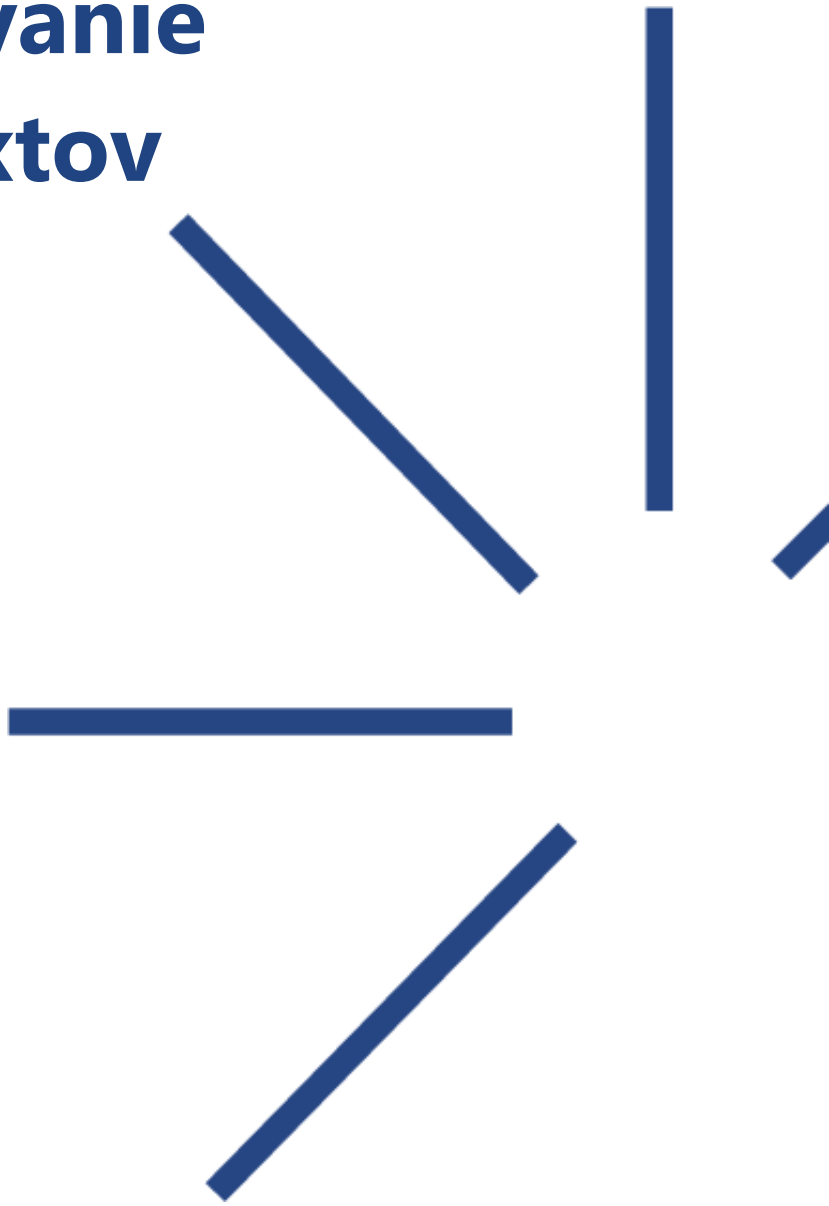
D16 - Dopytovací model pre vyhľadávanie právnych textov



Projekt financovaný Európskou úniou Next GenerationEU prostredníctvom Plánu obnovy a odolnosti SR v rámci projektu č. 09I05-03-V02-00049.



ÚVOD

Projekt „**Automatizácia analýzy právnych textov na základe strojového učenia**“ (ďalej len „**ALTAML**“) predstavuje kľúčovú iniciatívu zameranú na integráciu inovatívnych prístupov v oblasti spracovania údajov a následnej analýzy, so zameraní na oblasť práva, ktorá zahŕňa aj právo informačných a komunikačných technológií. Cieľom tohto projektu je vyvinúť a overiť účinné metódy automatizovanej analýzy právnych textov pomocou techník strojového učenia. To zahŕňa najmä vývoj nástrojov, ktoré uľahčujú spracovanie a analýzu veľkých objemov údajov vo forme rôznych právnych dokumentov, ako aj extrakciu relevantných informácií (atribútov) z takýchto dokumentov vrátane identifikácie kľúčových pojmov, odkazov na právne predpisy a ďalších atribútov.

