

Workshop de processamento de textos legislativos

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

↳ Tupi: abrangência, volume alto

↳ Ulysses Tesmão: corpus legal brasileiro

↳ Em inclusão

↳ 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
- Treinamento do Ulysses

PRÉ-PROCESSAMENTO

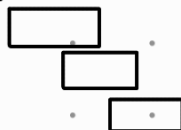
ML

PÓS-PROCESSAMENTO

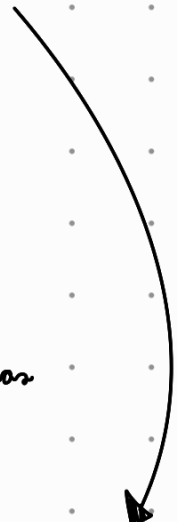
remover quebra
de linha, espaços
em branco agrupados

Tokenização

- janelas móveis



} Instâncias separadas



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

Retorna as sentenças

- Pooling
- Token wise prediction
- User output

- **Weak supervision**: modelo treinado em 100.000 + docs legal rotulados aprendendo a copiar/implementar regras

- **Active learning**: Fine tuning em 4.000 + instâncias selecionadas de acordo com uma métrica de dano do modelo
↳ corrigir falsos negs e falsos positivos

⇒ **Ulysses Document Comparer**: pipeline p/ recuperação de documentos legislativos

Retorna proposições legislativas (PL, PEC) ou outras consultas

↓ consulta
Após de uma query recupera documentos similares ou relevantes àquela busca

↳ 4 micros serviços:

- Look 4 similar (BM25)
 - ↳ Calcula score p/ cada documento
- Pré-processamento
 - lower case
 - remove stopwords, punctuation, accentuation

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

Precisão \neq
p/ documentos \neq

Score normalizado

- 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 informações
 - ↳ Tópicos pre-selecionados pelo usuário
 - ↳ Retorna tópicos agrupados por emendas

- Cluster Comments e Ulysses Segmentor

- Agrupa as emendas de PECs em tópicos
- Copiar com 269 ^{pequeno e único} emendas anotados em tópicos por um consultor

Modelo BERTopic

➡ Ulysses Sentence Models

↳ Bloco construtor

- Vetor de sentenças

- Treinamento supervisionado

➡ Map 2 Doc : mapeamento de opiniões 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

