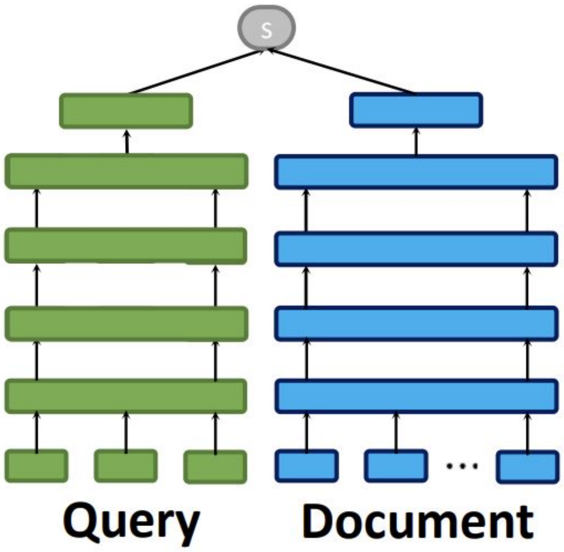
Towards Lossless Token Pruning in Late-Interaction Retrieval Models

Yuxuan Zong, Benjamin Piwowarski

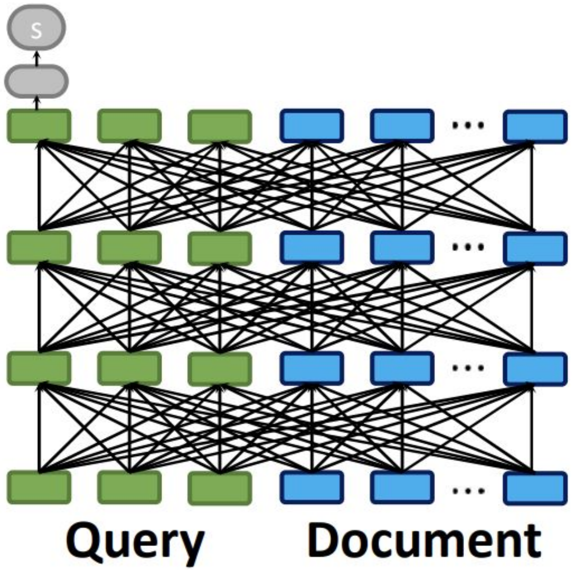
CNRS, ISIR, Sorbonne Université

Context

Information retrieval (IR) task: User gives a query, output a set of documents which are relevant → With the introduction of transformers, neural IR turns to be the state of the art

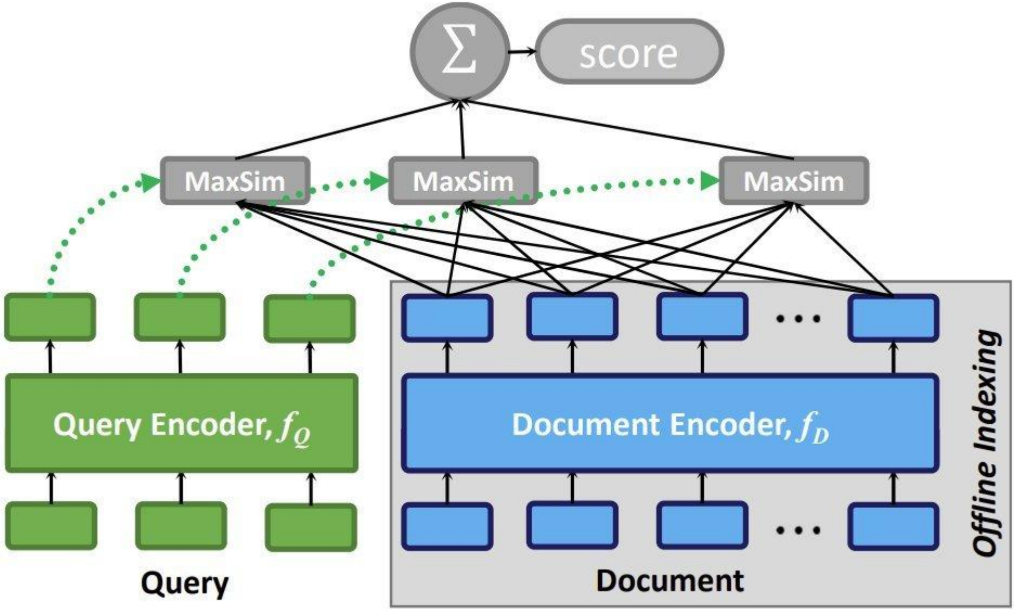


$$s = \frac{\mathbf{q} \cdot \mathbf{d}}{\|\mathbf{q}\|_2 \cdot \|\mathbf{d}\|_2}$$



$$s = \mathbf{W}\ \mathbf{e}_{[\mathrm{CLS}]}$$

Late Interaction Models



$$s = \sum_{\mathbf{q}\in\mathbf{Q}} \max_{\mathbf{d}\in\mathbf{D}} \mathbf{q}\cdot\mathbf{d}$$

However, late interaction models require a large index space

Related Work: Pruning Methods

- Clustering
- Pruning: Statistical or heuristic approach
  - Stopwords, IDF value, position-based methods (first-k), etc
  - Degrade the model performance with a high pruning ratio
- Pruning: “Learn to prune” approach
  - Gating mechanism: AligneR, LEAP_MV, etc
  - Learned pooling: ConstBERT
  - Better than static
- Is there some way to remove the tokens with no influence on the final score?

Intuition

Remove the $\mathbf{d} \in \mathbf{D}$ that never match any $\mathbf{q}$

$$\mathrm{ColBERT}(\mathbf{Q}, \mathbf{D}) = \sum_{\mathbf{q} \in \mathbf{Q}} \max_{\mathbf{d} \in \mathbf{D}} \mathbf{q} \cdot \mathbf{d}$$

Intuition

Remove the $\mathbf{d} \in \mathbf{D}$ that never match any $\mathbf{q}$

$$\mathrm{ColBERT}(\mathbf{Q}, \mathbf{D}) = \sum_{\mathbf{q} \in \mathbf{Q}} \boxed{\max_{\mathbf{d} \in \mathbf{D}} \mathbf{q} \cdot \mathbf{d}}$$

$$\forall \mathbf{q} \in \mathbb{R}^d, \boxed{\max_{\mathbf{d} \in \mathbf{D}} \mathbf{q} \cdot \mathbf{d}} = \max_{\mathbf{d} \in \mathbf{D}^+} \mathbf{q} \cdot \mathbf{d} \text{ with } \mathbf{D}^+ \subseteq \mathbf{D}$$

Intuition

Remove the $\mathbf{d} \in \mathbf{D}$ that never match any $\mathbf{q}$

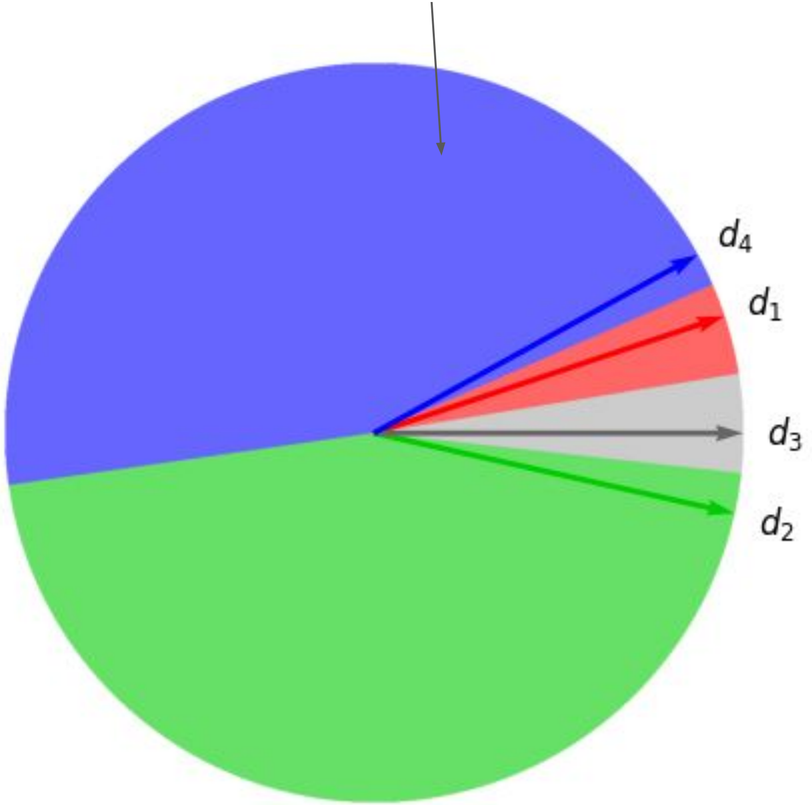
$$\mathrm{ColBERT}(\mathbf{Q}, \mathbf{D}) = \sum_{\mathbf{q} \in \mathbf{Q}} \boxed{\max_{\mathbf{d} \in \mathbf{D}} \mathbf{q} \cdot \mathbf{d}} \longrightarrow \mathrm{ColBERT}(\mathbf{Q}, \mathbf{D}) = \sum_{\mathbf{q} \in \mathbf{Q}} \max_{\mathbf{d} \in \mathbf{D}^+} \mathbf{q} \cdot \mathbf{d}$$

