Pattern Recognition Project
Information Retrieval

Mina GORANOVIC, Marija BRKIC

Université Paris-Saclay, M1 Artificial Intelligence

11/04/2025

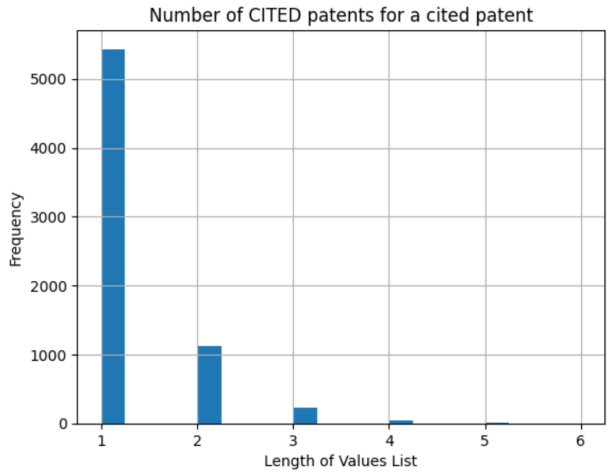
Task1

- **Objective: Finding relevant cited patents for citing patents**
- **Three datasets: citing dataset(train and test), cited dataset and mapping dataset**
- **Patents contain: title, abstract, claim1, claims, description**

	0	1	2	3	4
0	3712070A1	[c-en-0004]	3354576A1	[p0024, p0027, c-en-0012, c-en-0013]	A
1	3675165A1	[c-en-0001, c-en-0002, c-en-0003, c-en-0004, c...	3336831A2	[p0045, p0046, p0047, p0048, p0049, p0050, p00...	A
2	3599626A1	[c-en-0002, c-en-0003, c-en-0004, c-en-0005, c...	2453448A1	[p0029, p0030]	A
3	3705201A1	[c-en-0001, c-en-0002, c-en-0004, c-en-0006, c...	2468433A2	[p0011, p0012, p0013, p0014, p0015, p0016, p00...	X
4	3628210A1	[c-en-0001, c-en-0002, c-en-0003, c-en-0004, c...	3369366A1	[pa01]	A
...	...	...	...	...	...
8589	3623977A1	[c-en-0008, c-en-0009, c-en-0010, c-en-0011, c...	2518981A1	[p0021, p0022, p0023, p0024, p0025, p0026, p00...	A
8590	3721843A1	[c-en-0001, c-en-0002, c-en-0003, c-en-0004, c...	3213727A1	[p0015, p0016, p0017, p0018, p0019, p0020, p00...	X
8591	3708263A1	[c-en-0001, c-en-0002, c-en-0003, c-en-0004, c...	3217171A1	[pa01, p0010, p0014, p0003, p0009, p0016]	A
8592	3588557A1	[c-en-0001, c-en-0002, c-en-0003, c-en-0004, c...	2988328A1	[p0047, p0012]	A
8593	3657819A1	[c-en-0010]	3334179A1	[p0072, p0098]	A

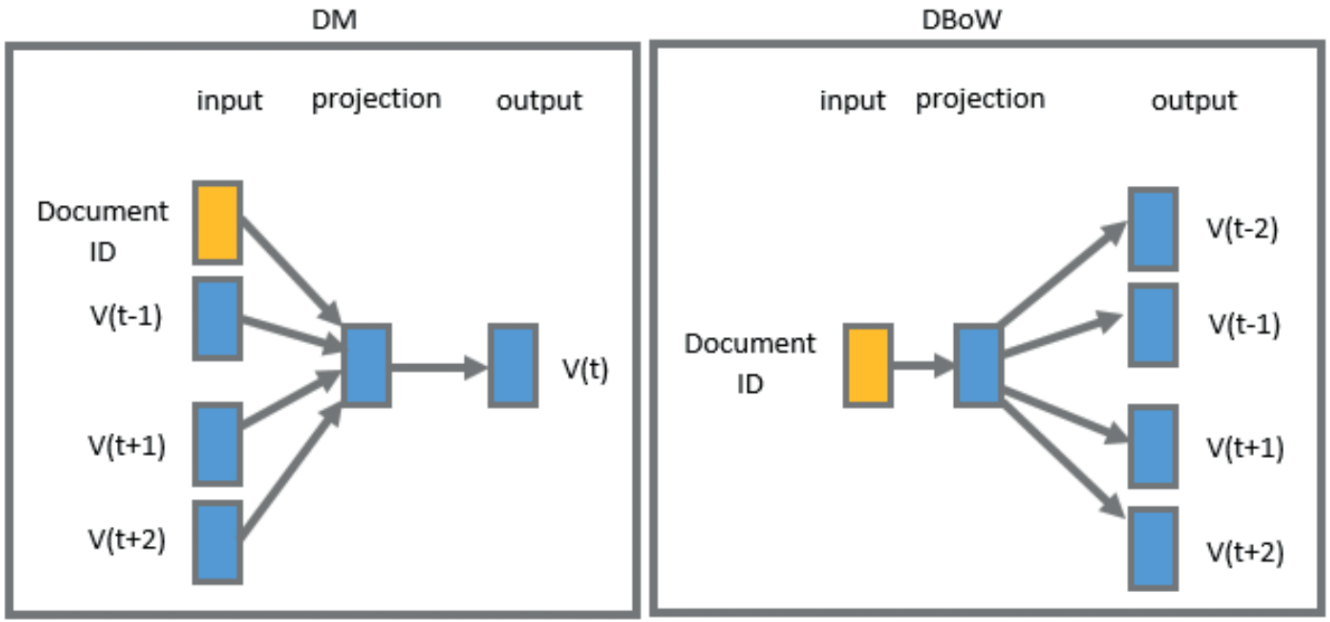
Task1: Evaluation Metrics

- $$\textbf{precission@k} = \frac{\textbf{TP}}{\textbf{K}}$$
- $$\textbf{recall@k} = \frac{\textbf{TP}}{\textbf{P}}$$
- **Mean ranking = how well are relevant documents ranked**



