Links: final do doc

Workshop de processamento de textos legislativos

↳ Ulysses: plataforma de processamento de textos legislativos
↳ DIVERSOS ALGS

↳ TUPI: abrangência, volume alto

Ulysses Tesemõ: corpus legal brasileiro
↳ Em edição
↳ Legislação federal
Notícias da câmara
Notícias de Ministérios
Senado Federal
Legislação Estadual

Ulysses Segmenter: segmentador de textos em Python
↳ Bloco construtor
- map 2 doc
- Ulysses Amendment Comparer
- Tratamento do Ulysses

PRÉ-PROCESSAMENTO
- remover quebra de linha, espaços em branco agrupados
- Tokenização

ML
- janelas móveis
} Instâncias separadas

PÓS PROCESSAMENTO

- Bert / Bi-LSTM
- Byte Pair Encoding
- 4 classes: - No operation
  - Ruído
  - S/ ruído

Retorna as sentenças

- Pooling
- Token wise prediction
- User output

- **Weak supervision** (aprende a copiar/implementar regra): modelo treinado em 100.000 + docs legs rotulados por meio de 30 regs
- **Active learning** (corrigir falsos negs e falsos positivos): Fine tuning em 4.000+ instâncias selecionadas de acordo com uma métrica de dúvida do modelo

Ulysses Document Comparer: pipeline p/ recuperação de documentos legislativos

Retorna proposições legislativas (PL, PEC) ou outras consultas

↓ consulta

Apartir de uma query recuperar documentos similares ou relevantes àquela busca

↳ 4 microsserviços:

- Look 4 similar (BM25)
  ↳ Calcula score p/ cada documento
  - Pré-processamento
    - lowercase
    - remover stopwords, pontuação, acentuação

- Save Relevance Feedback: usuário pode fornecer feedback sobre o retorno de documentos em 3 níveis
  - Relevante, irrelevante, pouco relevante
  - Score normalizado

Precisão ≠ p/ documentos ≠

- Improve Similarity: utilizar o feedback p/ melhorar recuperações futuras
  - ↑ o score de documentos relevantes e penaliza documentos irrelevantes

* Ulysses-RF corpus: utilizado p/ avaliar o sistema

Ulysses NER-BR Expand Query

- Reconhecimento de entidades nomeadas
- Retorna uma expansão explicação da query

Ulysses Amendment Comparer

- Otimizar o trabalho
- 2 abordagens: - Modelagem de tópicos
  - Inferência de tópicos pelo modelo
  - Agrupamento

- Recuperação de informação
  - Tópicos pre-selecionados pelo usuário
  - Retorna tópicos agrupados por emendas

- Cluster Comments e Ulysses Segmenter

- Agrupa as emendas de PECs em tópicos (modelo BERTopic)
- Corpus com 269 (pouco e pequeno) emendas anotadas em tópicos por um consultor

Ulysses Sentence Models
  - Bloco construtor

- Vetor de sentenças

- Treinamento supervisionado

Map 2 Doc : mapeamento de opinião pública em documentos

Links:

- https://github.com/Convenio-Camara-dos-Deputados
- https://github.com/ulysses-camara
- https://www.camara.leg.br/noticias/548730-camara-lanca-ulysses-robo-digital-que-articula-dados-legislativos/
- https://drive.google.com/file/d/1w3iCuoHOX49msq-Js8wRx13MX2dxlzxv/view?usp=sharing
https://drive.google.com/file/d/

- 1Dm082C2FtXgTbu7r_IsbrEhqB8PYUZ6z/view?usp=sharing