$$\forall \mathbf{q} \in \mathbb{R}^d, \boxed{\max_{\mathbf{d} \in \mathbf{D}} \mathbf{q} \cdot \mathbf{d}} = \max_{\mathbf{d} \in \mathbf{D}^+} \mathbf{q} \cdot \mathbf{d} \quad \text{with } \mathbf{D}^+ \subseteq \mathbf{D}$$

Adaptating ColBERT

$$\mathrm{ColBERT}(\mathbf{Q},\mathbf{D}) = \sum_{\mathbf{q}\in\mathbf{Q}} \max_{\mathbf{d}\in\mathbf{D}} \mathbf{q}\cdot\mathbf{d}$$

$$\underset{q\in\mathbb{R}^2}{\arg\max}\quad d_4 \cdot q$$



Adaptating ColBERT

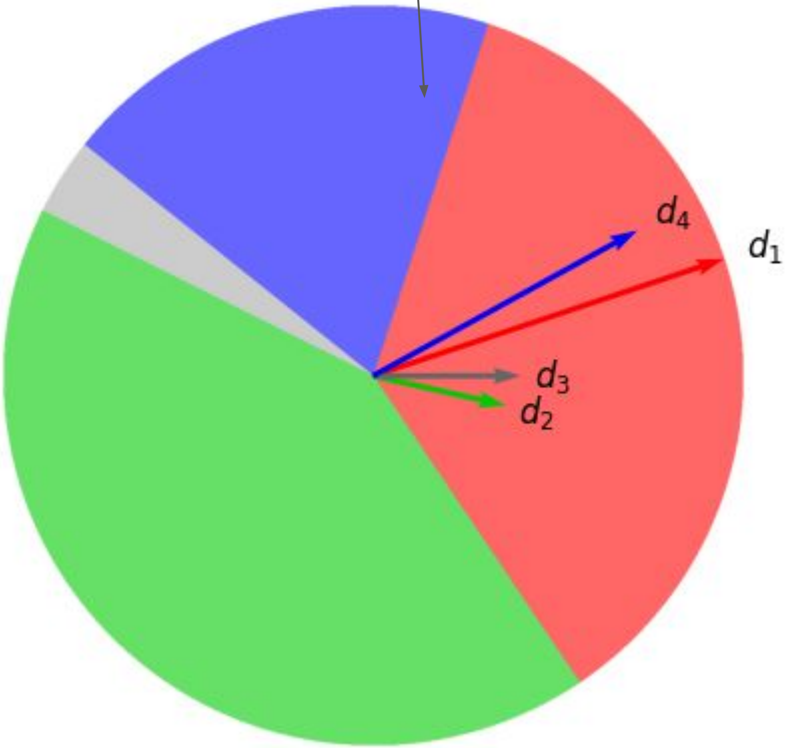
- Normalized token vectors → bounded token vectors

$$\underset{q \in \mathbb{R}^2}{\arg\max} \quad d_4 \cdot q$$

$$\mathrm{ColBERT}(\mathbf{Q}, \mathbf{D}) = \sum_{\mathbf{q} \in \mathbf{Q}} \max_{\mathbf{d} \in \mathbf{D}} \mathbf{q} \cdot \mathbf{d}$$

↓

$$\mathrm{ColBERT}_{\mathrm{P}}(\mathbf{Q}, \mathbf{D}) = \sum_{\mathbf{q} \in \mathbf{Q}} \max_{\mathbf{d} \in \mathbf{D}} \pi_\theta(\mathbf{q}) \cdot \pi_\theta(\mathbf{d})$$



Adaptating ColBERT

- Normalized token vectors → bounded token vectors
- Relued inner product value

$$\mathrm{ColBERT}(\mathbf{Q},\mathbf{D}) = \sum_{\mathbf{q}\in\mathbf{Q}} \max_{\mathbf{d}\in\mathbf{D}} \mathbf{q}\cdot\mathbf{d}$$

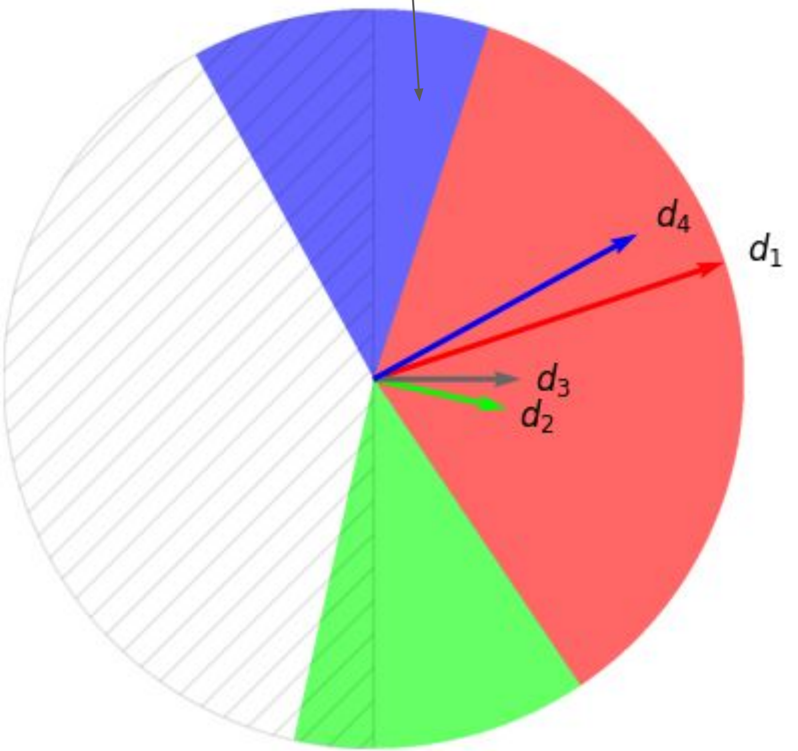
$$\downarrow$$

$$\mathrm{ColBERT_P}(\mathbf{Q},\mathbf{D}) = \sum_{\mathbf{q}\in\mathbf{Q}} \max_{\mathbf{d}\in\mathbf{D}} \pi_\theta(\mathbf{q})\cdot\pi_\theta(\mathbf{d})$$

$$\downarrow$$

$$\mathrm{ColBERT_P}(\mathbf{Q},\mathbf{D}) = \sum_{\mathbf{q}\in\mathbf{Q}} \max_{\mathbf{d}\in\mathbf{D}} [\pi_\theta(\mathbf{q})\cdot\pi_\theta(\mathbf{d})]_+$$

$$\arg\max_{q\in\mathbb{R}^2} \quad d_4\cdot q \quad \text{and } d_4\cdot q \geq 0$$



