Busca Semântica com Transformers

COVID-19 Bert Literature Search Engine

Isabela, João Gabriel, Juliana, Pedro, Vinícius

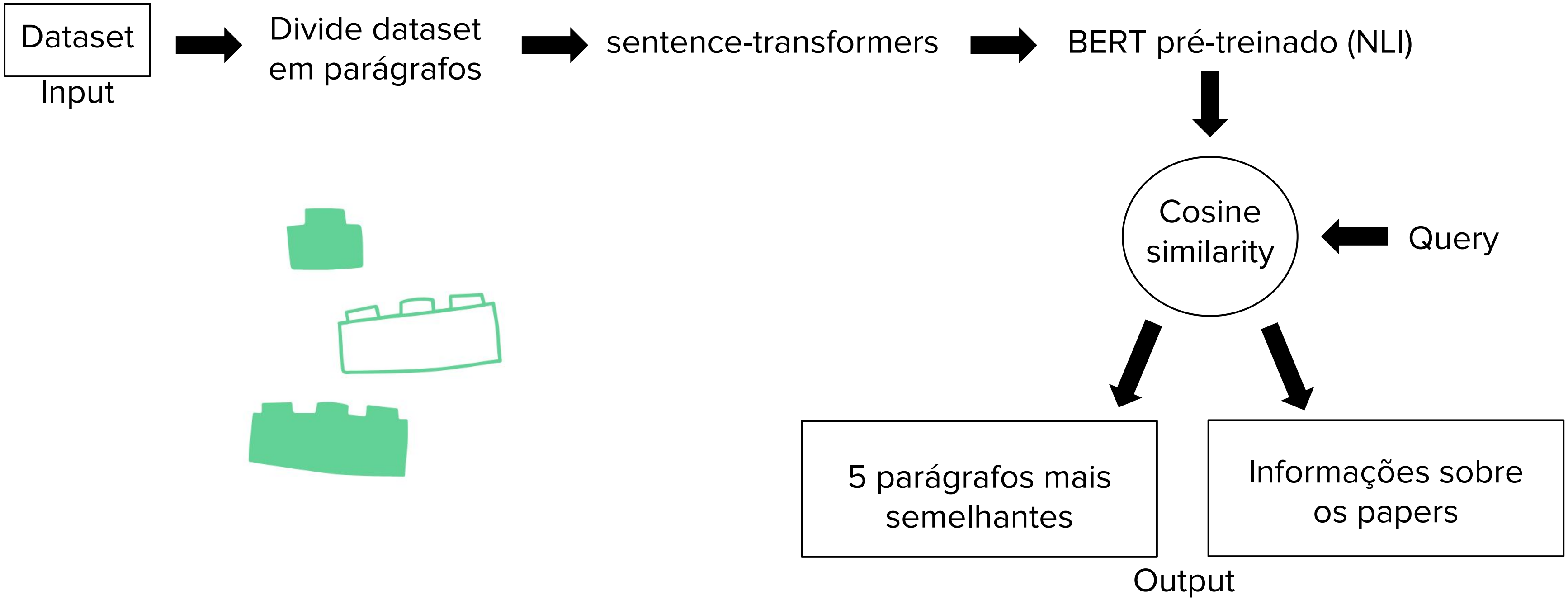
Agenda

1. Abordagem
2. Sentence Transformers
3. Aplicação em Bases de Conhecimento

Abordagem

1. Extrair embeddings de parágrafos
2. Extrair embeddings do termo de busca
3. Distância de cosseno para obtenção de dados mais semelhantes

Execução



Dataset

Input

Divide dataset em parágrafos

sentence-transformers

BERT pré-treinado (NLI)