Cieľom projektu ALTAML je teda prispieť k efektívnejšiemu prístupu k právnym informáciám a k zrýchleniu právnych procesov, čím sa zabezpečí vyššia miera právnej istoty a zlepší sa dostupnosť právnych textov, výsledkov ich analýzy a relevantných právnych informácií pre odbornú aj širokú verejnosť.

V rámci pracovného balíka KPB3 bol vyvinutý **D16 – Model dotazovania na filtrovanie právneho textu** využívajúci kombináciu extrahovaných atribútov, fulltextového a sémantického vyhľadávania. Cieľom tohto výstupu je definovať spôsob vyhľadávania nad právnymi textami, najmä nad súdnymi rozhodnutiami a právnymi predpismi, pričom navrhnutý model je možné rozšíriť aj o ďalšie typy právnych dokumentov. Model umožňuje kombinovať:

- fulltextové vyhľadávanie v textoch súdnych rozhodnutí a právnych predpisov,
- vyhľadávanie podľa extrahovaných odkazov na právne predpisy,
- vyhľadávanie podľa extrahovaných odkazov na iné súdne rozhodnutia,
- vyhľadávanie podľa automaticky identifikovaných kľúčových pojmov,
- filtrovanie podľa kategorizovaných atribútov (napr. oblasť práva, podoblasť práva, súd alebo dátum rozhodnutia),
- sémantické vyhľadávanie pomocou vektorových reprezentácií súdnych rozhodnutí a právnych predpisov,
- hybridné vyhľadávanie kombinujúce fulltextové, atribútové a sémantické vyhľadávanie.

Navrhnutý dopytovací model vychádza z využitia Elasticsearch Query DSL a nadväzuje na výstup **D15 – Návrh repozitára**, v ktorom bol Elasticsearch definovaný ako finálna vyhľadávacia vrstva systému ALTAML.

Nad týmto dopytovacím modelom je následne možné budovať vyššie vrstvy používateľskej interakcie a inteligentnej asistencie pri formulovaní dopytov. Táto funkcionalita je riešená samostatne v rámci nadväzujúceho výstupu projektu **D17 – Sprostredkovacia vrstva vyhľadávania**, ktorého úlohou je transformovať

používateľské požiadavky formulované v prirodzenom jazyku do dopytov nad právnymi dokumentmi.

1. Konfigurácia indexov Elasticsearch

V prostredí Elasticsearch predstavuje index logické úložisko dokumentov určené na vyhľadávanie a filtrovanie dát, ktoré je možné prirovnať k tabuľke v relačných SQL databázach. Každý index obsahuje dokumenty rovnakého typu spolu s definíciou ich štruktúry (*mapping*) a spôsobu spracovania textových polí (*settings* a analyzéry). Konfigurácia indexu určuje dátové typy jednotlivých polí, spôsob spracovania textu a podporované typy vyhľadávania.

Navrhnutý dopytovací model využíva samostatné Elasticsearch indexy pre jednotlivé typy právnych dokumentov. V nasledujúcich podkapitolách je prezentovaná ich konfigurácia.

1.1. Konfigurácia indexu súdnych rozhodnutí

Index súdnych rozhodnutí obsahuje jednotlivé súdne rozhodnutia reprezentované ako samostatné dokumenty. Dokumenty obsahujú fulltextové polia, štruktúrované atribúty a vektorové reprezentácie využívané pri rôznych typoch vyhľadávania. Konfigurácia indexu pozostáva z častí *settings* a *mappings*.

Časť *settings* definuje spôsob spracovania textu pri indexovaní a vyhľadávaní. Dôležitou súčasťou tejto konfigurácie sú analyzéry, ktoré určujú spôsob tokenizácie textu a aplikácie textových transformácií, napríklad prevod na malé písmená alebo lematizáciu.

Nasledujúca ukážka Ukážka 1 prezentuje konfiguráciu časti *settings* použitej pri indexe súdnych rozhodnutí:

```
"settings": {
  "analysis": {
    "tokenizer": {
      "standard": { "type": "standard" }
    },
    "filter": {
      "lowercase": { "type": "lowercase" },
      "hunspell_sk": {
        "type": "hunspell",
        "locale": "sk",
        "dictionary": "/usr/share/elasticsearch/config/hunspell/sk"
      }
    },
    "analyzer": {
      "slovak_lemmatizer": {
        "type": "custom",
        "tokenizer": "standard",
```

```
        "filter": ["lowercase", "hunspell_sk"]
      }
     }
    },
    "number_of_shards": 1,
    "number_of_replicas": 1
  }
```

Ukážka 1 - Konfigurácia časti settings indexu súdnych rozhodnutí

Pre fulltextové polia je použitý vlastný analyzér *slovak_lemmatizer* založený na slovenskom **Hunspell** slovníku. Použitie Hunspell analyzéra umožňuje dosiahnuť zhodu aj medzi rovnakými slovami v rôznych gramatických tvaroch, čo je dôležité pri vyhľadávaní v slovenských textoch.

