Busca Semântica com Sentence-Transformers

Continuação dos Experimentos

Juliana Resplande

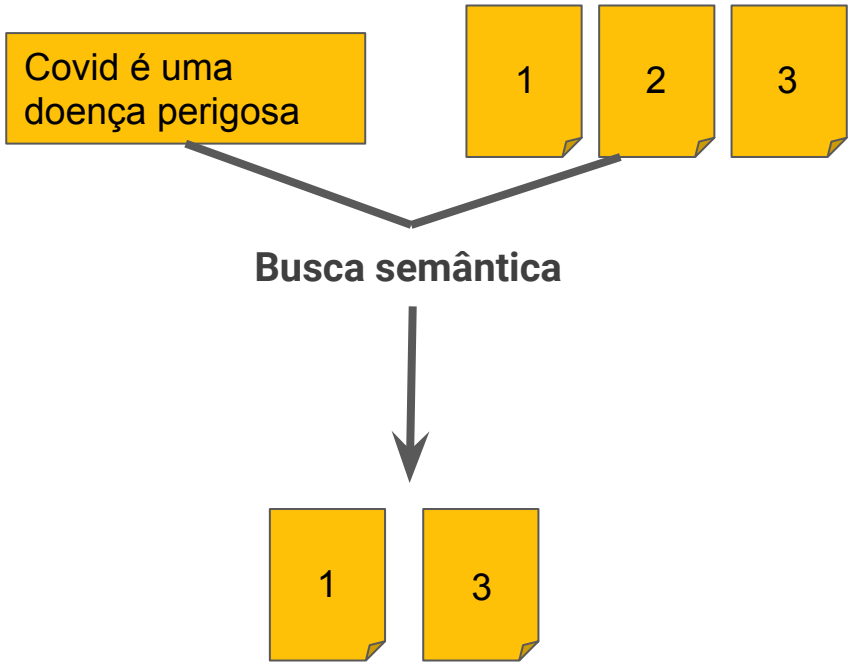
Agenda

1. Busca
2. Sentence Transformers
3. Novos experimentos
4. Conclusões

Busca semântica

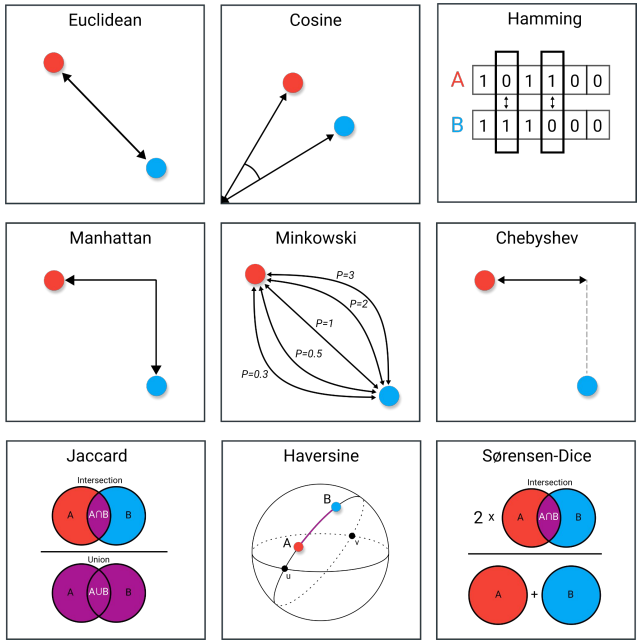
Busca semântica

- Information Retrieval (IR)
- Funcionamento:
  a. Gera as **representações semântica**s
  b. Compara o vetor da representação pela **distância**
- Dataset: MS-Marco



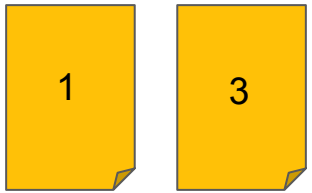
Considerações

- Como medir a **distância**?



- Considerar o **ranking** ou não?
  - Tarefa de similaridade
  - Tarefa de Ranqueamento

“O documento 1 é mais importante que o documento 3?”