Dominance

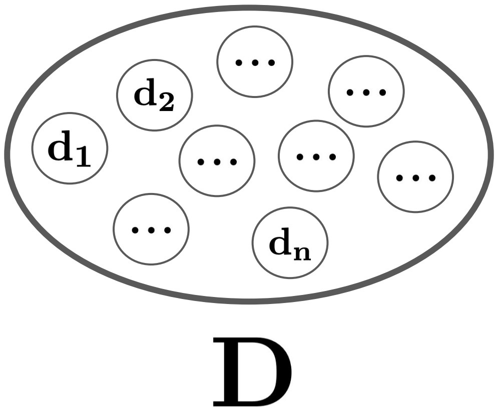
DEFINITION 1 (LOCAL DOMINANCE). *A vector* $\boldsymbol{d}^- \in \mathbb{R}^d$ *is* dominated *by a set of vectors* $\mathbf{D}$*, which we denote by* $\boldsymbol{d}^- \prec \mathbf{D}$*, if and only if for any* $\boldsymbol{q} \in \mathbb{R}^d$,

$$\text{either } \boldsymbol{q} \cdot \boldsymbol{d}^- \leq 0 \text{ or } \exists \boldsymbol{d}^+ \in \boldsymbol{D} \text{ such that } \boldsymbol{q} \cdot \boldsymbol{d}^+ > \boldsymbol{q} \cdot \boldsymbol{d}^- \qquad (3)$$

Dominance

Definition 1 (Local dominance). *A vector* $\mathbf{d}^- \in \mathbb{R}^d$ *is* dominated *by a set of vectors* $\mathbf{D}$*, which we denote by* $\mathbf{d}^- \prec \mathbf{D}$*, if and only if for any* $\mathbf{q} \in \mathbb{R}^d$,

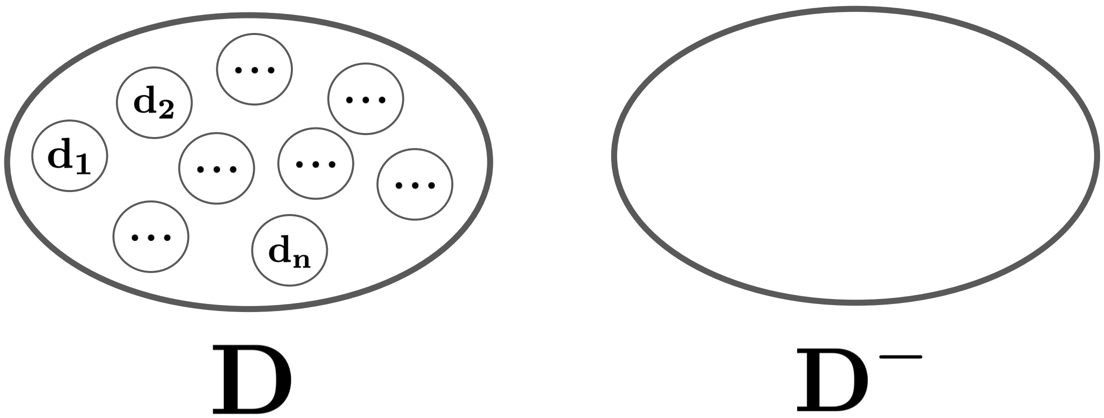
$$\text{either } \mathbf{q} \cdot \mathbf{d}^- \leq 0 \text{ or } \exists \mathbf{d}^+ \in \mathbf{D} \text{ such that } \mathbf{q} \cdot \mathbf{d}^+ > \mathbf{q} \cdot \mathbf{d}^- \quad (3)$$



Dominance

Definition 1 (Local dominance). *A vector* $\boldsymbol{d}^- \in \mathbb{R}^d$ *is* dominated *by a set of vectors* $\mathbf{D}$*, which we denote by* $\mathbf{d}^- \prec \mathbf{D}$*, if and only if for any* $\boldsymbol{q} \in \mathbb{R}^d$*,*

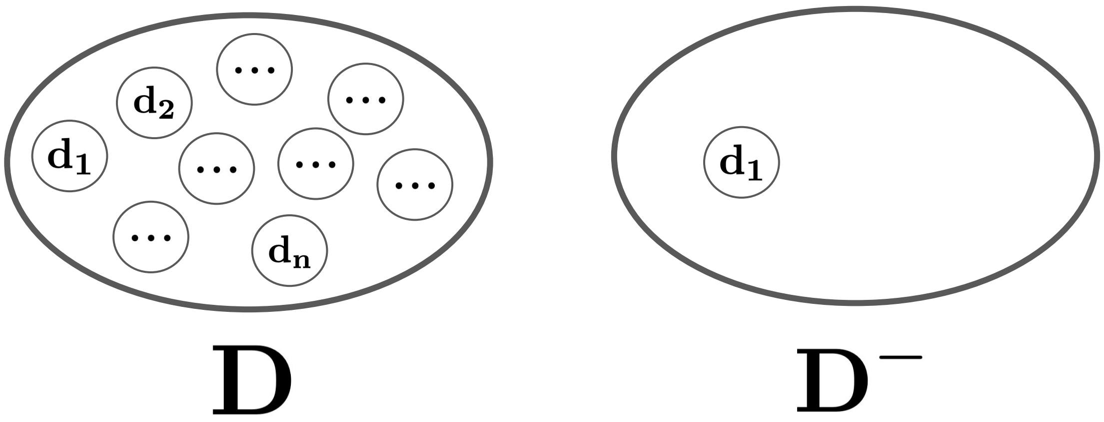
$$\text{either } \boldsymbol{q} \cdot \boldsymbol{d}^- \leq 0 \text{ or } \exists \boldsymbol{d}^+ \in \mathbf{D} \text{ such that } \boldsymbol{q} \cdot \boldsymbol{d}^+ > \boldsymbol{q} \cdot \boldsymbol{d}^- \tag{3}$$



Dominance

Definition 1 (Local dominance). *A vector* $\boldsymbol{d}^- \in \mathbb{R}^d$ *is* dominated *by a set of vectors* $\mathbf{D}$*, which we denote by* $\boldsymbol{d}^- \prec \mathbf{D}$*, if and only if for any* $\boldsymbol{q} \in \mathbb{R}^d$,

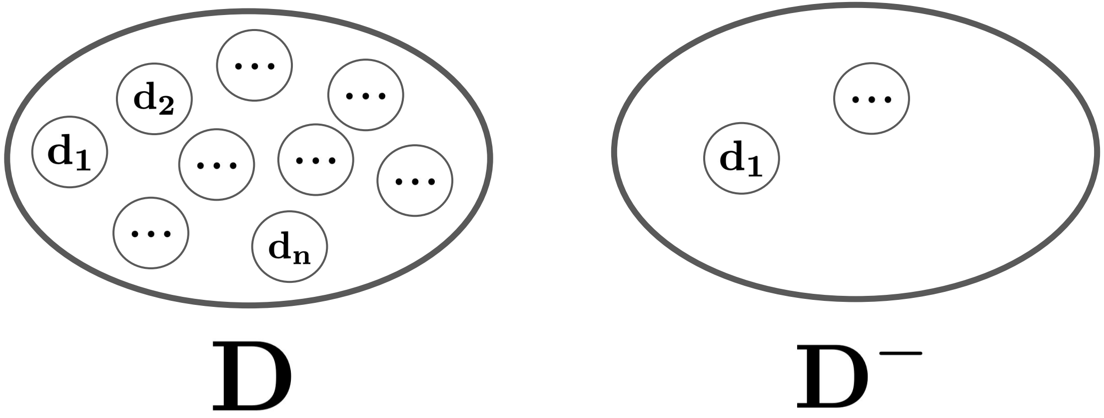
$$\textit{either } \boldsymbol{q} \cdot \boldsymbol{d}^- \leq 0 \textit{ or } \exists \boldsymbol{d}^+ \in \mathbf{D} \textit{ such that } \boldsymbol{q} \cdot \boldsymbol{d}^+ > \boldsymbol{q} \cdot \boldsymbol{d}^- \tag{3}$$



Dominance

Definition 1 (Local dominance). *A vector* $\boldsymbol{d}^{-} \in \mathbb{R}^{d}$ *is* dominated *by a set of vectors* $\mathbf{D}$*, which we denote by* $\boldsymbol{d}^{-} \prec \mathbf{D}$*, if and only if for any* $\boldsymbol{q} \in \mathbb{R}^{d}$*,*

