the context for each occurrence of a given word. For instance, whereas the vector for "running" will have the same word2vec vector representation for both of its occurrences in the sentences "He is running a company" and "He is running a marathon", BERT will provide a contextualized embedding that will be different according to the sentence

**What is weak-supervision?**

* Weak supervision is an area of machine learning where noisy, limited, or imprecise sources are used to provide supervision signals for labelling large amounts of training data in a supervised learning task. It is a way of using lower-quality labels more efficiently and/or at a higher abstraction level than It eliminates the need for obtaining hand-labelled data sets, which can be expensive or impractical. The cheaper weak labels, though imperfect, can be used to create a strong predictive model.

## ****What is the BERT model?****

BERT (Bidirectional Encoder Representations from Transformers) is an [NLP](https://en.wikipedia.org/wiki/Natural_language_processing) model introduced by [Google Research](https://research.google/) in 2018. It pretrains deep bidirectional representations from the unlabeled text by jointly conditioning on both left and right context in all the network layers. So, the pre-trained BERT model can be fine-tuned by adding only one additional output layer. It is found to be useful for several Natural Language Processing (NLP) and Natural Language Understanding ([NLU](https://en.wikipedia.org/wiki/Natural-language_understanding)) tasks, such as question answering and language inference, without requiring significant task-specific modifications to be done to the model architecture.

TAPAS is a weakly supervised question answering model. It reasons over tables without generating logical forms. It predicts a minimal program by selecting a subset of the table cells and a possible aggregation operation to be executed on top of them. Consequently, it can learn operations from natural language, without specifying them in some formalism.

TAPAS is an extension of BERT’s architecture. It has some additional embeddings which can capture tabular structure. Besides, it has two classification layers – one for cells’ selection and the other for predicting a corresponding aggregation operator. It flattens the tabular data into a sequence of words. It then splits those words into word pieces called ‘tokens’ and concatenates the question tokens before the table tokens.

It adds a separator token between the question and the table, but not between cells or rows. The token embeddings are combined with table-aware positional embeddings before feeding them to the model

**TAPAS MODEL DESCRIPTION:**

TAPAS is a BERT-like transformers model pretrained on a large corpus of English data from Wikipedia in a self-supervised fashion. This means it was pretrained on the raw tables and associated texts only, with no humans labelling them in any way (which is why it can use lots of publicly available data) with an automatic process to generate inputs and labels from those texts. More precisely, it was pretrained with two objectives:

* Masked language modeling (MLM): taking a (flattened) table and associated context, the model randomly masks 15% of the words in the input, then runs the entire (partially masked) sequence through the model. The model then has to predict the masked words. This is different from traditional recurrent neural networks (RNNs) that usually see the words one after the other, or from autoregressive models like GPT which internally mask the future tokens. It allows the model to learn a bidirectional representation of a table and associated text.