Časť *mappings* (Ukážka 2) definuje jednotlivé polia dokumentov, ich dátové typy a spôsob ich využitia pri vyhľadávaní. Dátové typy určujú, či bude pole využívané na fulltextové vyhľadávanie, filtrovanie, dátumové dopyty alebo sémantické vyhľadávanie pomocou vektorových reprezentácií.

```
  "mappings": {
    "properties": {
      "document_pretty": {
        "type": "text",
        "analyzer": "slovak_lemmatizer"
      },
      "datum_vydania_rozhodnutia": {
        "type": "date",
        "format": "yyyy-MM-dd"
      },
      "oblast_pravnej_upravy": {
        "type": "keyword"
      },
      "text_chunks": {
        "type": "nested",
        "properties": {
          "part": {
            "type": "keyword"
          },
          "text": {
            "type": "text",
            "analyzer": "slovak_lemmatizer"
          },
```

```
        "embedding": {
          "type": "dense_vector",
          "dims": 384,
          "index": true,
          "similarity": "cosine"
        }
      },
      //ďalšie atribúty...
      }
    }
  }
```

Ukážka 2 - Konfigurácia časti mappings indexu súdnych rozhodnutí

V konfigurácii indexu sú jednotlivým poliam priradené dátové typy podľa spôsobu ich využitia pri vyhľadávaní (Tabuľka 1). Typ **text** je určený pre fulltextové polia analyzované pri vyhľadávaní, typ **keyword** pre presné filtrovanie a agregácie a typ **date** pre dátumové polia podporujúce rozsahové dopyty. Typ **nested** sa používa pre vnorené štruktúrované objekty, napríklad odkazy na právne predpisy alebo iné súdne rozhodnutia. Pre sémantické vyhľadávanie sú využité polia typu **dense_vector**, ktoré obsahujú vektorové reprezentácie právnych dokumentov.

Tabuľka 1 - Atribúty indexu súdnych rozhodnutí, ich dátové typy a účel pri vyhľadávaní a filtrovaní

Atribút	**Typ**	**Účel**
document_pretty	text	fulltextové vyhľadávanie nad textom rozhodnutia
datum_vydania_rozhodnutia	date	filtrovanie podľa dátumu
ecli	keyword	presné filtrovanie podľa ECLI identifikátora rozhodnutia
spisova_znacka	keyword	presné filtrovanie podľa spisovej značky
forma_rozhodnutia	keyword	filtrovanie podľa formy rozhodnutia
oblast_pravnej_upravy	keyword	filtrovanie podľa právnej oblasti
podoblast_pravnej_upravy	keyword	filtrovanie podľa právnej podoblasti
laws	nested	filtrovanie podľa identifikátorov odkazovaných právnych predpisov
dockets	nested	filtrovanie podľa odkazovaných rozhodnutí
key_phrases	keyword	filtrovanie podľa extrahovaných kľúčových pojmov
sud_nazov	keyword	filtrovanie podľa názvu súdu
sud_typ	keyword	filtrovanie podľa typu súdu

sudca_meno	text	fulltextové vyhľadávanie podľa mena sudcu
text_chunks	nested	textové časti rozhodnutia využívané pri sémantickom vyhľadávaní
text_chunks.part	keyword	filtrovanie podľa časti rozhodnutia
text_chunks.text	text	fulltextové vyhľadávanie v konkrétnych častiach rozhodnutia
text_chunks.embedding	dense_vector	sémantické vyhľadávanie podľa významovej podobnosti k častiam rozhodnutí

Keďže súdne rozhodnutia môžu byť rozsiahle, vektorová reprezentácia sa nevytvára iba pre celé rozhodnutie, ale pre menšie textové časti (chunky). Každý chunk obsahuje samotný text, informáciu o časti rozhodnutia, z ktorej pochádza (napr. hlavička, úvod, výrok, odôvodnenie alebo poučenie), a jeho vektorovú reprezentáciu, čo umožňuje sémantické vyhľadávanie podobných častí rozhodnutí.