Introdução

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

Sentence Transformers

Cross-Encoder

- **BERT**
- **Entrada:**
  - Concatenação das sentenças
- Função de regressão depende do modelo
- Devagar

X

Bi-encoder

- **SBERT**
- **Entrada:**
  - Sentenças paralelamente
- Representações independentes
- Pode usar similaridade cosseno
- Rápido

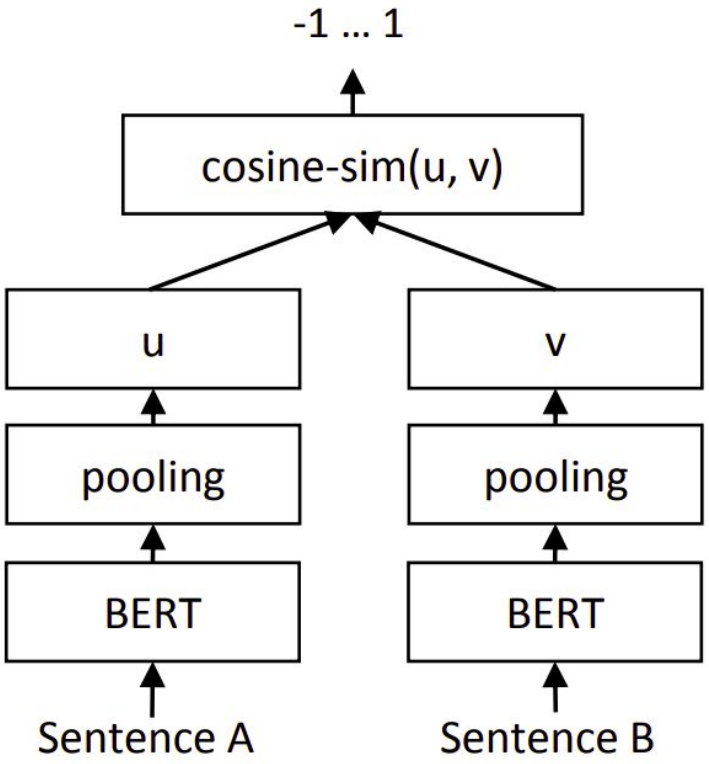
Sentence embeddings

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



Inferência Textual (NLI)

- **Classificação** de pares de sentenças
- Exemplos: QQP, QNLI, SNLI, MNLI, RTE
- **Treino em NLI** gera boas representações semânticas de texto

"Se consegue correlacionar bem quaisquer dois pares de sentenças, provavelmente correlacionará bem a query com o documento"

A soccer game with multiple males playing.

Some men are playing a sport.

entailment SNLI

What programming languages do what?

What do all programming languages do?

duplicate QQP

Sentence-Transformers

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

Modos de treino iniciais:

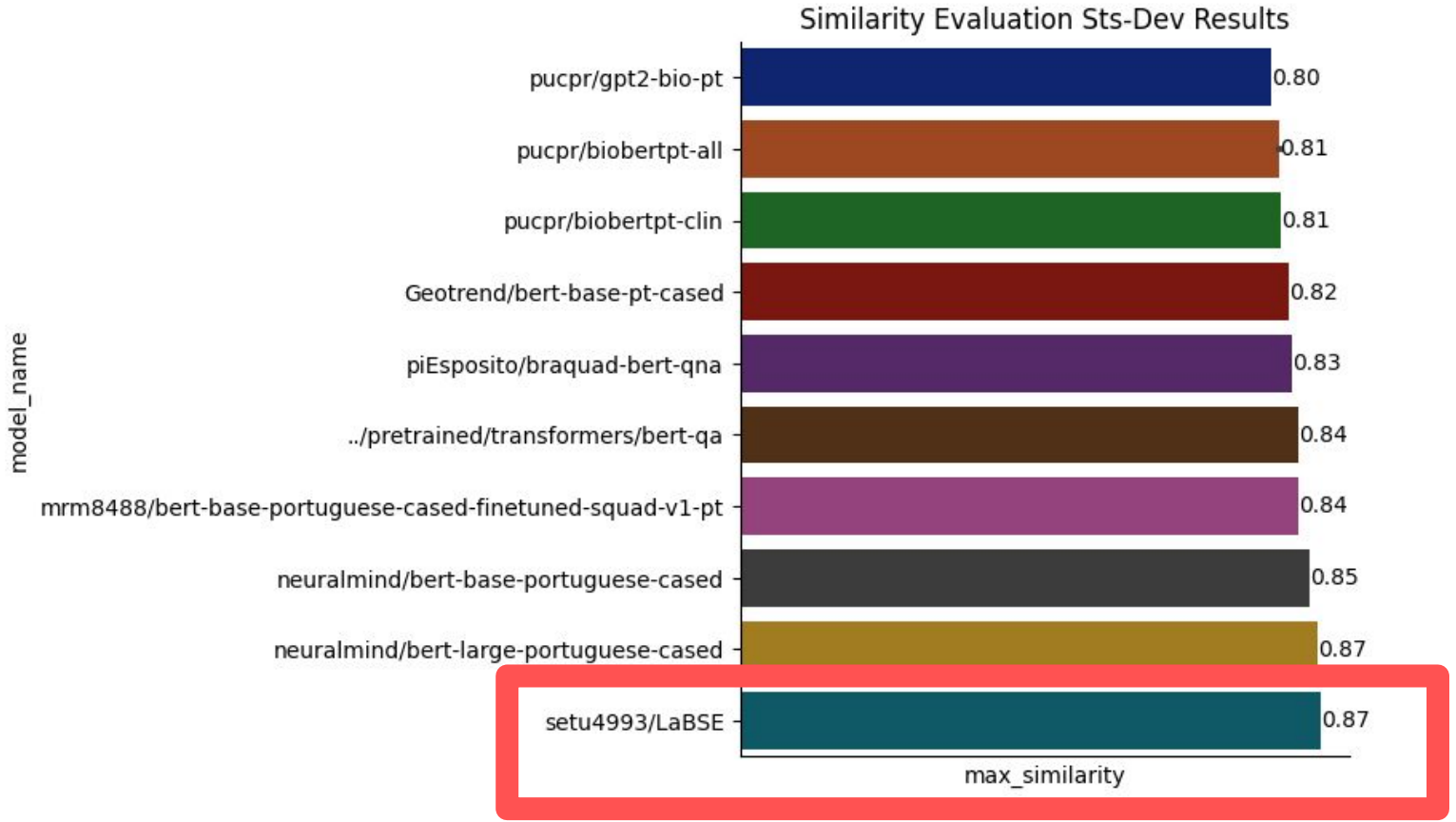
- Treinar no NLI
- Treinar no NLI e depois no STS-B
- Treinar em multi-tasking NLI e STS-B
- *Knowledge Distilation* em tradução