Cosine similarity

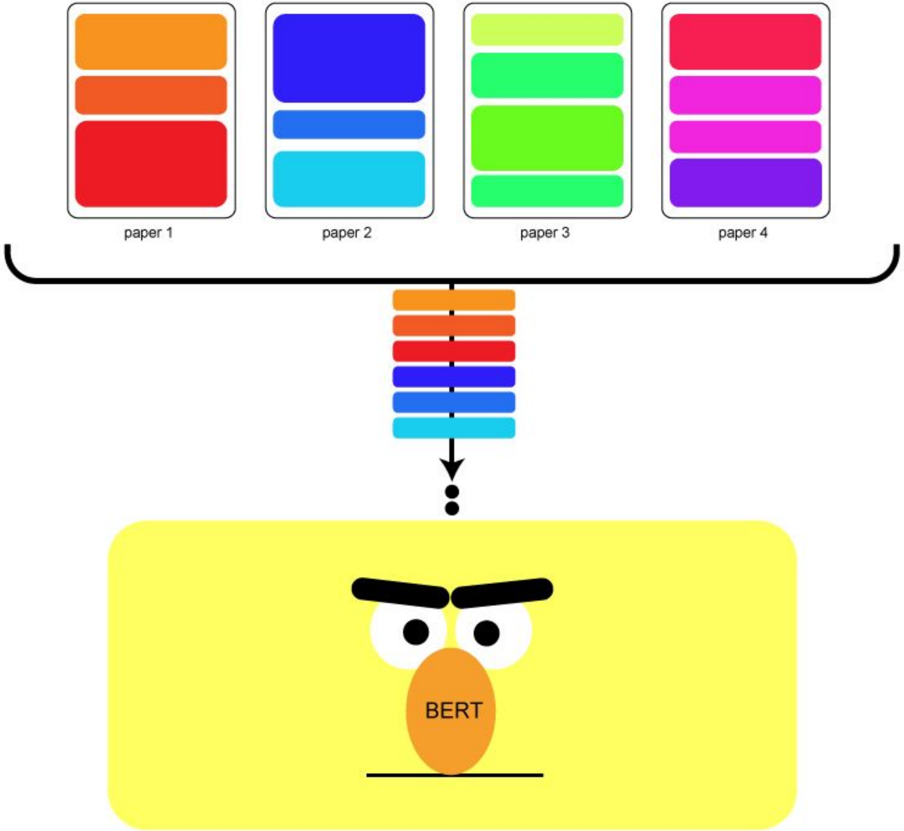
Query

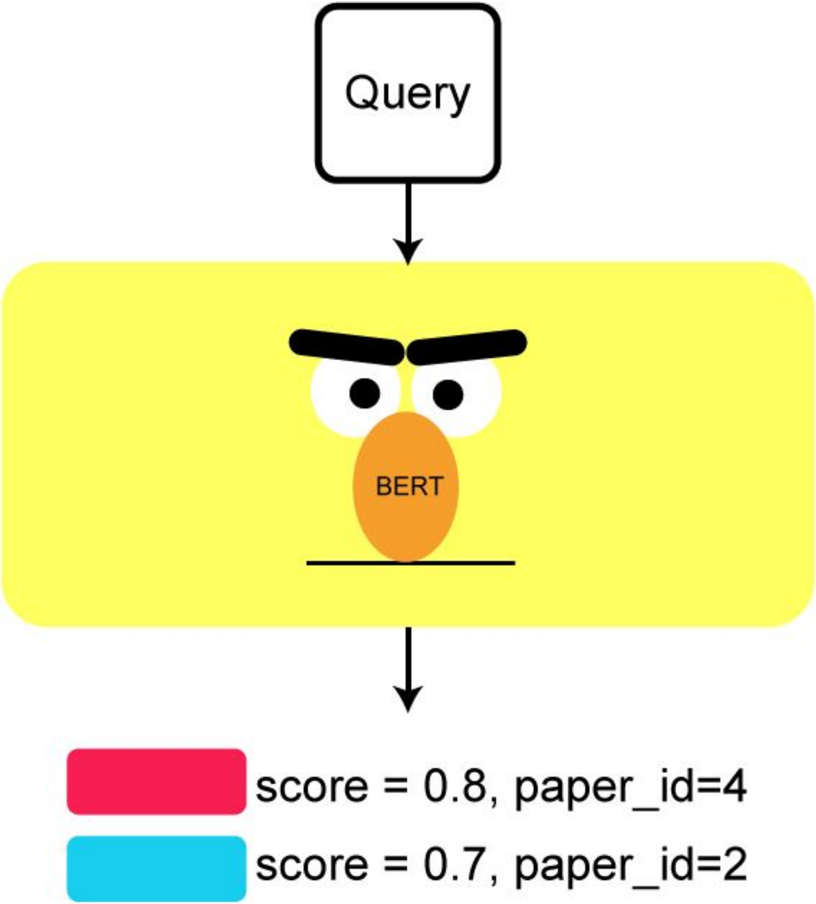
5 parágrafos mais semelhantes

Informações sobre os papers

Output

BERT





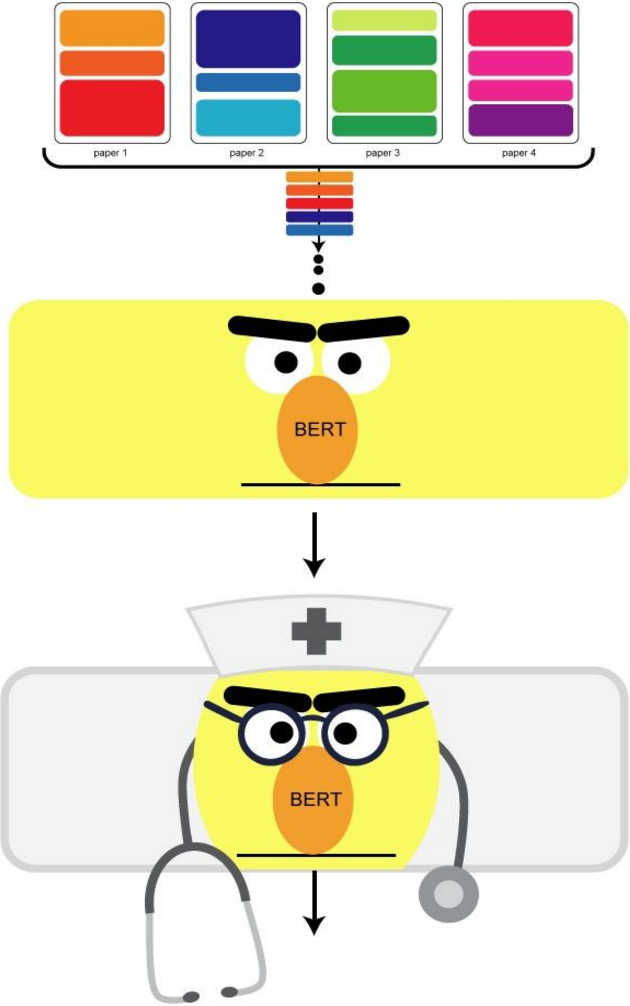
Query

BERT

🟥 score = 0.8, paper_id=4

🟦 score = 0.7, paper_id=2

Fine-tuning



Sentence Transformers

Busca semântica com BERT

Cross Encoder: **concatenação de duas sentenças** são alimentadas na DNN

1. **Não** calcula Sentence Embeddings **independentes**
2. **Complexa função** de regressão **depende do modelo**

Encontrar o par de sentenças mais similar em **10.000 x 10.000** sentenças:

65 horas com V100 GPU

Busca semântica com Sentence-BERT

Utiliza **siamese and triplet network**: cada sentença **paralelamente** alimentada

1. **Gera** Sentence Embeddings **independentes**
2. Pode ser comparado usando **cosine similarity e outras funções**

Encontrar o par de sentenças mais similar em **10.000** sentenças:

5 segundos com V100 GPU

Natural Language Inference

- Determinar se uma **hipótese** é **verdadeira** (entailment), **falsa** (contradiction) ou **indeterminada** (neutral) **dado uma premissa**
- Também denominado de Recognizing Textual Entailment (RTE)
- Treino em **NLI** gera **sentence embeddings universais**.
- SBERT foi **treinado na combinação** de Stanford Natural Language Inference **(SNLI)** e Multi-Genre Natural Language Inference **(MNLI).**

A soccer game with multiple males playing.

Some men are playing a sport.

entailment

SNLI

The Old One always comforted Ca'daan,

except today.