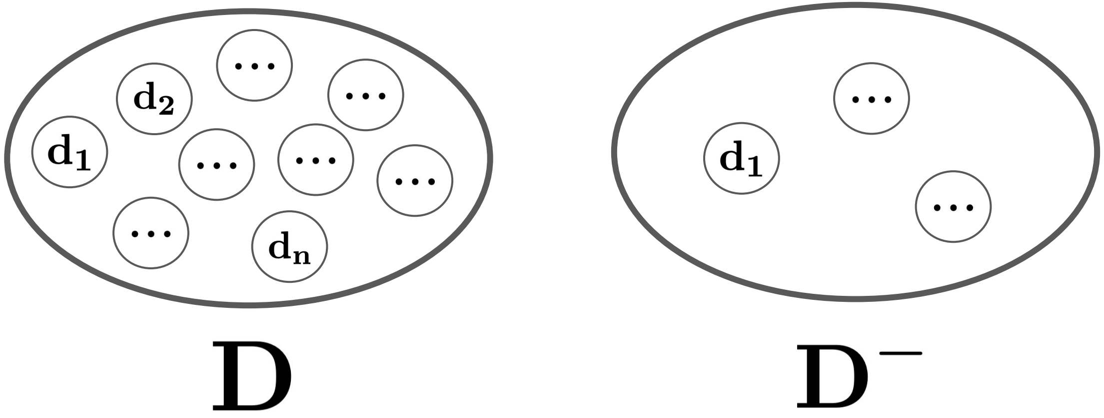
$$\text{either } \boldsymbol{q} \cdot \boldsymbol{d}^{-} \leq 0 \text{ or } \exists \boldsymbol{d}^{+} \in \mathbf{D} \text{ such that } \boldsymbol{q} \cdot \boldsymbol{d}^{+} > \boldsymbol{q} \cdot \boldsymbol{d}^{-} \tag{3}$$



Dominance

Definition 1 (Local dominance). *A vector* $\boldsymbol{d}^- \in \mathbb{R}^d$ *is* dominated *by a set of vectors* $\mathbf{D}$*, which we denote by* $\boldsymbol{d}^- \prec \mathbf{D}$*, if and only if for any* $\boldsymbol{q} \in \mathbb{R}^d$*,*

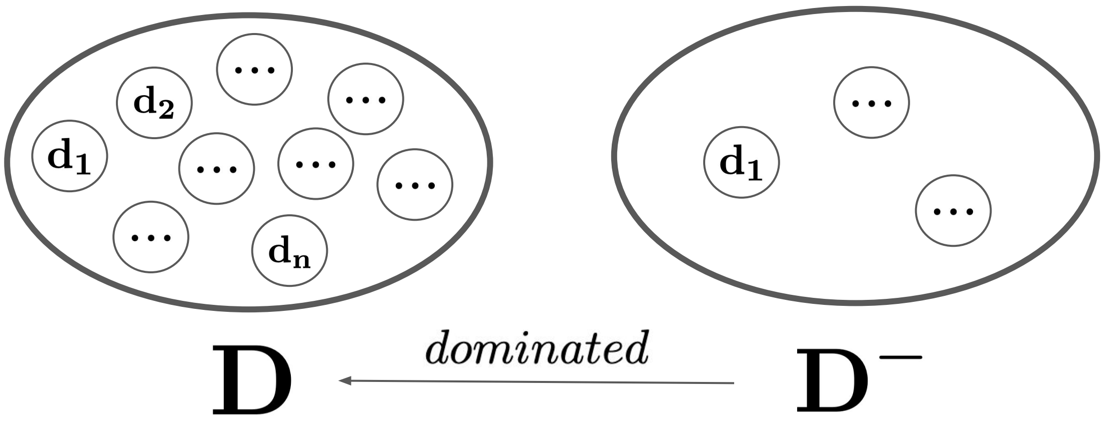
$$\text{either } \boldsymbol{q} \cdot \boldsymbol{d}^- \le 0 \text{ or } \exists \boldsymbol{d}^+ \in \mathbf{D} \text{ such that } \boldsymbol{q} \cdot \boldsymbol{d}^+ > \boldsymbol{q} \cdot \boldsymbol{d}^- \tag{3}$$



Dominance

Definition 1 (Local dominance). *A vector* $\boldsymbol{d}^- \in \mathbb{R}^d$ *is dominated by a set of vectors* $\mathbf{D}$*, which we denote by* $\boldsymbol{d}^- \prec \mathbf{D}$*, if and only if for any* $\boldsymbol{q} \in \mathbb{R}^d$,

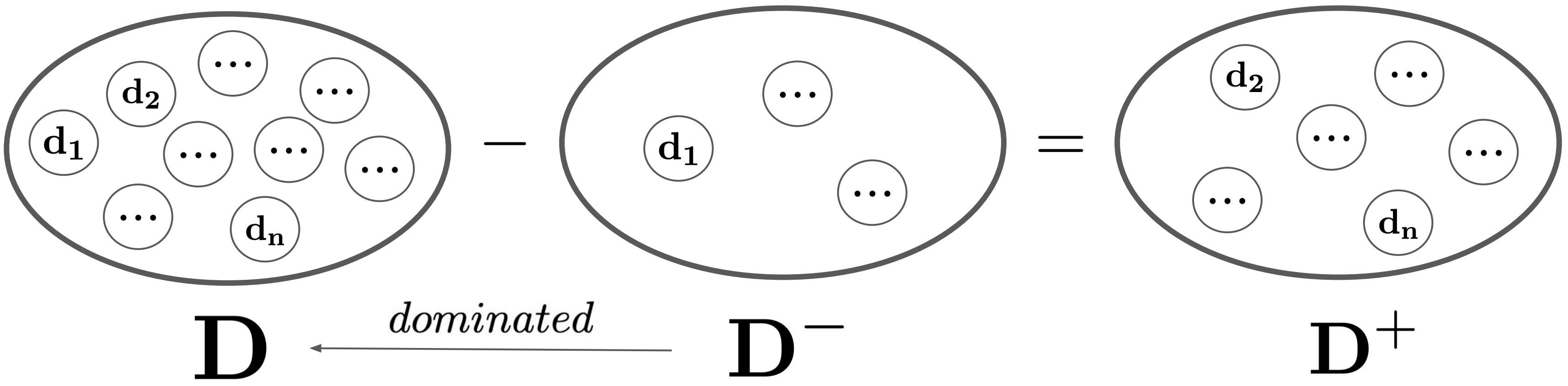
$$\textit{either } \boldsymbol{q} \cdot \boldsymbol{d}^- \leq 0 \textit{ or } \exists \boldsymbol{d}^+ \in \mathbf{D} \textit{ such that } \boldsymbol{q} \cdot \boldsymbol{d}^+ > \boldsymbol{q} \cdot \boldsymbol{d}^- \tag{3}$$



Dominance

DEFINITION 1 (LOCAL DOMINANCE). *A vector* $\boldsymbol{d}^- \in \mathbb{R}^d$ *is* dominated *by a set of vectors* $\mathbf{D}$, *which we denote by* $\boldsymbol{d}^- \prec \mathbf{D}$, *if and only if for any* $\boldsymbol{q} \in \mathbb{R}^d$,

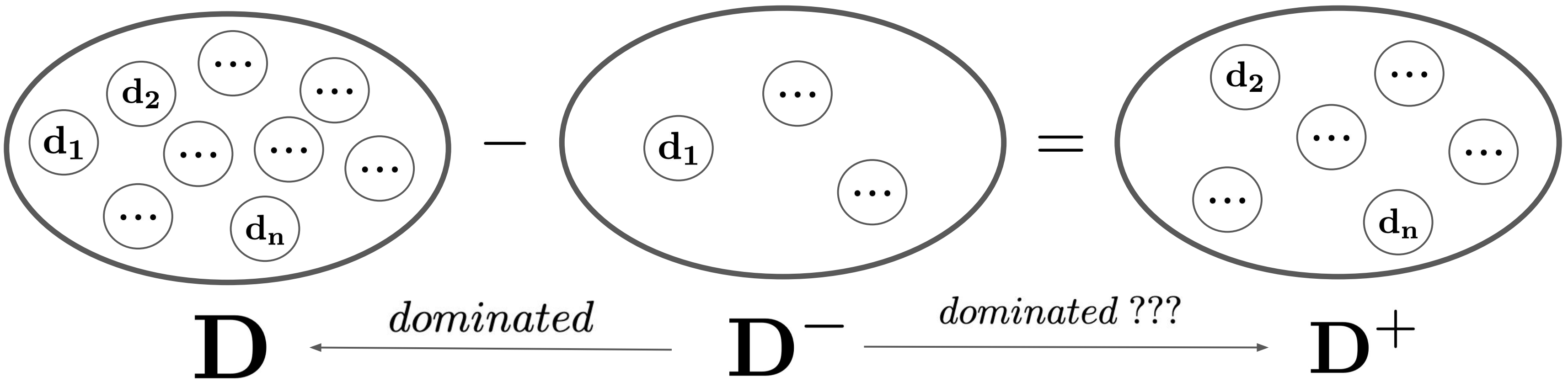
$$\text{either } \boldsymbol{q} \cdot \boldsymbol{d}^- \leq 0 \text{ or } \exists \boldsymbol{d}^+ \in \mathbf{D} \text{ such that } \boldsymbol{q} \cdot \boldsymbol{d}^+ > \boldsymbol{q} \cdot \boldsymbol{d}^- \tag{3}$$