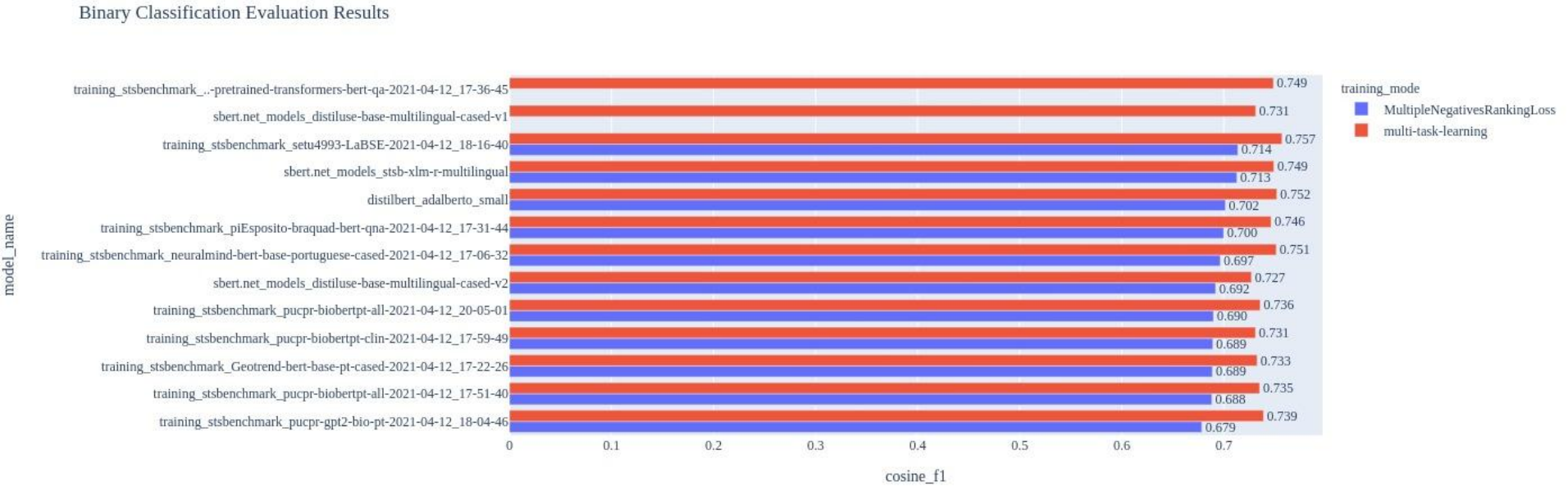
Novos experimentos

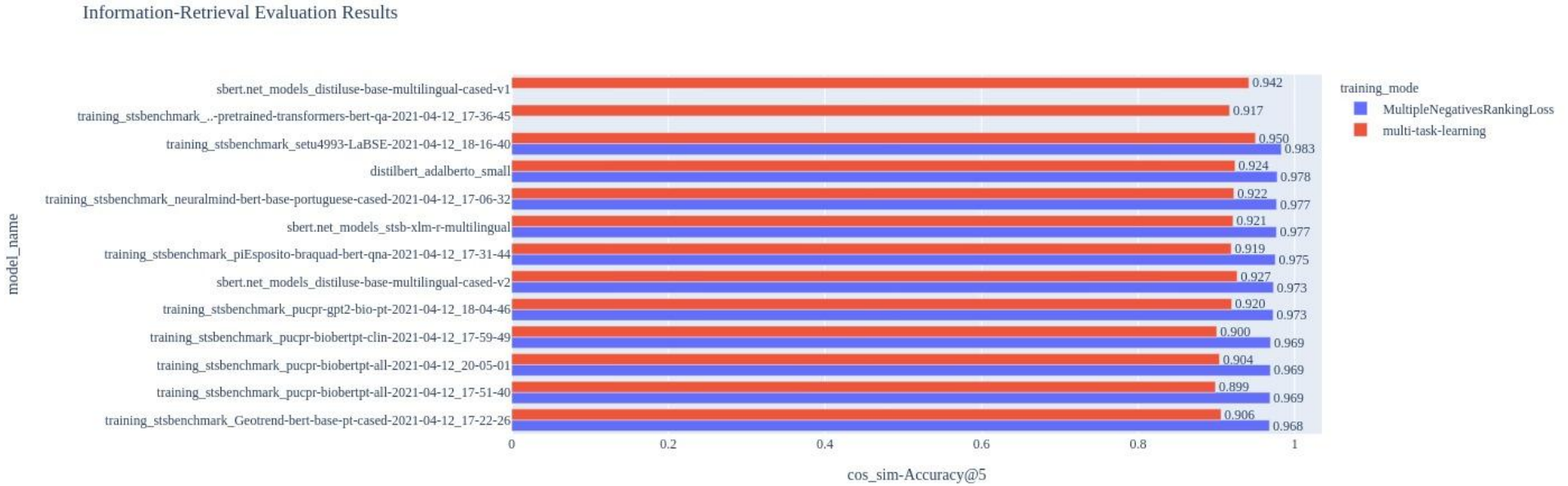
Treinos mais voltados para busca semântica