Ca'daan knew the Old One very well.

neutral

MNLI

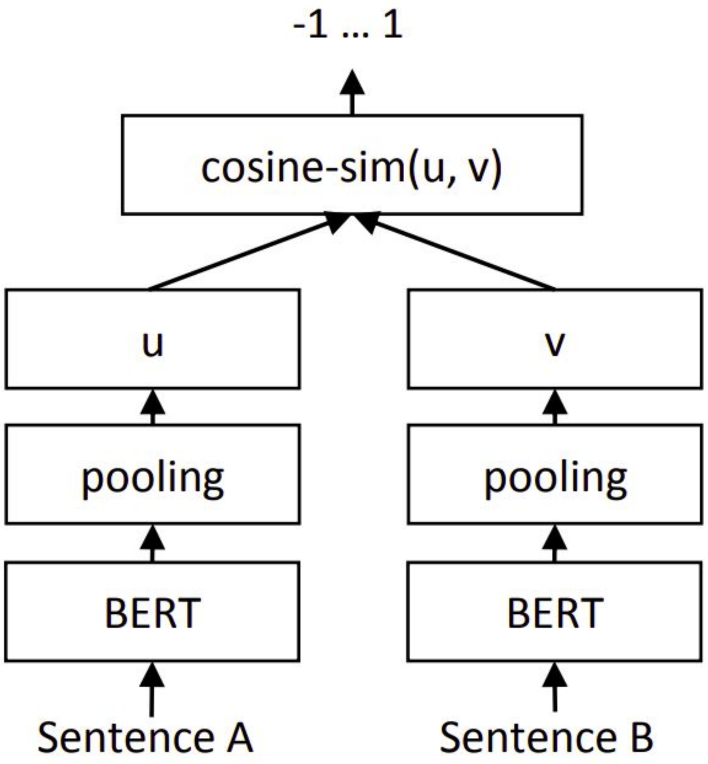
Sentence embeddings

Práticas comuns:

- Embeddings **individuais** usando BERT
- **Inicialização aleatória**
- Tamanho fixo por meio de média ou CLS

Sentence BERT

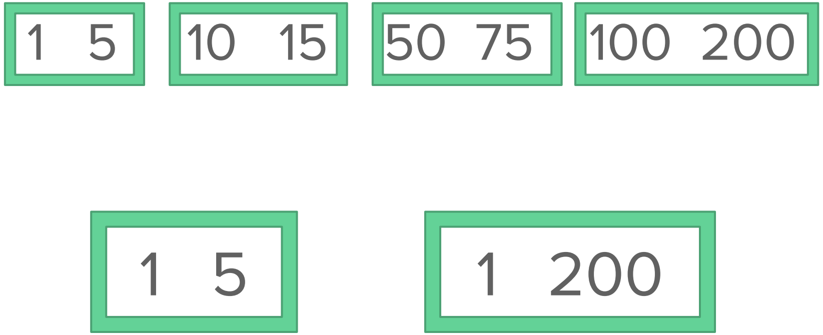
- BERT **"siamês" para a tarefa**
- Carrega o **BERT pré-treinado**
- Tamanho fixo por meio de **média**, máximo, CLS



Smart Batching

Sentenças de tamanho: Batch:

1 5 10 15 50 75 100 200 2



1 5

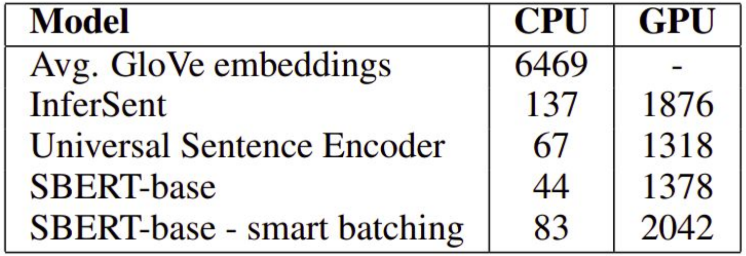
10 15

50 75

100 200

1 5

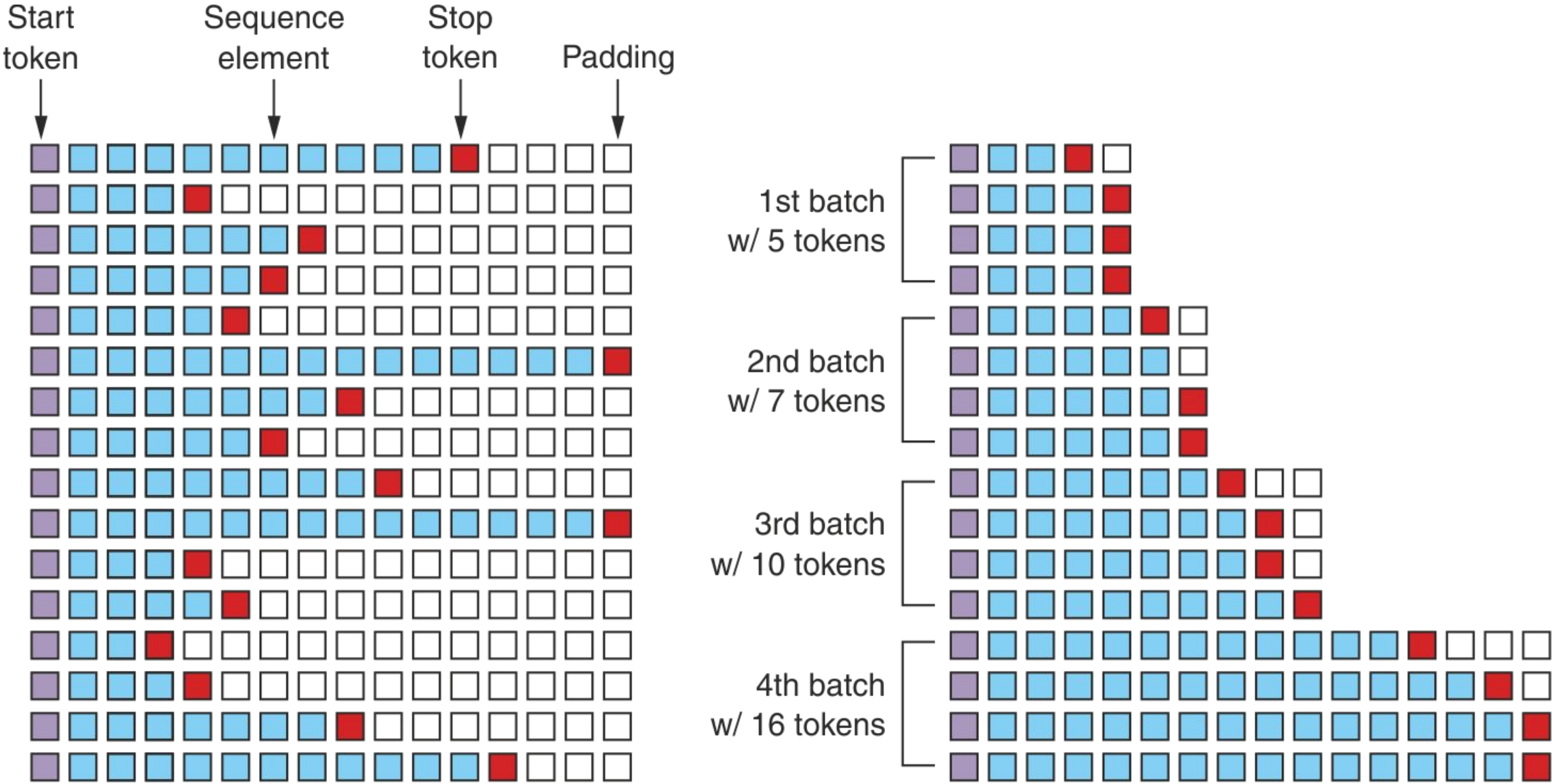
1 200



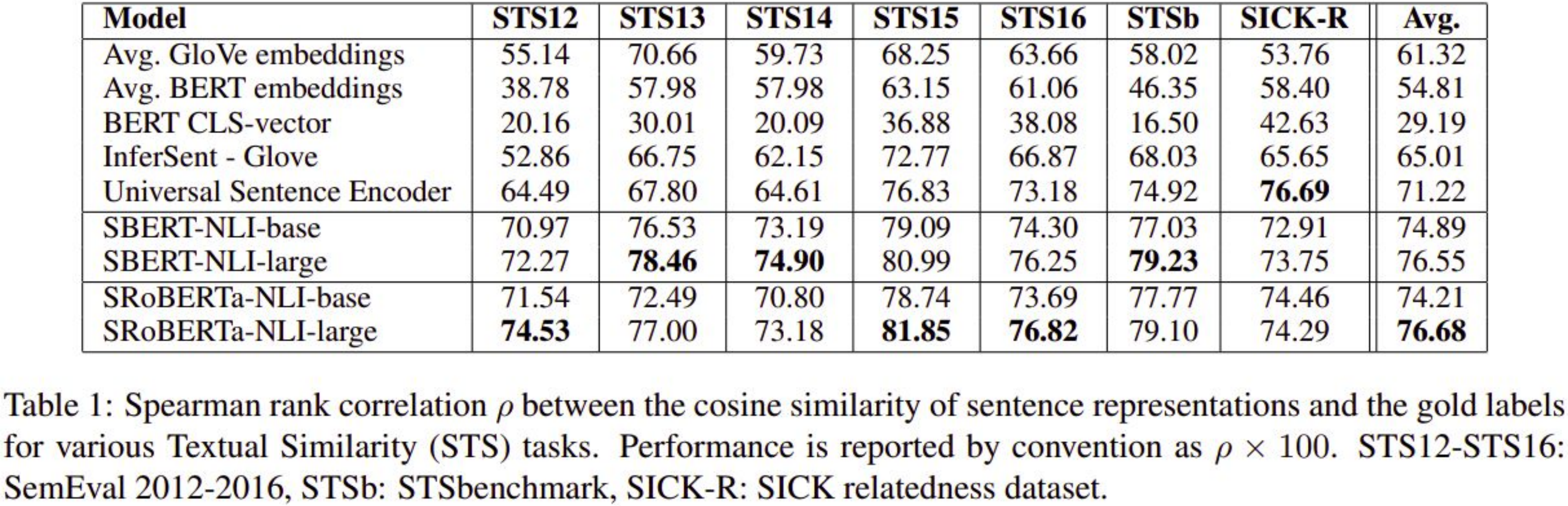
Model	CPU	GPU
Avg. GloVe embeddings	6469	-
InferSent	137	1876
Universal Sentence Encoder	67	1318
SBERT-base	44	1378
SBERT-base - smart batching	83	2042

Table 7: Computation speed (sentences per second) of sentence embedding methods. Higher is better.

Cerca de 9% mais rápido que o InferSent e 55% mais rápido que o Universal Sentence Encoder


Start token
Sequence element
Stop token
Padding
1st batch w/ 5 tokens
2nd batch w/ 7 tokens
3rd batch w/ 10 tokens
4th batch w/ 16 tokens