Pri sémantickom dopyte sa používateľský výraz alebo otázka najprv prevedie rovnakým jazykovým modelom do vektorovej reprezentácie a následne sa porovnáva s vektormi jednotlivých chunkov uložených v Elasticsearch. Takýmto spôsobom je možné vyhľadávať významovo podobné rozhodnutia alebo ich časti aj v prípadoch, keď používateľský dopyt neobsahuje rovnaké slová ako samotný dokument. Zároveň je možné obmedziť vyhľadávanie iba na konkrétne časti rozhodnutí, napríklad iba na odôvodnenie alebo výrok rozhodnutia. Ukážka takéhoto sémantického dopytu je prezentovaná v podkapitole 2.1.

1.2. Konfigurácia indexu právnych predpisov

Keďže právne predpisy majú odlišnú štruktúru ako súdne rozhodnutia, sú uložené v samostatnom Elasticsearch indexe.

Konfigurácia časti *settings* vrátane analyzéra *slovak_lemmatizer* je pri indexe právnych predpisov riešená obdobným spôsobom ako pri indexe súdnych rozhodnutí. Z dôvodu prehľadnosti už nie je znovu uvedená ukážka konfigurácie *mappings* a jednotlivé atribúty indexu sú priamo zhrnuté v nasledujúcej tabuľke Tabuľka 2.

Tabuľka 2 - Atribúty indexu právnych predpisov, ich dátové typy a účel pri vyhľadávaní a filtrovaní

Atribút	**Typ**	**Účel**
law_id	keyword	presné filtrovanie podľa identifikátora právneho predpisu
law_titles	text	fulltextové vyhľadávanie podľa hierarchických nadpisov právneho predpisu

valid_from	date	filtrovanie podľa začiatku platnosti
valid_to	date	filtrovanie podľa konca platnosti
forma_rozhodnutia	keyword	filtrovanie podľa formy rozhodnutia
text	text	fulltextové vyhľadávanie v texte právneho predpisu
key_phrases	keyword	filtrovanie podľa extrahovaných kľúčových pojmov
embedding	dense_vector	sémantické vyhľadávanie podľa významovej podobnosti k právnym predpisom

Na rozdiel od súdnych rozhodnutí sa pri právnych predpisoch nepoužíva delenie textu na chunky. Sémantické vyhľadávanie je realizované pomocou jednej vektorovej reprezentácie vytvorenej pre celý právny predpis a uloženej v poli embedding.

2. Typy podporovaných dopytov

Navrhnutý dopytovací model umožňuje realizovať rôzne typy vyhľadávania nad súdnymi rozhodnutiami a právnymi predpismi a kombinovať ich v rámci jedného Elasticsearch DSL dopytu. Keďže jednotlivé typy dokumentov sú uložené v samostatných Elasticsearch indexoch, dopyty sú vykonávané nad príslušným indexom, napríklad prostredníctvom volania GET /court_decisions/_search alebo GET /laws/_search

V nasledujúcich podkapitolách sú prezentované ukážky vybraných typov dopytov.

2.1. Ukážky dopytov nad súdnymi rozhodnutiami

V nasledujúcich ukážkach sú prezentované vybrané typy dopytov demonštrujúce možnosti navrhnutého dopytovacieho modelu nad súdnymi rozhodnutiami.

Dopyt uvedený v Ukážka 3 vyhľadá súdne rozhodnutia, ktoré buď obsahujú pojem „nekalá súťaž“, alebo obsahujú extrahovaný odkaz na § 49 zákona č. 513/1991 Zb. Obchodný zákonník, ktorým je definované podplácanie.