Dominance

Definition 1 (Local dominance). *A vector* $\boldsymbol{d}^- \in \mathbb{R}^d$ *is* dominated *by a set of vectors* $\mathbf{D}$*, which we denote by* $\boldsymbol{d}^- \prec \mathbf{D}$*, if and only if for any* $\boldsymbol{q} \in \mathbb{R}^d$,

$$\text{either } \boldsymbol{q} \cdot \boldsymbol{d}^- \leq 0 \text{ or } \exists \boldsymbol{d}^+ \in \boldsymbol{D} \text{ such that } \boldsymbol{q} \cdot \boldsymbol{d}^+ > \boldsymbol{q} \cdot \boldsymbol{d}^- \tag{3}$$

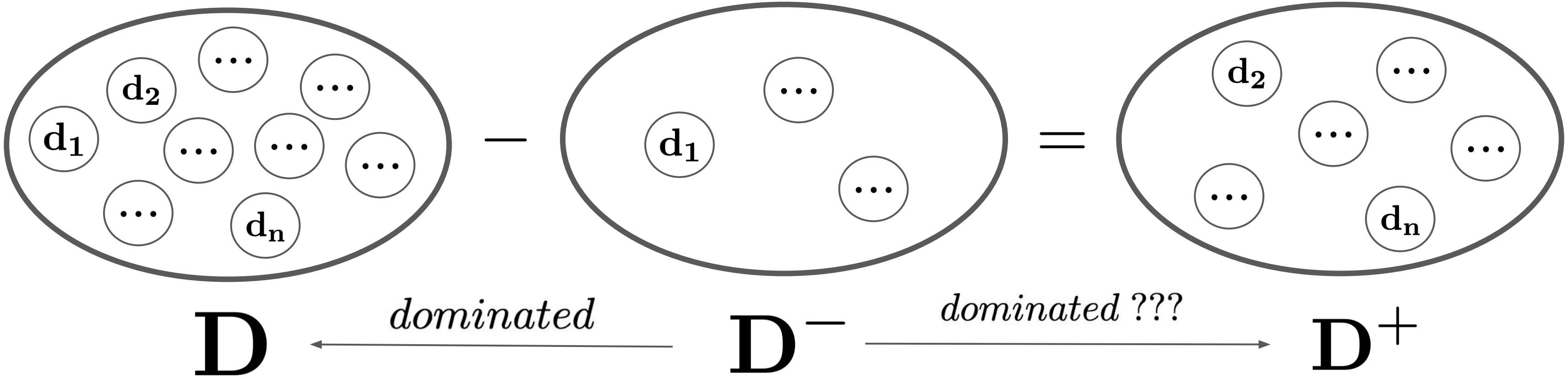


Dominance

Lemma 1 (Global dominance). *Let* $\mathbf{D}^{+}$ *and* $\mathbf{D}^{-}$ *be a partition of the set of document vectors* $\mathbf{D} \subset \mathbb{R}^{d}$ *such that any document vector* $\mathbf{d} \in \mathbf{D}^{-}$ *is dominated by* $\mathbf{D}$*. Formally,*

$$\mathbf{D}^{-} = \{\mathbf{d} \in \mathbf{D} | \mathbf{d} \prec \mathbf{D}\} \;\; \textit{and}\; \mathbf{D}^{+} = \mathbf{D} \setminus \mathbf{D}^{-}$$

Then, for any document $\mathbf{d}^{-} \in \mathbf{D}^{-}$, $\mathbf{d}^{-}$ *is dominated by* $\mathbf{D}^{+}$.

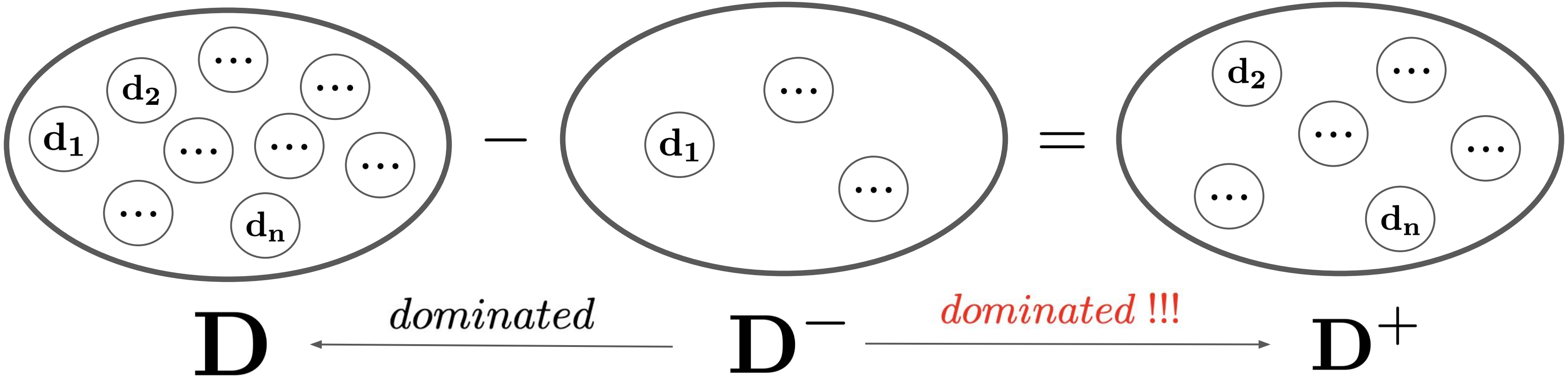


Dominance

Lemma 1 (Global dominance). *Let* $\mathbf{D}^{+}$ *and* $\mathbf{D}^{-}$ *be a partition of the set of document vectors* $\mathbf{D} \subset \mathbb{R}^{d}$ *such that any document vector* $\mathbf{d} \in \mathbf{D}^{-}$ *is dominated by* $\mathbf{D}$*. Formally,*

$$\mathbf{D}^{-} = \{\mathbf{d} \in \mathbf{D} | \mathbf{d} \prec \mathbf{D}\} \ \textit{and}\ \mathbf{D}^{+} = \mathbf{D} \setminus \mathbf{D}^{-}$$

Then, for any document $\mathbf{d}^{-} \in \mathbf{D}^{-}$, $\mathbf{d}^{-}$ *is dominated by* $\mathbf{D}^{+}$.