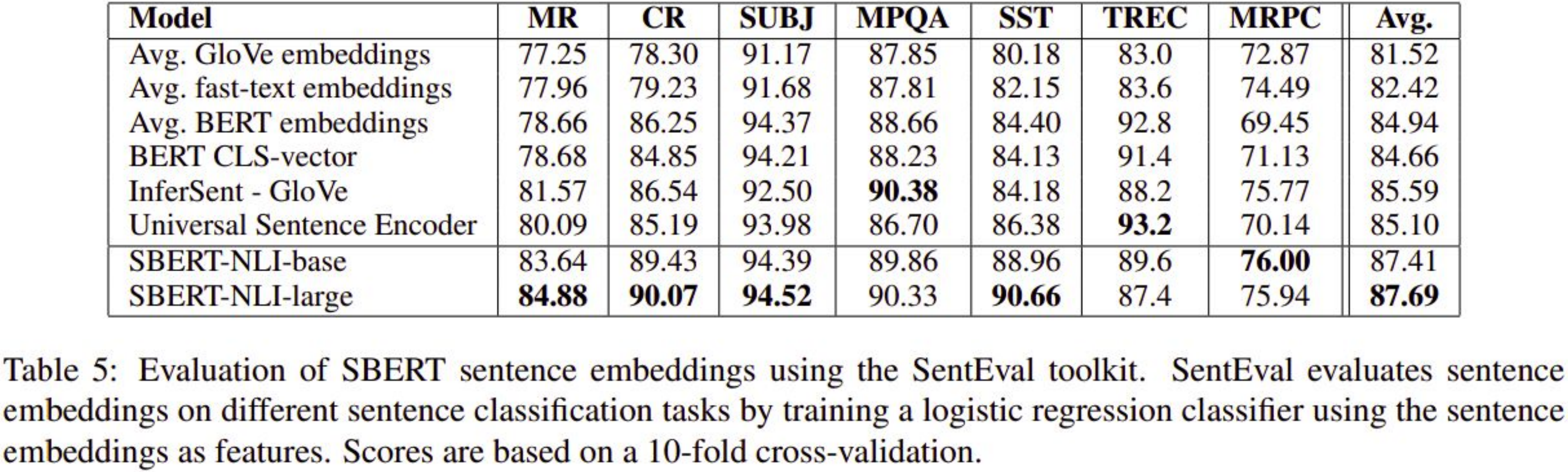
Similaridade Semântica



Model	STS12	STS13	STS14	STS15	STS16	STSb	SICK-R	Avg.
Avg. GloVe embeddings	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
Avg. BERT embeddings	38.78	57.98	57.98	63.15	61.06	46.35	58.40	54.81
BERT CLS-vector	20.16	30.01	20.09	36.88	38.08	16.50	42.63	29.19
InferSent - Glove	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
Universal Sentence Encoder	64.49	67.80	64.61	76.83	73.18	74.92	**76.69**	71.22
SBERT-NLI-base	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
SBERT-NLI-large	72.27	**78.46**	**74.90**	80.99	76.25	**79.23**	73.75	76.55
SRoBERTa-NLI-base	71.54	72.49	70.80	78.74	73.69	77.77	74.46	74.21
SRoBERTa-NLI-large	**74.53**	77.00	73.18	**81.85**	**76.82**	79.10	74.29	**76.68**

Table 1: Spearman rank correlation ρ between the cosine similarity of sentence representations and the gold labels for various Textual Similarity (STS) tasks. Performance is reported by convention as $\rho \times 100$. STS12-STS16: SemEval 2012-2016, STSb: STSbenchmark, SICK-R: SICK relatedness dataset.

Representação de sentenças



Model	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	Avg.
Avg. GloVe embeddings	77.25	78.30	91.17	87.85	80.18	83.0	72.87	81.52
Avg. fast-text embeddings	77.96	79.23	91.68	87.81	82.15	83.6	74.49	82.42
Avg. BERT embeddings	78.66	86.25	94.37	88.66	84.40	92.8	69.45	84.94
BERT CLS-vector	78.68	84.85	94.21	88.23	84.13	91.4	71.13	84.66
InferSent - GloVe	81.57	86.54	92.50	**90.38**	84.18	88.2	75.77	85.59
Universal Sentence Encoder	80.09	85.19	93.98	86.70	86.38	**93.2**	70.14	85.10
SBERT-NLI-base	83.64	89.43	94.39	89.86	88.96	89.6	**76.00**	87.41
SBERT-NLI-large	**84.88**	**90.07**	**94.52**	90.33	**90.66**	87.4	75.94	**87.69**

Table 5: Evaluation of SBERT sentence embeddings using the SentEval toolkit. SentEval evaluates sentence embeddings on different sentence classification tasks by training a logistic regression classifier using the sentence embeddings as features. Scores are based on a 10-fold cross-validation.

Sentence-Transformers

- Repositório oficial do Sentence-BERT
- Suporte aos transformers do HuggingFace
- Possui busca semântica implementada

Modos de treino:

- Treinar no NLI
- Treinar no NLI e depois no STS-B
- Treinar em multi-tasking NLI e STS-B

□ Portuguese-BERT disponível no HuggingFace

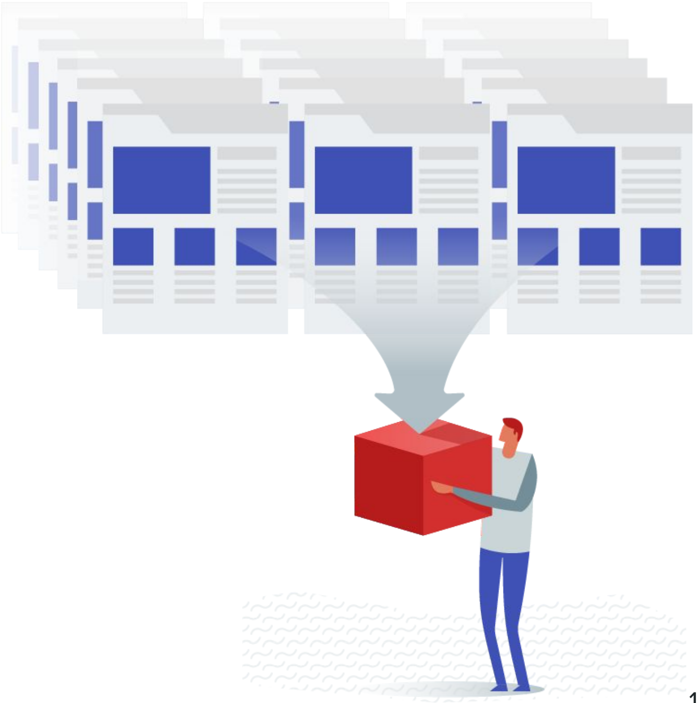
□ Dataset Avaliação de Similaridade Semântica e Textual (ASSIN) composto por tarefas análogas ao “NLI” e ao “STS-B”

□ SNLI e MNLI juntos formam cerca de 1M de sentenças enquanto ASSIN possui cerca de 10K