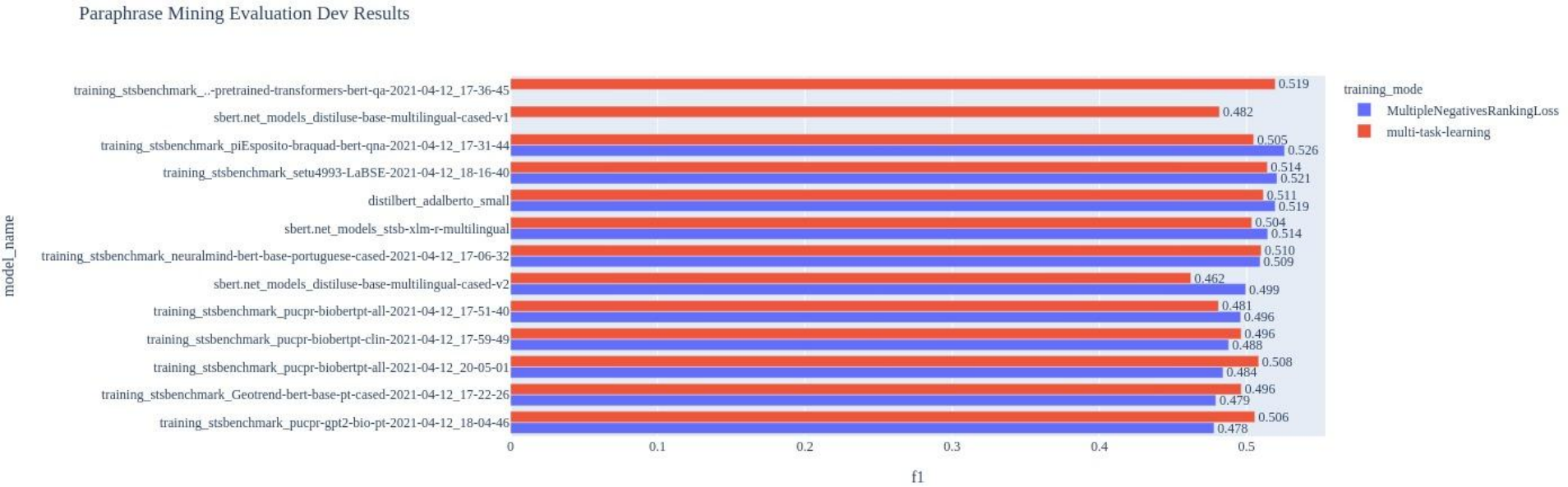
1. Treino pipeline STS + Quora Duplicate Questions
    a. Treino normal STS
    b. Treino no dataset do Quora modificado para busca semântica
        i. Duplicate Questions Classification: Given two questions, are these questions duplicates?
        ii. Duplicate Questions Mining: Given a large set (like 100k) of questions, identify all question pairs that are duplicates
        iii. Duplicate Questions Information Retrieval: Given a large corpus (350k+) of questions. For a new, unseen question, find the most related (i.e. duplicate) questions in this corpus.
2. Treino Ms-Marco
3. Treino DPR
4. Treino não-supervisionado

Resultados


Similarity Evaluation Sts-Dev Results
model_name
pucpr/gpt2-bio-pt
pucpr/biobertpt-all
pucpr/biobertpt-clin
Geotrend/bert-base-pt-cased
piEsposito/braquad-bert-qna
../pretrained/transformers/bert-qa
mrm8488/bert-base-portuguese-cased-finetuned-squad-v1-pt
neuralmind/bert-base-portuguese-cased
neuralmind/bert-large-portuguese-cased
setu4993/LaBSE
0.80
0.81
0.81
0.82
0.83
0.84
0.84
0.85
0.87
0.87
max_similarity