Approximate Pruning

- Dominance ←→ Linear Programming
    - Pruning based on the LP
    - Almost no solutions on full vectors
        - Apply SVD to reduce the dimension
- LP is a time consuming problem
    - Norm based pruning

Training Regularisation

- Nuclear norm regularization

$$\mathcal{L}_r^{(\mathrm{nuc})} = \frac{1}{\min(n, d)} \|\mathbf{D}\|_{nuc}$$

Training Regularisation

- Nuclear norm regularization

$$\mathcal{L}_r^{(\mathrm{nuc})} = \frac{1}{\min(n, d)} \|\mathbf{D}\|_{nuc}$$

- Document tokens similarity matrix regularization

$$\mathcal{L}_r^{(\mathrm{sim})} = -\frac{1}{n(n-1)} \sum_{\mathbf{d} \in \mathbf{D}} \sum_{\mathbf{d}' \in \mathbf{D} \setminus \{\mathbf{d}\}} [\mathbf{d} \cdot \mathbf{d}']_+$$

Training Regularisation

- Nuclear norm regularization

$$\mathcal{L}_r^{(\text{nuc})} = \frac{1}{\min(n, d)} \|\mathbf{D}\|_{nuc}$$

- Document tokens similarity matrix regularization

$$\mathcal{L}_r^{(\text{sim})} = -\frac{1}{n(n-1)} \sum_{\mathbf{d} \in \mathbf{D}} \sum_{\mathbf{d}' \in \mathbf{D} \setminus \{\mathbf{d}\}} \frac{[\mathbf{d} \cdot \mathbf{d}']_+}{\|\mathbf{d}\|_2 + \varepsilon}$$

Training Regularisation

- Nuclear norm regularization

$$\mathcal{L}_r^{(\mathrm{nuc})} = \frac{1}{\min(n, d)} \|\mathbf{D}\|_{nuc}$$

- Document tokens similarity matrix regularization

$$\mathcal{L}_r^{(\mathrm{sim})} = -\frac{1}{n(n-1)} \sum_{\mathbf{d} \in \mathbf{D}} (1 - \|\mathbf{d}\|_2) \sum_{\mathbf{d}' \in \mathbf{D} \setminus \{\mathbf{d}\}} \frac{[\mathbf{d} \cdot \mathbf{d}']_+}{\|\mathbf{d}\|_2 + \varepsilon}$$

Training Regularisation

- Nuclear norm regularization

$$\mathcal{L}_r^{(\mathrm{nuc})} = \frac{1}{\min(n, d)} \|\mathbf{D}\|_{nuc}$$

- Document tokens similarity matrix regularization

$$\mathcal{L}_r^{(\mathrm{sim})} = -\frac{1}{n(n-1)} \sum_{\mathbf{d} \in \mathbf{D}} (1 - \|\mathbf{d}\|_2) \sum_{\mathbf{d}' \in \mathbf{D} \setminus \{\mathbf{d}\}} \frac{[\mathbf{d} \cdot \mathbf{d}']_+}{\|\mathbf{d}\|_2 + \varepsilon}$$

- L1 norm

$$\mathcal{L}_r^{(L1)} = \frac{1}{n} \sum_{\mathbf{d} \in \mathbf{D}} \|\mathbf{d}\|_1$$

Training Regularisation

- Nuclear norm regularization

$$\mathcal{L}_r^{(\mathrm{nuc})} = \frac{1}{\min(n, d)} \|\mathbf{D}\|_{nuc}$$

- Document tokens similarity matrix regularization

$$\mathcal{L}_r^{(\mathrm{sim})} = -\frac{1}{n(n-1)} \sum_{\mathbf{d} \in \mathbf{D}} (1 - \|\mathbf{d}\|_2) \sum_{\mathbf{d}' \in \mathbf{D} \setminus \{\mathbf{d}\}} \frac{[\mathbf{d} \cdot \mathbf{d}']_+}{\|\mathbf{d}\|_2 + \varepsilon}$$

- L1 norm

$$\mathcal{L}_r^{(L1)} = \frac{1}{n} \sum_{\mathbf{d} \in \mathbf{D}} \|\mathbf{d}\|_1$$

Final Loss:

$$\mathcal{L}_{final} = \mathcal{L}_{IR} + \alpha \mathcal{L}_r$$

Experiments

- Dataset: MSMARCO for in-domain, BEIR and LoTTE for Out-of-domain
- Reranking based on SPLADE
- Checkpoint selecting: a weighted harmonic mean (effectiveness v.s efficiency)
- Baselines
  - w/o pruning models: SPLADE, ColBERT (end-to-end), ColBERT (rerank)
  - w/ pruning models:
    - Static models: IDF pruning, first-k pruning, stopwords pruning, etc
    - Dynamic models: AligneR, Const-BERT

In-Domain Evaluation Result

model		Remain	Dev set	DL 19	DL 20
w/o Pruning	BM25	-	18.7	50.6	47.5
	SPLADEv2	-	35.8	70.6	68.7
	ColBERTv2 end-to-end	-	39.7	74.5	-
	ColBERTv2 reranking	-	40.0	74.4	75.6
w/ Pruning	First p	50%	37.7	72.3	-
	IDF	50%	32.6	70.2	-
	Stopwords	67%	-	70.9	67.8
	Att. Score	50%	36.0	72.0	-
	AligneR$_{base}$	40%	38.1	-	-
	ConstBERT	47%	39.0	73.1	73.3

model		Remain	Dev set	DL 19	DL 20
LP-based pruning with threshold 0.7					
Ours $\alpha\mathcal{L}_r^{(L1)}$	$\alpha = 0.1$	25%	37.1	72.8†	70.6
	$\alpha = 0.05$	46%	37.9	72.6†	71.4
	$\alpha = 0.01$	83%	39.2	73.6†	72.1
Ours $\alpha\mathcal{L}_r^{(sim)}$	$\alpha = 2$	24%	38.4	72.5†	72.6
	$\alpha = 1.25$	29%	39.3	73.7†	72.8
	$\alpha = 0.8$	**32%**	**39.7**†	**73.8**†	**73.3**
	$\alpha = 0.5$	37%	39.5	73.9†	73.7
	$\alpha = 0.1$	**52%**	**39.9**†	**74.2**†	**73.8**
Our $\alpha\mathcal{L}_r^{(nuc)}$	$\alpha = 0.3$	38%	38.5	71.9†	73.9†
	$\alpha = 0.2$	40%	39.2	73.9†	73.1
	$\alpha = 0.1$	56%	39.4	73.9†	73.2
L1-norm based pruning with threshold 0.5					
Ours $\alpha\mathcal{L}_r^{(L1)}$	$\alpha = 0.1$	9%	37.5	73.0†	70.6
	$\alpha = 0.05$	14%	37.9	72.6†	71.4
	$\alpha = 0.01$	**31%**	**39.2**	**73.6**†	**72.1**
Ours $\alpha\mathcal{L}_r^{(sim)}$	$\alpha = 2$	25%	38.7	72.8†	72.6
	$\alpha = 1.25$	29%	39.4	73.1†	72.9
	$\alpha = 0.8$	33%	39.7†	73.8†	73.2
	$\alpha = 0.5$	37%	39.5	73.9†	73.7
	$\alpha = 0.1$	51%	39.9†	74.2†	73.8
Our $\alpha\mathcal{L}_r^{(nuc)}$	$\alpha = 0.3$	18%	38.5	71.9†	73.9†
	$\alpha = 0.2$	25%	39.2	73.9†	73.1
	$\alpha = 0.1$	34%	39.4	73.9†	73.2

In-Domain Evaluation Result

model		Remain	Dev set	DL 19	DL 20
w/o Pruning	BM25	-	18.7	50.6	47.5
	SPLADEv2	-	35.8	70.6	68.7
	ColBERTv2 end-to-end	-	39.7	74.5	-
	ColBERTv2 reranking	-	40.0	74.4	75.6
w/ Pruning	First p	50%	37.7	72.3	-
	IDF	50%	32.6	70.2	-
	Stopwords	67%	-	70.9	67.8
	Att. Score	50%	36.0	72.0	-
	AligneR$_{base}$	40%	38.1	-	-
	ConstBERT	47%	39.0	73.1	73.3

model		Remain	Dev set	DL 19	DL 20
LP-based pruning with threshold 0.7					
Ours $\alpha\mathcal{L}_r^{(L1)}$	$\alpha = 0.1$	25%	37.1	72.8†	70.6
	$\alpha = 0.05$	46%	37.9	72.6†	71.4
	$\alpha = 0.01$	83%	39.2	73.6†	72.1
Ours $\alpha\mathcal{L}_r^{(sim)}$	$\alpha = 2$	24%	38.4	72.5†	72.6
	$\alpha = 1.25$	29%	39.3	73.7†	72.8
	$\alpha = 0.8$	**32%**	**39.7**†	**73.8**†	**73.3**
	$\alpha = 0.5$	37%	39.5	73.9†	73.7
	$\alpha = 0.1$	**52%**	**39.9**†	**74.2**†	**73.8**
Our $\alpha\mathcal{L}_r^{(nuc)}$	$\alpha = 0.3$	38%	38.5	71.9†	73.9†
	$\alpha = 0.2$	40%	39.2	73.9†	73.1
	$\alpha = 0.1$	56%	39.4	73.9†	73.2
L1-norm based pruning with threshold 0.5					
Ours $\alpha\mathcal{L}_r^{(L1)}$	$\alpha = 0.1$	9%	37.5	73.0†	70.6
	$\alpha = 0.05$	14%	37.9	72.6†	71.4
	$\alpha = 0.01$	**31%**	**39.2**	**73.6**†	**72.1**
Ours $\alpha\mathcal{L}_r^{(sim)}$	$\alpha = 2$	25%	38.7	72.8†	72.6
	$\alpha = 1.25$	29%	39.4	73.1†	72.9
	$\alpha = 0.8$	33%	39.7†	73.8†	73.2
	$\alpha = 0.5$	37%	39.5	73.9†	73.7
	$\alpha = 0.1$	51%	39.9†	74.2†	73.8
Our $\alpha\mathcal{L}_r^{(nuc)}$	$\alpha = 0.3$	18%	38.5	71.9†	73.9†
	$\alpha = 0.2$	25%	39.2	73.9†	73.1
	$\alpha = 0.1$	34%	39.4	73.9†	73.2

