Развитие метода ранжирования вопросно-ответных пар по релевантности

Выполнил: Файтельсон А.А.

Предметная область

Основные объекты предметной области:

- **База знаний**
- **Запросы от пользователей**
- **Результат работы модели**

Основные процессы предметной области:

- **Получение релевантного ответа от модели**

Объективное противоречие

“Сравнение” текстов на основе TF-IDF выдает недостаточно точные ответы.
Пример:
Вопрос: Как получить УВБД участникам СВО?

Результат от модели:
Как военнослужащему ЧВК получить удостоверение ветерана боевых действий

Объективное противоречие

- **Желаемый ответ:**
  Выдача УВБД участникам СВО, получившим ранение

Цель и задачи

Цель:

- Сравнить разные подходы к решению и посмотреть какая из них будет выдавать более точный результат.

Задачи:

- Выделить конкретные и точные метрики для определения подходящего результата.
- Создание бенчмарков на основе метрик.

Подходы к решению задач

- Использование более продвинутой функции поиска BM-25
- Использование Sentence Transformers для семантического поиска

Метрика TF-IDF

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

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

Функция BM25

$$w_j(\overline{d},C) = \frac{(k_1+1)tf_j}{k_1((1-b)+b\frac{dl}{avdl})+tf_j}\log\frac{N-df_j+0.5}{df_j+0.5}$$

d - документ

C – коллекция документов

w_j(d,C) – вес j-го терма в документе d коллекции C

tf_j – частота j-го терма в документе d коллекции C (TF)

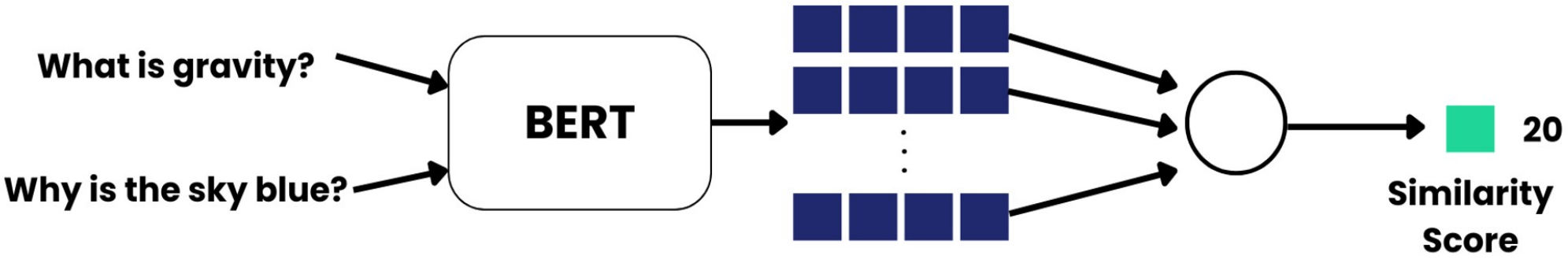
df_j – количество документов коллекции, содержащих j-й терм