Binary Classification Evaluation Results
model_name
training_stsbenchmark_..-pretrained-transformers-bert-qa-2021-04-12_17-36-45
sbert.net_models_distiluse-base-multilingual-cased-v1
training_stsbenchmark_setu4993-LaBSE-2021-04-12_18-16-40
sbert.net_models_stsb-xlm-r-multilingual
distilbert_adalberto_small
training_stsbenchmark_piEsposito-braquad-bert-qna-2021-04-12_17-31-44
training_stsbenchmark_neuralmind-bert-base-portuguese-cased-2021-04-12_17-06-32
sbert.net_models_distiluse-base-multilingual-cased-v2
training_stsbenchmark_pucpr-biobertpt-all-2021-04-12_20-05-01
training_stsbenchmark_pucpr-biobertpt-clin-2021-04-12_17-59-49
training_stsbenchmark_Geotrend-bert-base-pt-cased-2021-04-12_17-22-26
training_stsbenchmark_pucpr-biobertpt-all-2021-04-12_17-51-40
training_stsbenchmark_pucpr-gpt2-bio-pt-2021-04-12_18-04-46
0.749
0.731
0.757
0.714
0.749
0.713
0.752
0.702
0.746
0.700
0.751
0.697
0.727
0.692
0.736
0.690
0.731
0.689
0.733
0.689
0.735
0.688
0.739
0.679
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
cosine_f1
training_mode
MultipleNegativesRankingLoss
multi-task-learning


Information-Retrieval Evaluation Results
model_name
sbert.net_models_distiluse-base-multilingual-cased-v1
training_stsbenchmark_..-pretrained-transformers-bert-qa-2021-04-12_17-36-45
training_stsbenchmark_setu4993-LaBSE-2021-04-12_18-16-40
distilbert_adalberto_small
training_stsbenchmark_neuralmind-bert-base-portuguese-cased-2021-04-12_17-06-32
sbert.net_models_stsb-xlm-r-multilingual
training_stsbenchmark_piEsposito-braquad-bert-qna-2021-04-12_17-31-44
sbert.net_models_distiluse-base-multilingual-cased-v2
training_stsbenchmark_pucpr-gpt2-bio-pt-2021-04-12_18-04-46
training_stsbenchmark_pucpr-biobertpt-clin-2021-04-12_17-59-49
training_stsbenchmark_pucpr-biobertpt-all-2021-04-12_20-05-01
training_stsbenchmark_pucpr-biobertpt-all-2021-04-12_17-51-40
training_stsbenchmark_Geotrend-bert-base-pt-cased-2021-04-12_17-22-26
0.942
0.917
0.950
0.983
0.924
0.978
0.922
0.977
0.921
0.977
0.919
0.975
0.927
0.973
0.920
0.973
0.900
0.969
0.904
0.969
0.899
0.969
0.906
0.968
0
0.2
0.4
0.6
0.8
1
cos_sim-Accuracy@5
training_mode
MultipleNegativesRankingLoss
multi-task-learning


Paraphrase Mining Evaluation Dev Results
model_name
training_stsbenchmark_..-pretrained-transformers-bert-qa-2021-04-12_17-36-45
sbert.net_models_distiluse-base-multilingual-cased-v1
training_stsbenchmark_piEsposito-braquad-bert-qna-2021-04-12_17-31-44
training_stsbenchmark_setu4993-LaBSE-2021-04-12_18-16-40
distilbert_adalberto_small
sbert.net_models_stsb-xlm-r-multilingual
training_stsbenchmark_neuralmind-bert-base-portuguese-cased-2021-04-12_17-06-32
sbert.net_models_distiluse-base-multilingual-cased-v2
training_stsbenchmark_pucpr-biobertpt-all-2021-04-12_17-51-40
training_stsbenchmark_pucpr-biobertpt-clin-2021-04-12_17-59-49
training_stsbenchmark_pucpr-biobertpt-all-2021-04-12_20-05-01
training_stsbenchmark_Geotrend-bert-base-pt-cased-2021-04-12_17-22-26
training_stsbenchmark_pucpr-gpt2-bio-pt-2021-04-12_18-04-46
0.519
0.482
0.505
0.526
0.514
0.521
0.511
0.519
0.504
0.514
0.510
0.509
0.462
0.499
0.481
0.496
0.496
0.488
0.508
0.484
0.496
0.479
0.506
0.478
0
0.1
0.2
0.3
0.4
0.5
f1
training_mode
MultipleNegativesRankingLoss
multi-task-learning

Conclusões

Conclusões/Trabalhos Futuros

- Importância de usar tarefas e métricas de ranqueamento
- LABSE: modelos multilíngues em detrimento de modelos monolíngues
- Finalizar o treino com o MS-Marco
- Treinar DPR
- Treino Não-Supervisionado
- Melhorar treino DeCLUTR