In-Domain Evaluation Result

model		Remain	Dev set	DL 19	DL 20
w/o Pruning	BM25	-	18.7	50.6	47.5
	SPLADEv2	-	35.8	70.6	68.7
	ColBERTv2 end-to-end	-	39.7	74.5	-
	ColBERTv2 reranking	-	40.0	74.4	75.6
w/ Pruning	First p	50%	37.7	72.3	-
	IDF	50%	32.6	70.2	-
	Stopwords	67%	-	70.9	67.8
	Att. Score	50%	36.0	72.0	-
	AligneR$_{base}$	40%	38.1	-	-
	ConstBERT	47%	39.0	73.1	73.3

model		Remain	Dev set	DL 19	DL 20
LP-based pruning with threshold 0.7					
Ours $\alpha\mathcal{L}_r^{(L1)}$	$\alpha = 0.1$	25%	37.1	72.8†	70.6
	$\alpha = 0.05$	46%	37.9	72.6†	71.4
	$\alpha = 0.01$	83%	39.2	73.6†	72.1
Ours $\alpha\mathcal{L}_r^{(sim)}$	$\alpha = 2$	24%	38.4	72.5†	72.6
	$\alpha = 1.25$	29%	39.3	73.7†	72.8
	$\alpha = 0.8$	**32%**	**39.7**†	**73.8**†	**73.3**
	$\alpha = 0.5$	37%	39.5	73.9†	73.7
	$\alpha = 0.1$	**52%**	**39.9**†	**74.2**†	**73.8**
Our $\alpha\mathcal{L}_r^{(nuc)}$	$\alpha = 0.3$	38%	38.5	71.9†	73.9†
	$\alpha = 0.2$	40%	39.2	73.9†	73.1
	$\alpha = 0.1$	56%	39.4	73.9†	73.2
L1-norm based pruning with threshold 0.5					
Ours $\alpha\mathcal{L}_r^{(L1)}$	$\alpha = 0.1$	9%	37.5	73.0†	70.6
	$\alpha = 0.05$	14%	37.9	72.6†	71.4
	$\alpha = 0.01$	**31%**	**39.2**	**73.6**†	**72.1**
Ours $\alpha\mathcal{L}_r^{(sim)}$	$\alpha = 2$	25%	38.7	72.8†	72.6
	$\alpha = 1.25$	29%	39.4	73.1†	72.9
	$\alpha = 0.8$	33%	39.7†	73.8†	73.2
	$\alpha = 0.5$	37%	39.5	73.9†	73.7
	$\alpha = 0.1$	51%	39.9†	74.2†	73.8
Our $\alpha\mathcal{L}_r^{(nuc)}$	$\alpha = 0.3$	18%	38.5	71.9†	73.9†
	$\alpha = 0.2$	25%	39.2	73.9†	73.1
	$\alpha = 0.1$	34%	39.4	73.9†	73.2

Out-of-Domain Evaluation Result

	model	Remain	CF	DB	FQ	NF	NQ	QU	SD	SF	TC	TO	Avg.
w/o. Pruning	BM25	-	21.3	31.3	23.6	32.5	32.9	78.9	15.8	66.5	65.6	36.7	40.5
	SPLADEv2	-	21.3	42.0	31.6	32.8	50.8	81.1	14.0	65.9	65.5	25.5	43.0
	ColBERTv2 end-to-end	-	17.6	44.6	35.6	33.8	56.2	85.2	15.4	69.3	73.8	26.3	45.8
	ColBERTv2 reranking	-	20.3	45.9	35.7	34.3	56.8	85.7	14.4	68.5	75.5	33.7	47.0
w/. Pruning	First p	50%	15.5	44.0	32.4	32.4	53.2	78.7	15.2	61.6	72.1	26.1	43.1
	IDF	50%	17.0	45.1	34.0	32.6	55.0	85.3	15.5	64.8	71.4	26.2	44.7
	Att. Score	50%	16.7	44.4	33.6	32.6	54.7	84.4	15.5	64.0	70.0	26.2	44.2
Ours $\alpha\mathcal{L}_r^{(L1)}$	$\alpha = 0.01$	-	18.6	44.3	34.3	$33.8^{\dagger}$	54.0	84.5	$14.2^{\dagger}$	$67.0^{\dagger}$	71.6	31.8	45.4
	Remain %		34%	39%	30%	29%	35%	76%	32%	30%	32%	28%	39%
Ours $\alpha\mathcal{L}_r^{(sim)}$	$\alpha = 0.8$	-	19.6	42.9	$35.0^{\dagger}$	$33.8^{\dagger}$	54.7	82.6	13.9	$66.9^{\dagger}$	$74.3^{\dagger}$	$33.1^{\dagger}$	45.7
	Remain %		36%	34%	36%	38%	36%	40%	36%	38%	37%	40%	37%
	$\alpha = 0.1$	-	$20.2^{\dagger}$	44.7	$35.0^{\dagger}$	$34.4^{\dagger}$	55.5	84.4	14.0	$67.4^{\dagger}$	71.1	$32.8^{\dagger}$	45.9
	Remain %		60%	58%	57%	64%	57%	58%	58%	69%	65%	65%	61%

Out-of-Domain Evaluation Result

model		Remain	CF	DB	FQ	NF	NQ	QU	SD	SF	TC	TO	Avg.
w/o. Pruning	BM25	-	21.3	31.3	23.6	32.5	32.9	78.9	15.8	66.5	65.6	36.7	40.5
	SPLADEv2	-	21.3	42.0	31.6	32.8	50.8	81.1	14.0	65.9	65.5	25.5	43.0
	ColBERTv2 end-to-end	-	17.6	44.6	35.6	33.8	56.2	85.2	15.4	69.3	73.8	26.3	45.8
	ColBERTv2 reranking	-	20.3	45.9	35.7	34.3	56.8	85.7	14.4	68.5	75.5	33.7	47.0
w/. Pruning	First p	50%	15.5	44.0	32.4	32.4	53.2	78.7	15.2	61.6	72.1	26.1	43.1
	IDF	50%	17.0	45.1	34.0	32.6	55.0	85.3	15.5	64.8	71.4	26.2	44.7
	Att. Score	50%	16.7	44.4	33.6	32.6	54.7	84.4	15.5	64.0	70.0	26.2	44.2
Ours $\alpha\mathcal{L}_r^{(L1)}$	$\alpha = 0.01$	-	18.6	44.3	34.3	33.8†	54.0	84.5	14.2†	67.0†	71.6	31.8	45.4
	Remain %		34%	39%	30%	29%	35%	76%	32%	30%	32%	28%	39%
Ours $\alpha\mathcal{L}_r^{(sim)}$	$\alpha = 0.8$	-	19.6	42.9	35.0†	33.8†	54.7	82.6	13.9	66.9†	74.3†	33.1†	45.7
	Remain %		36%	34%	36%	38%	36%	40%	36%	38%	37%	40%	37%
	$\alpha = 0.1$	-	20.2†	44.7	35.0†	34.4†	55.5	84.4	14.0	67.4†	71.1	32.8†	45.9
	Remain %		60%	58%	57%	64%	57%	58%	58%	69%	65%	65%	61%