Dopyt využíva podmienku *should*, ktorá reprezentuje logickú operáciu OR medzi fulltextovým vyhľadávaním a filtrovaním podľa extrahovaných právnych predpisov. Parameter *minimum_should_match* určuje minimálny počet podmienok v časti *should*, ktoré musia byť splnené. Hodnota 1 znamená, že dokument musí spĺňať aspoň jednu z uvedených podmienok.

```
{
  "query": {
    "bool": {
      "should": [
        {
          "match": {
            "document_pretty": "nekalá súťaž"
          }
        },
        {
          "nested": {
            "path": "laws",
            "query": {
              "bool": {
                "must": [
                  {
                    "term": {
                      "laws.law": "513/1991"
                    }
                  },
                  {
                    "term": {
                      "laws.paragraph": "49"
                    }
                  }
                ]
              }
            }
          }
        }
      ],
      "minimum_should_match": 1
    }
  }
}
```

Ukážka 3 - Elasticsearch dopyt vyhľadávajúci rozhodnutia podľa pojmu „nekalá súťaž“ alebo § 49 Obchodného zákonníka

Ukážka 4 prezentuje dopyt vyhľadávajúci súdne rozhodnutia, ktorým bol automaticky priradený aspoň jeden z kľúčových pojmov „klamlivá reklama“ alebo „obchodné tajomstvo“. Operátor *terms* umožňuje vyhľadať dokumenty obsahujúce aspoň jednu z uvedených hodnôt. Na rozdiel od fulltextového vyhľadávania sa v tomto prípade nevyhľadáva priamo v texte rozhodnutia, ale v štruktúrovanom atribúte (key_phrases) obsahujúcom extrahované pojmy.

```
{
  "query": {
    "terms": {
      "key_phrases": [
        "klamlivá reklama",
        "obchodné tajomstvo"
      ]
    }
  }
}
```

```
}
```

Ukážka 4 - Elasticsearch dopyt vyhľadávajúci rozhodnutia podľa kľúčových pojmov „klamlivá reklama“ a „obchodné tajomstvo“

Ukážka 5 prezentuje sémantický dopyt vyhľadávajúci súdne rozhodnutia obsahujúce **významovo podobné** pasáže k pojmu „nekalá súťaž“, pričom vyhľadávanie je obmedzené iba na časť rozhodnutia predstavujúcu odôvodnenie (ang. reasoning).

```
{
  "size": 10,
  "query": {
    "nested": {
      "path": "text_chunks",
      "query": {
        "bool": {
          "filter": [
            {
              "term": {
                "text_chunks.part": "reasoning"
              }
            }
          ],
          "must": [
            {
              "knn": {
                "field": "text_chunks.embedding",
                "query_vector":[0.12,...,-0.03,0.44], /*384-rozmerný vektor */
                "k": 30,
                "num_candidates": 300
              }
            }
          ]
        }
      },
      "inner_hits": {
        "size": 3
      }
    }
  }
}
```

Ukážka 5 - Elasticsearch knn dopyt pre sémantické vyhľadávanie v časti odôvodnenia súdnych rozhodnutí

Pri sémantickom vyhľadávaní je používateľský dopyt, napríklad pojem „nekalá súťaž“, najprv prevedený rovnakým jazykovým modelom, aký bol použitý pri vytváraní poľa text_chunks.embedding, do vektorovej reprezentácie (*query_vector*). Ukážka 6 demonštruje vytvorenie vektorovej reprezentácie používateľského dopytu v jazyku Python pomocou knižnice sentence-transformers.

```
from sentence_transformers import SentenceTransformer

model = SentenceTransformer("sentence-transformers/paraphrase-multilingual-
```

```
MiniLM-L12-v2")
query_vector = model.encode("nekalá súťaž").tolist()
```

Ukážka 6 - Vytvorenie embedding vektora používateľského dopytu pomocou knižnice sentence-transformers

Dimenzionalita vektora závisí od použitého jazykového modelu. V uvedenom príklade model *paraphrase-multilingual-MiniLM-L12-v2* vytvára 384-rozmerné vektorové reprezentácie. Použitý jazykový model je možné v budúcnosti nahradiť iným modelom, avšak v takom prípade je potrebné znovu vytvoriť vektorové reprezentácie všetkých indexovaných dokumentov, aby boli embeddingy dokumentov a používateľských dopytov vytvorené v rovnakom vektorovom priestore.

Vytvorená vektorová reprezentácia sa následne porovnáva kosínovou podobnosťou s vektormi jednotlivých textových častí rozhodnutí uložených v Elasticsearch. Takýto prístup umožňuje vyhľadávať významovo podobné pasáže aj v prípadoch, keď sa v dokumente nenachádzajú rovnaké slová ako v používateľskom dopyte.

Použitie atribútu text_chunks.part zároveň umožňuje obmedziť sémantické vyhľadávanie iba na konkrétnu časť rozhodnutia, napríklad na odôvodnenie, výrok alebo poučenie.

Parameter *size* určuje maximálny počet súdnych rozhodnutí vrátených používateľovi.