dl – длина документа

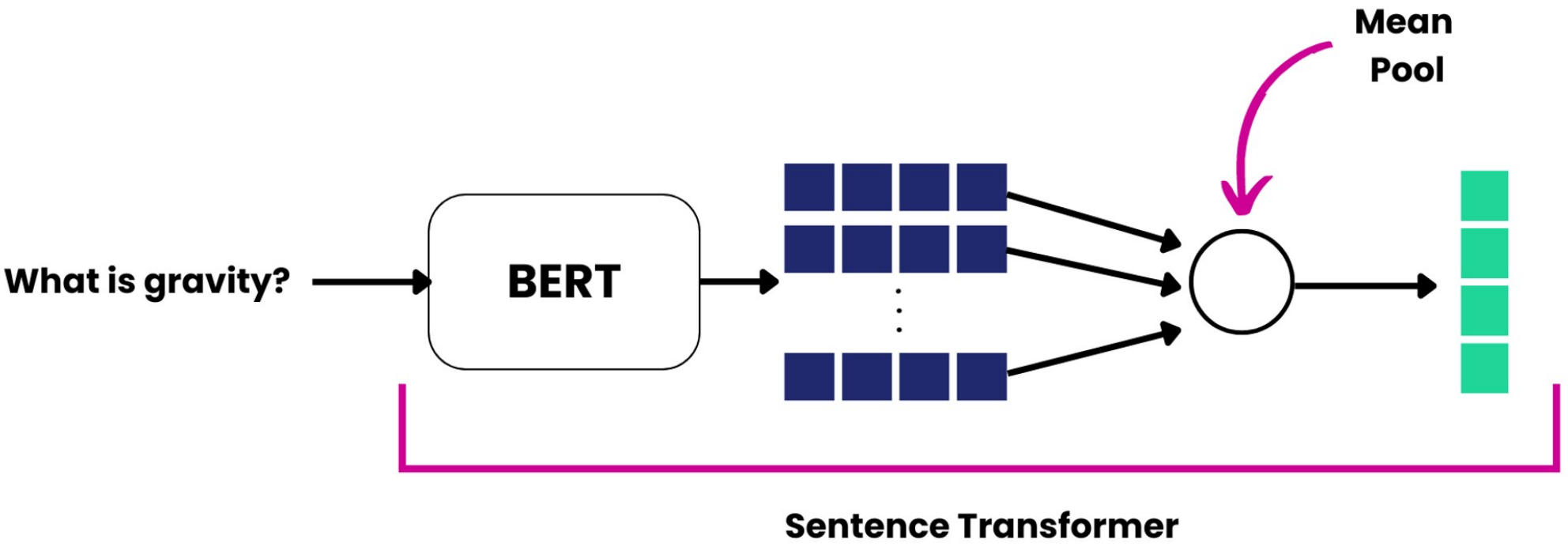
avdl – средняя длина документов в коллекции

k_1, b – коэффициенты (обычно k_1 = 2, b = 0.75)

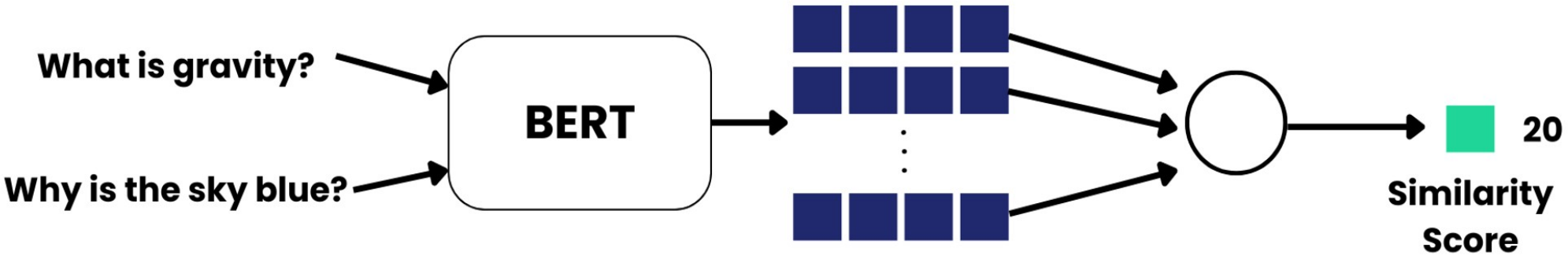
Архитектура Sentence Transformers



Архитектура Sentence Transformers



Архитектура Sentence Transformers



Стек технологий

- SentenceTransformer: sentence_transformers, torch
- TF-IDF: pymorphy2, nltk, sklearn
- BM25: nltk, rank_bm25

Датасет

- Была предоставлена база знаний.
- Она содержит большое кол-во txt файлов с ответами.

Модель должна выдавать релевантный ответ из базы знаний.

Надеюсь на ваши вопросы