Aplicação em Bases de Conhecimento

Recuperação a partir de Bases de Conhecimento

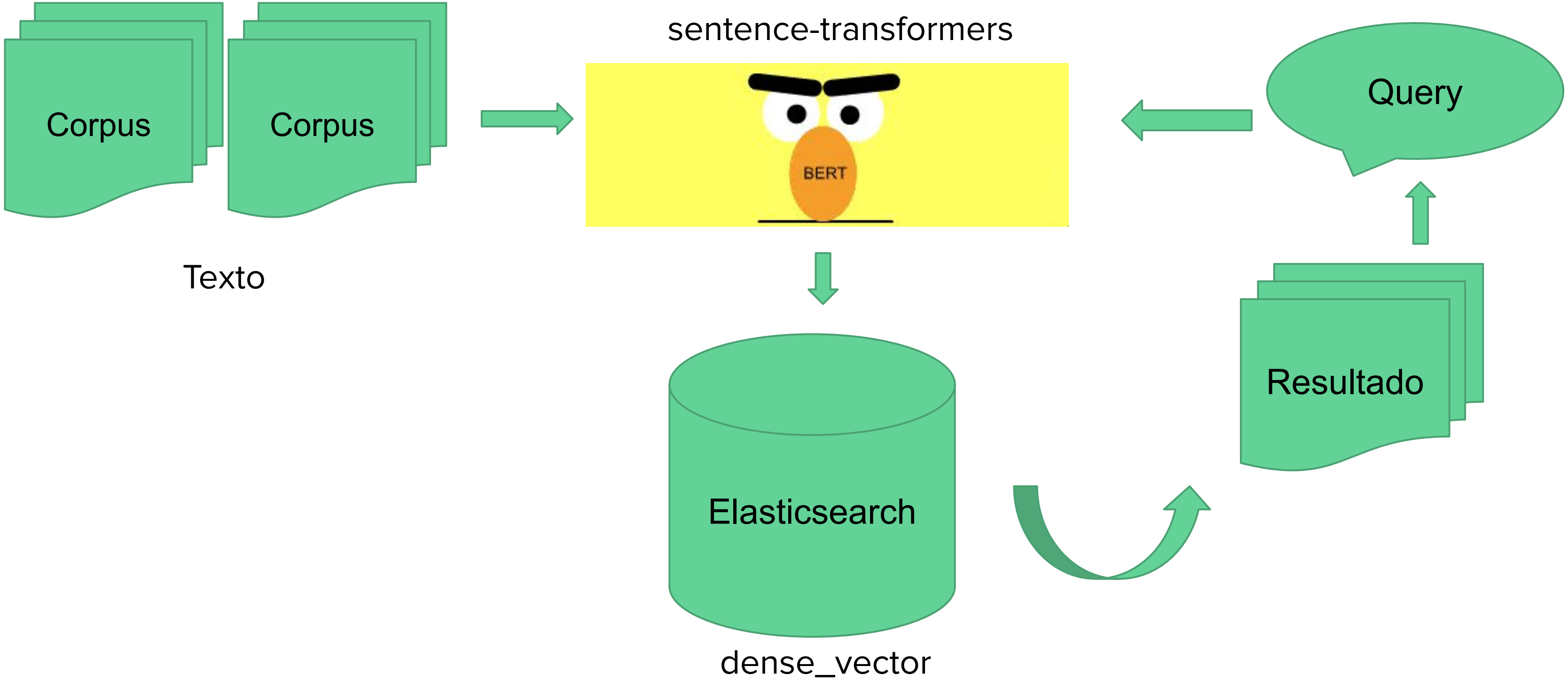
- **Aquisição dos Dados**
  - **Crawling / Scraping**
  - **Extração de Texto**
- **Pré-Processamento**
- Enriquecimento dos Dados
  - Modelagem de Tópicos
  - Extração de Entidades (NER)
  - Extração de Relações (RE)
  - Vinculação de Entidades (EL)
- **Recuperação de Informações**
  - **Busca Semântica**
  - **Perguntas e Respostas (Q&A)**