Elasticsearch realizuje *knn* dopyty pomocou približného vyhľadávania najbližších susedov (ANN – Approximate Nearest Neighbor). Parameter *num_candidates* určuje počet takýchto kandidátnych textových chunkov zaradených do užšieho výberu, z ktorého sa následne vyberie *k* najpodobnejších výsledkov. Vyššia hodnota *num_candidates* môže zvýšiť presnosť výsledkov, ale zároveň zvyšuje výpočtovú náročnosť vyhľadávania.

Atribút *inner_hits* určuje počet textových častí rozhodnutia s najvyšším skóre, ktoré budú spolu s výsledným rozhodnutím vrátené používateľovi. V uvedenom príklade bude vrátených maximálne 10 rozhodnutí a pre každé rozhodnutie maximálne 3 najrelevantnejšie textové chunky.

Okrem samostatného fulltextového a sémantického vyhľadávania je možné oba prístupy kombinovať aj v rámci jedného **hybridného dopytu**. Takýto prístup umožňuje zohľadniť nielen presnú textovú zhodu s používateľským dopytom, ale aj významovú podobnosť medzi dokumentmi a používateľským dopytom.

Ukážka 7 prezentuje hybridné vyhľadávanie kombinujúce klasické fulltextové skóre BM25 so sémantickou podobnosťou založenou na vektorových reprezentáciách. Výsledné skóre dokumentu vzniká váženou kombináciou oboch prístupov.

```
{
  "query": {
    "script_score": {
      "query": {
        "match": {
```

```
                    "document_pretty": "nekalá súťaž"
                  }
                },
                "script": {
                  "source": """
                    0.7 * _score +
                    0.3 * cosineSimilarity(
                      params.query_vector,
                      'embedding'
                    )
                  """,
                  "params": {
                    "query_vector": [0.12, -0.03, 0.44]
                  }
                }
              }
            }
          }
```

Ukážka 7 - Hybridné vyhľadávanie kombinujúce fulltextové skóre BM25 a kosínovú podobnosť embeddingov

V uvedenom príklade predstavuje hodnota *_score* skóre fulltextového vyhľadávania BM25 a funkcia *cosineSimilarity* vyjadruje významovú podobnosť medzi používateľským dopytom a vektorovou reprezentáciou dokumentu. Váhy 0.7 a 0.3 určujú relatívny vplyv fulltextového a sémantického vyhľadávania na výsledné skóre dokumentu. Vyššia váha priradená fulltextovému skóre zvýhodňuje dokumenty obsahujúce presnú textovú zhodu s používateľským dopytom, zatiaľ čo vyššia váha sémantickej podobnosti zvýhodňuje dokumenty významovo podobné používateľskému dopytu aj bez výskytu rovnakých slov.

2.2. Ukážky dopytov nad právnymi predpismi

Navrhnutý dopytovací model podporuje aj vyhľadávanie nad právnymi predpismi uloženými v samostatnom Elasticsearch indexe. Pri vyhľadávaní je možné kombinovať fulltextové vyhľadávanie v texte právnych predpisov, filtrovanie podľa obdobia platnosti, vyhľadávanie podľa extrahovaných pojmov aj sémantické vyhľadávanie pomocou vektorových reprezentácií právnych predpisov.

Ukážka 8 prezentuje dopyt vyhľadávajúci právne predpisy obsahujúce pojem „nekalá súťaž“, ktoré boli platné k 1.1.2026.

```
{
  "query": {
    "bool": {
      "must": [
        {
          "match": {
            "text": "nekalá súťaž"
          }
        }
      ],
      "filter": [
```

```
                    {
                      "range": {
                        "valid_from": {
                          "lte": "2026-01-01"
                        }
                      }
                    },
                    {
                      "range": {
                        "valid_to": {
                          "gte": "2026-01-01"
                        }
                      }
                    }
                  ]
                }
              }
            }
```

Ukážka 8 - Elasticsearch dopyt kombinujúci fulltextové vyhľadávanie a filtrovanie podľa obdobia platnosti právneho predpisu

Sémantické vyhľadávanie pomocou vektorových reprezentácií a vyhľadávanie podľa extrahovaných kľúčových pojmov je pri právnych predpisoch realizované obdobným spôsobom ako pri súdnych rozhodnutiach. Rozdiel spočíva najmä v odlišnej štruktúre indexovaných dokumentov a v použití jednej vektorovej reprezentácie pre celý právny predpis.