Task1: Doc2Vec

- **Similar to Word2Vec, adds paragraph ID**
- **Distributed Memory and Distributed Bag of Words**



Task1: Doc2Vec

- **claims from citing documents and description for cited documents**

```
Recall at 10: 0.4793
Recall at 20: 0.5679
Recall at 50: 0.6848
Recall at 100: 0.7624
Mean ranking: 36.0658
Mean average precision: 0.2837
Number of patents measured: 6831
Number of patents not in the citation: 0
```

- **claim1 from citing documents and claims for cited documents**

```
Number of documents without claim 1: 0
Number of documents without claim 1: 0
Number of documents without claims: 3
Removing 3 documents without required text
Recall at 10: 0.2439
Recall at 20: 0.3065
Recall at 50: 0.4152
Recall at 100: 0.4979
Mean ranking: 61.7052
Mean average precision: 0.1325
Number of patents measured: 6831
Number of patents not in the citation: 0
```

- **claim1 from citing documents and claim1 for cited documents**

```
Number of documents without claim 1: 0
Number of documents without claim 1: 0
Number of documents without claim 1: 3
Removing 3 documents without required text
Recall at 10: 0.0849
Recall at 20: 0.1206
Recall at 50: 0.1875
Recall at 100: 0.2514
Mean ranking: 82.723
Mean average precision: 0.0419
Number of patents measured: 6831
Number of patents not in the citation: 0
```

Task1: Doc2Vec

'Device (1) for controlling the braking of a trailer, comprising: - at least one control line (2) connectable to a source of a work fluid a
t a first pressure; - a braking line (3) connectable to the service braking system (4) of the trailer and communicating with said control l
ine (2); - at least one additional line (5) connectable to a source of a work fluid at a second pressure; - at least one emergency line (6)
connectable to said additional line (5) and connectable to the emergency and/or parking brake (7) of the trailer of the type of a hydraulic
ally released spring brake; - at least one discharge line (8) of the work fluid communicating with a collection tank (9); - first valve mea
ns operable between a braking position, wherein said additional line (5) is isolated from said discharge line (8), and an emergency positio
n, wherein said additional line (5) is communicating with said discharge line (8);'

'The present invention relates to a device for controlling the braking of a trailer. It is known that in the case of a trailer towed by a p
rime mover their braking systems are operatively connected so that the braking of the prime mover actuated by the operator also causes the
braking of the towed trailer.'

'The present invention relates to device for the towing vehicle-trailer connection. As is known, to date towing vehicles are connected to t
he relative trailer through a connection device comprising a male coupling associated with the towing vehicle and a relative female couplin
g associated with the trailer.'

Task1: BERT Encoder

- **Sentence Transformer BERT 'all-MiniLM-L6-v2' model from hugging face**

- **claims from citing documents and description for cited documents**

```
Number of documents without claims: 0
Number of documents without claims: 0
Number of documents without description: 0
Recall at 10: 0.5214
Recall at 20: 0.6158
Recall at 50: 0.7378
Recall at 100: 0.8104
Mean ranking: 31.2318
Mean average precision: 0.3208
Number of patents measured: 6831
Number of patents not in citation: 0
```

- **claim1 from citing documents and claims for cited documents**

```
Number of documents without claim 1: 0
Number of documents without claim 1: 0
Number of documents without claims: 3
Removing 3 documents without required text
Recall at 10: 0.493
Recall at 20: 0.5847
Recall at 50: 0.6995
Recall at 100: 0.7747
Mean ranking: 34.5613
Mean average precision: 0.3053
Number of patents measured: 6831
Number of patents not in citation: 0
```

- **claim1 from citing documents and claim1 for cited documents**

```
Number of documents without claim 1: 0
Number of documents without claim 1: 0
Number of documents without claim 1: 3
Removing 3 documents without required text
Recall at 10: 0.4763
Recall at 20: 0.5609
Recall at 50: 0.6795
Recall at 100: 0.7568
Mean ranking: 36.5708
Mean average precision: 0.2922
Number of patents measured: 6831
Number of patents not in citation: 0
```

- **Test recall@100 = 0.824, mAP = 0.351**

Task2: Approaches Attempted

- **Cross-Encoder (e5-large-v2):**
  - **Pairwise scoring with full interaction**
  - **Input: Title + Abstract (due to token limit)**
- **Doc2Vec:**
  - **Dense vectors from full patent text**
  - **Reranking based on cosine similarity**
- **Hybrid: Doc2Vec + BM25:**
  - **Combined sparse BM25 rank and dense cosine similarity**
- **Contrastive Learning (Triplet Loss):**
  - **Fine-tuned dense encoder on (query, pos, neg) triplets**
- **Token Selection Strategies:**
  - **Title + Abstract, Claims only, Full text, LLM-derived**

Task2: Limitations Observed

- **Cross-Encoder:**
  - **Truncated input (512 tokens) excluded claims and description**
- **Doc2Vec:**
  - **Semantic matching alone failed to capture legal phrasing**
- **Hybrid: Doc2Vec + BM25:**
  - **Initial =0.5 underperformed; tuning improved MAP**
- **Contrastive Learning (Triplet Loss):**
  - **Weak negatives, overfit to metadata like filing year**
- **Token Selection Strategies:**
  - **No consistent winner across representations**

Q&A

Thank you !

Questions?