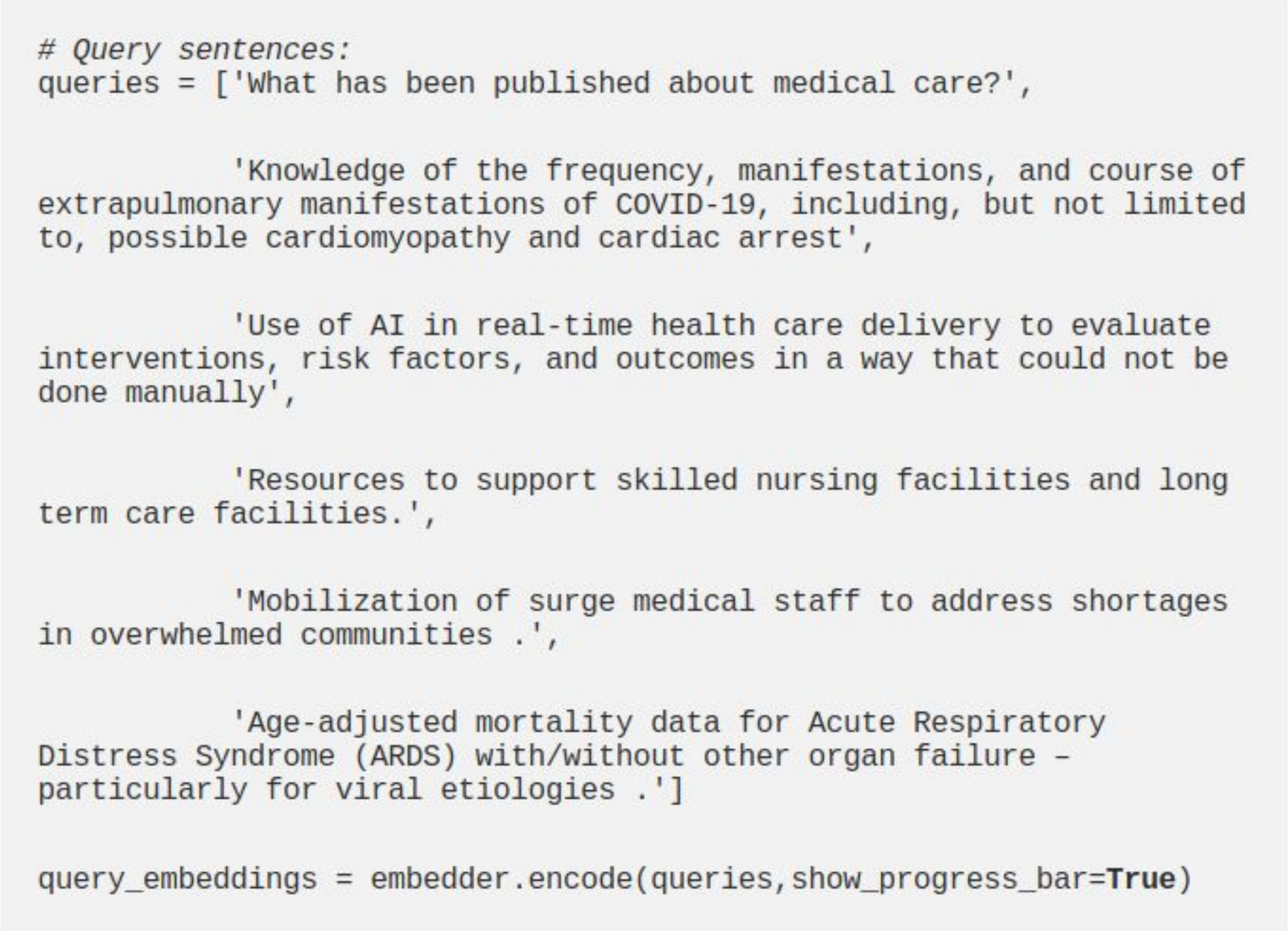


Pré-Processamento

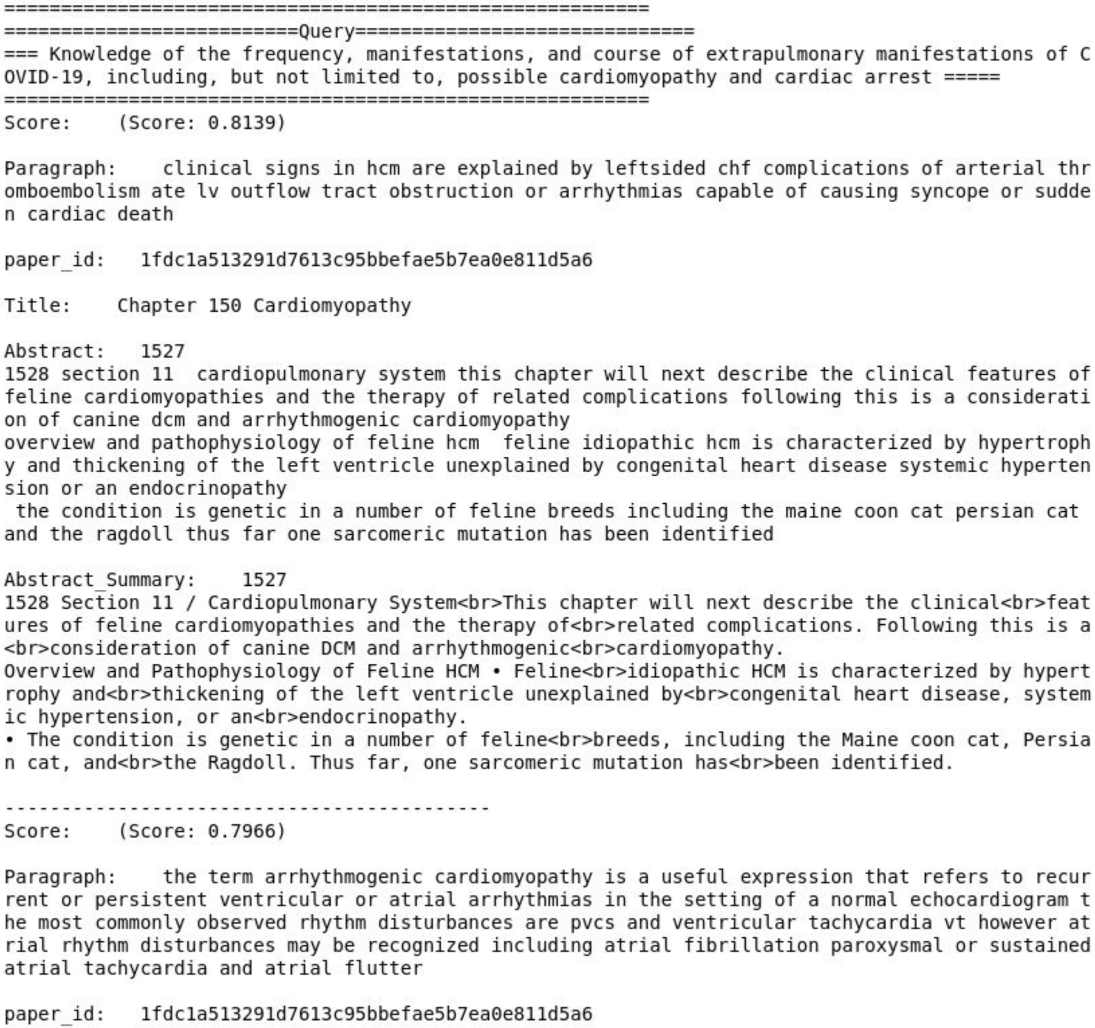
1. Busca de material
   a. Extração de PDFs
      i. PDFMiner
   b. Extração de HTMLs
      i. BeautifulSoup
2. Limpeza
3. Segmentação por Parágrafos
4. Indexação