Out-of-Domain Evaluation Result

model		Remain	CF	DB	FQ	NF	NQ	QU	SD	SF	TC	TO	Avg.
w/o. Pruning	BM25	-	21.3	31.3	23.6	32.5	32.9	78.9	15.8	66.5	65.6	36.7	40.5
	SPLADEv2	-	21.3	42.0	31.6	32.8	50.8	81.1	14.0	65.9	65.5	25.5	43.0
	ColBERTv2 end-to-end	-	17.6	44.6	35.6	33.8	56.2	85.2	15.4	69.3	73.8	26.3	45.8
	ColBERTv2 reranking	-	20.3	45.9	35.7	34.3	56.8	85.7	14.4	68.5	75.5	33.7	47.0
w/. Pruning	First p	50%	15.5	44.0	32.4	32.4	53.2	78.7	15.2	61.6	72.1	26.1	43.1
	IDF	50%	17.0	45.1	34.0	32.6	55.0	85.3	15.5	64.8	71.4	26.2	44.7
	Att. Score	50%	16.7	44.4	33.6	32.6	54.7	84.4	15.5	64.0	70.0	26.2	44.2
Ours $\alpha\mathcal{L}_r^{(L1)}$	$\alpha = 0.01$	-	18.6	44.3	34.3	$33.8^\dagger$	54.0	84.5	$14.2^\dagger$	$67.0^\dagger$	71.6	31.8	45.4
	Remain %		34%	39%	30%	29%	35%	76%	32%	30%	32%	28%	39%
Ours $\alpha\mathcal{L}_r^{(sim)}$	$\alpha = 0.8$	-	19.6	42.9	$35.0^\dagger$	$33.8^\dagger$	54.7	82.6	13.9	$66.9^\dagger$	$74.3^\dagger$	$33.1^\dagger$	45.7
	Remain %		36%	34%	36%	38%	36%	40%	36%	38%	37%	40%	37%
	$\alpha = 0.1$	-	$20.2^\dagger$	44.7	$35.0^\dagger$	$34.4^\dagger$	55.5	84.4	14.0	$67.4^\dagger$	71.1	$32.8^\dagger$	45.9
	Remain %		60%	58%	57%	64%	57%	58%	58%	69%	65%	65%	61%

In domain pruning ratio

31%

32%

52%

Qualitative Analysis

Example passage	Norm pruning: $\theta_N = 0.5$	LP pruning: $\theta_{LP} = 0.7$
Doc-Sim Regularization	What is the **Population** of **New Delhi**. According to **estimated** figures from **Census** of India, **Population** of New Delhi in **2016** is 18.6 million. Delhi's is witnessing a huge **growth** in its population every **year**. **Population** of Delhi **city** is estimated to cross 25 million in **2020**.	What is the **Population** of **New Delhi**. According to **estimated figures** from **Census** of India, **Population** of New Delhi in **2016** is 18.6 million. Delhi's is witnessing a huge **growth** in its population every **year**. Population of Delhi city is estimated to cross 25 million in **2020**.
L1-Norm Regularization	What is the **Population** of **New Delhi**. According to **estimated figures** from **Census** of India, **Population** of New Delhi in **2016** is 18.6 million. Delhi's is witnessing a huge **growth** in its population every **year**. **Population** of Delhi **city** is estimated to cross 25 **million** in **2020**.	What is the **Population** of **New Delhi**. According **to estimated figures** from **Census of India**, **Population of New Delhi in 2016** is **18.6 million**. Delhi's is witnessing a **huge growth in** its **population** every **year**. **Population** of **Delhi city** is **estimated** to **cross** 25 **million** in **2020**.

Remaining Tokens in Bold

Qualitative Analysis

Example passage	Norm pruning: $\theta_N = 0.5$	LP pruning: $\theta_{LP} = 0.7$
Doc-Sim Regularization	What is the **Population** of **New Delhi**. According to **estimated** figures from **Census** of India, **Population** of New Delhi in **2016** is 18.6 million. Delhi's is witnessing a huge **growth** in its population every **year**. **Population** of Delhi **city** is estimated to cross 25 million in **2020**.	What is the **Population** of **New Delhi**. According to **estimated figures** from **Census** of India, **Population** of New Delhi in **2016** is 18.6 million. Delhi's is witnessing a huge **growth** in its population every **year**. Population of Delhi city is estimated to cross 25 million in **2020**.
L1-Norm Regularization	What is the **Population** of **New Delhi**. According to **estimated figures** from **Census** of India, **Population** of New Delhi in **2016** is 18.6 million. Delhi's is witnessing a huge **growth** in its population every **year**. **Population** of Delhi **city** is estimated to cross 25 **million** in **2020**.	What is the **Population** of **New Delhi**. According **to estimated figures** from **Census of India**, **Population of New Delhi in 2016** is **18.6 million**. Delhi's is witnessing a **huge growth in** its **population** every **year**. **Population** of **Delhi city** is **estimated** to **cross** 25 **million** in **2020**.

Remaining Tokens in Bold

Conclusion

- Propose the definition of the **dominance**
- **Adapt** ColBERT score functions
- Propose approximate pruning and regularisation
- Experiment based on experimaestro and XpmIR.

Code for the project

Thank you!

Annexe - Dominance → Linear Programming (LP)

DEFINITION 1 (LOCAL DOMINANCE). *A vector* $\boldsymbol{d}^- \in \mathbb{R}^d$ *is* dominated *by a set of vectors* $\mathbf{D}$*, which we denote by* $\boldsymbol{d}^- \prec \mathbf{D}$*, if and only if for any* $\boldsymbol{q} \in \mathbb{R}^d$,

$$\text{either } \boldsymbol{q} \cdot \boldsymbol{d}^- \leq 0 \text{ or } \exists \boldsymbol{d}^+ \in \mathbf{D} \text{ such that } \boldsymbol{q} \cdot \boldsymbol{d}^+ > \boldsymbol{q} \cdot \boldsymbol{d}^- \quad (3)$$

not dominated

LEMMA 2 (FARKAS' LEMMA). *Let* $\boldsymbol{A} \in \mathbb{R}^{d \times n}$ *and* $\mathbf{b} \in \mathbb{R}^m$*, then exactly one of the following assertion is true (the vector inequalities mean that they must hold for any component of the vector):*

(1) There exists $\boldsymbol{x} \in \mathbb{R}^n$ *such that* $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ *and* $\mathbf{x} \geq 0$.
(2) There exists $\mathbf{y} \in \mathbb{R}^d$ *such that* $\mathbf{A}^\top \mathbf{y} \geq 0$ *and* $\boldsymbol{b}^\top \mathbf{y} < 0$.

Setting $\mathbf{A} = \begin{pmatrix} \mathbf{d} - \mathbf{d}_1 & \dots & \mathbf{d} - \mathbf{d}_n \end{pmatrix}$

$\boldsymbol{b} = -\mathbf{d}$

$$\exists \boldsymbol{x} \in \mathbb{R}^n, \begin{pmatrix} \mathbf{d} - \mathbf{d}_1 & \dots & \mathbf{d} - \mathbf{d}_n \end{pmatrix} \boldsymbol{x} = -\boldsymbol{d} \text{ and } \mathbf{x} \geq 0$$

Final LP formulation

Approximate Pruning

- Dominance ←→ Linear Programming
  - But, some vectors have a very small component that prevent LP to have solutions

$$\mathbf{D} = \mathbf{U} \underbrace{\begin{bmatrix} 1.65 \\ 1.13 \\ 0.97 \\ 0.26 \\ \end{bmatrix}}_{\mathbf{\Sigma}} \underbrace{\begin{bmatrix} v_{11} & v_{12} & v_{13} & v_{14} & v_{15} \\ v_{21} & v_{22} & v_{23} & v_{24} & v_{25} \\ v_{31} & v_{32} & v_{33} & v_{34} & v_{35} \\ v_{41} & v_{42} & v_{43} & v_{44} & v_{45} \\ & & & & \end{bmatrix}}_{\mathbf{V}^*}$$

Apply LP only on truncated V*

- LP is a time consuming problem
  - Norm based pruning

Limitation and Future Work

- From reranking pipeline to end-to-end retrieval
- LP’s heavy computational cost