Elasticsearch - dense vectors / cosineSimilarity





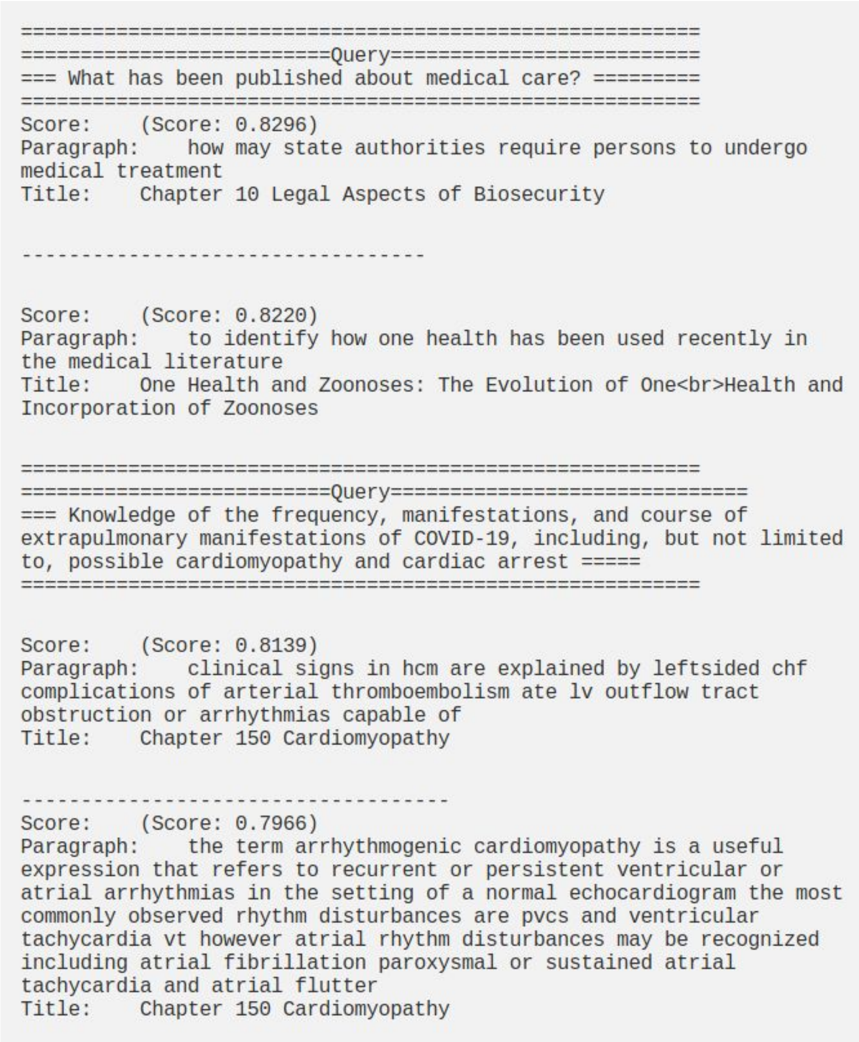
```
# Query sentences:
queries = ['What has been published about medical care?',

           'Knowledge of the frequency, manifestations, and course of
extrapulmonary manifestations of COVID-19, including, but not limited
to, possible cardiomyopathy and cardiac arrest',

           'Use of AI in real-time health care delivery to evaluate
interventions, risk factors, and outcomes in a way that could not be
done manually',

           'Resources to support skilled nursing facilities and long
term care facilities.',

           'Mobilization of surge medical staff to address shortages
in overwhelmed communities .',

           'Age-adjusted mortality data for Acute Respiratory
Distress Syndrome (ARDS) with/without other organ failure –
particularly for viral etiologies .']

query_embeddings = embedder.encode(queries,show_progress_bar=True)
```



```
=========================================================
==========================Query==============================
=== Knowledge of the frequency, manifestations, and course of extrapulmonary manifestations of C
OVID-19, including, but not limited to, possible cardiomyopathy and cardiac arrest =====
=========================================================
Score:    (Score: 0.8139)

Paragraph:    clinical signs in hcm are explained by leftsided chf complications of arterial thr
omboembolism ate lv outflow tract obstruction or arrhythmias capable of causing syncope or sudde
n cardiac death

paper_id:   1fdc1a513291d7613c95bbefae5b7ea0e811d5a6

Title:    Chapter 150 Cardiomyopathy

Abstract:   1527
1528 section 11  cardiopulmonary system this chapter will next describe the clinical features of
feline cardiomyopathies and the therapy of related complications following this is a considerati
on of canine dcm and arrhythmogenic cardiomyopathy
overview and pathophysiology of feline hcm  feline idiopathic hcm is characterized by hypertroph
y and thickening of the left ventricle unexplained by congenital heart disease systemic hyperten
sion or an endocrinopathy
 the condition is genetic in a number of feline breeds including the maine coon cat persian cat
and the ragdoll thus far one sarcomeric mutation has been identified

Abstract_Summary:    1527
1528 Section 11 / Cardiopulmonary System<br>This chapter will next describe the clinical<br>feat
ures of feline cardiomyopathies and the therapy of<br>related complications. Following this is a
<br>consideration of canine DCM and arrhythmogenic<br>cardiomyopathy.
Overview and Pathophysiology of Feline HCM • Feline<br>idiopathic HCM is characterized by hypert
rophy and<br>thickening of the left ventricle unexplained by<br>congenital heart disease, system
ic hypertension, or an<br>endocrinopathy.
• The condition is genetic in a number of feline<br>breeds, including the Maine coon cat, Persia
n cat, and<br>the Ragdoll. Thus far, one sarcomeric mutation has<br>been identified.

-------------------------------------------
Score:    (Score: 0.7966)

Paragraph:    the term arrhythmogenic cardiomyopathy is a useful expression that refers to recur
rent or persistent ventricular or atrial arrhythmias in the setting of a normal echocardiogram t
he most commonly observed rhythm disturbances are pvcs and ventricular tachycardia vt however at
rial rhythm disturbances may be recognized including atrial fibrillation paroxysmal or sustained
atrial tachycardia and atrial flutter

paper_id:   1fdc1a513291d7613c95bbefae5b7ea0e811d5a6
```

Q&A

- Question and Answering
- Linguagem mais natural
- Modelo de Similaridade para filtragem dos documentos da base
- Modelo de Q&A treinado para responder
- Transformers + SQuAD



```
=========================================================
==========================Query==========================
=== What has been published about medical care? =========
=========================================================
Score:    (Score: 0.8296)
Paragraph:    how may state authorities require persons to undergo
medical treatment
Title:    Chapter 10 Legal Aspects of Biosecurity


----------------------------------


Score:    (Score: 0.8220)
Paragraph:    to identify how one health has been used recently in
the medical literature
Title:    One Health and Zoonoses: The Evolution of One<br>Health and
Incorporation of Zoonoses


=========================================================
==========================Query==============================
=== Knowledge of the frequency, manifestations, and course of
extrapulmonary manifestations of COVID-19, including, but not limited
to, possible cardiomyopathy and cardiac arrest =====
=========================================================


Score:    (Score: 0.8139)
Paragraph:    clinical signs in hcm are explained by leftsided chf
complications of arterial thromboembolism ate lv outflow tract
obstruction or arrhythmias capable of
Title:    Chapter 150 Cardiomyopathy


------------------------------------
Score:    (Score: 0.7966)
Paragraph:    the term arrhythmogenic cardiomyopathy is a useful
expression that refers to recurrent or persistent ventricular or
atrial arrhythmias in the setting of a normal echocardiogram the most
commonly observed rhythm disturbances are pvcs and ventricular
tachycardia vt however atrial rhythm disturbances may be recognized
including atrial fibrillation paroxysmal or sustained atrial
tachycardia and atrial flutter
Title:    Chapter 150 Cardiomyopathy
```

Dúvidas?

- https://towardsdatascience.com/covid-19-bert-literature-search-engine-4d06cdac08bd
- https://towardsdatascience.com/covid-19-finetuned-bert-literature-search-engine-93ff9755a502
- https://github.com/UKPLab/sentence-transformers