Machine Learning - Tech Tendencies

Gabriel Cassol Bach, Bruno Emanuel Zenatti

Contexto Geral

- Dataset de solicitações de patentes feitas no EUA entre 2011 e 2016 (~4.5 milhões de pedidos de patentes)

- Perceber tendências e padrões interessantes nas novas tecnologias.

Foram utilizadas duas abordagens de Machine Learning:

- Primeira abordagem: SentenceTransformers + Clusterização: perceber novas tendências de tecnologia.

- Segunda abordagem: RandomForest: estimar se uma patente vai ser aprovada ou rejeitada.

Ideia da Primeira Abordagem

- Utilizar o título e o resumo de cada patente para gerar vetores com alta capacidade de representação dos documentos (SentenceTransformers).

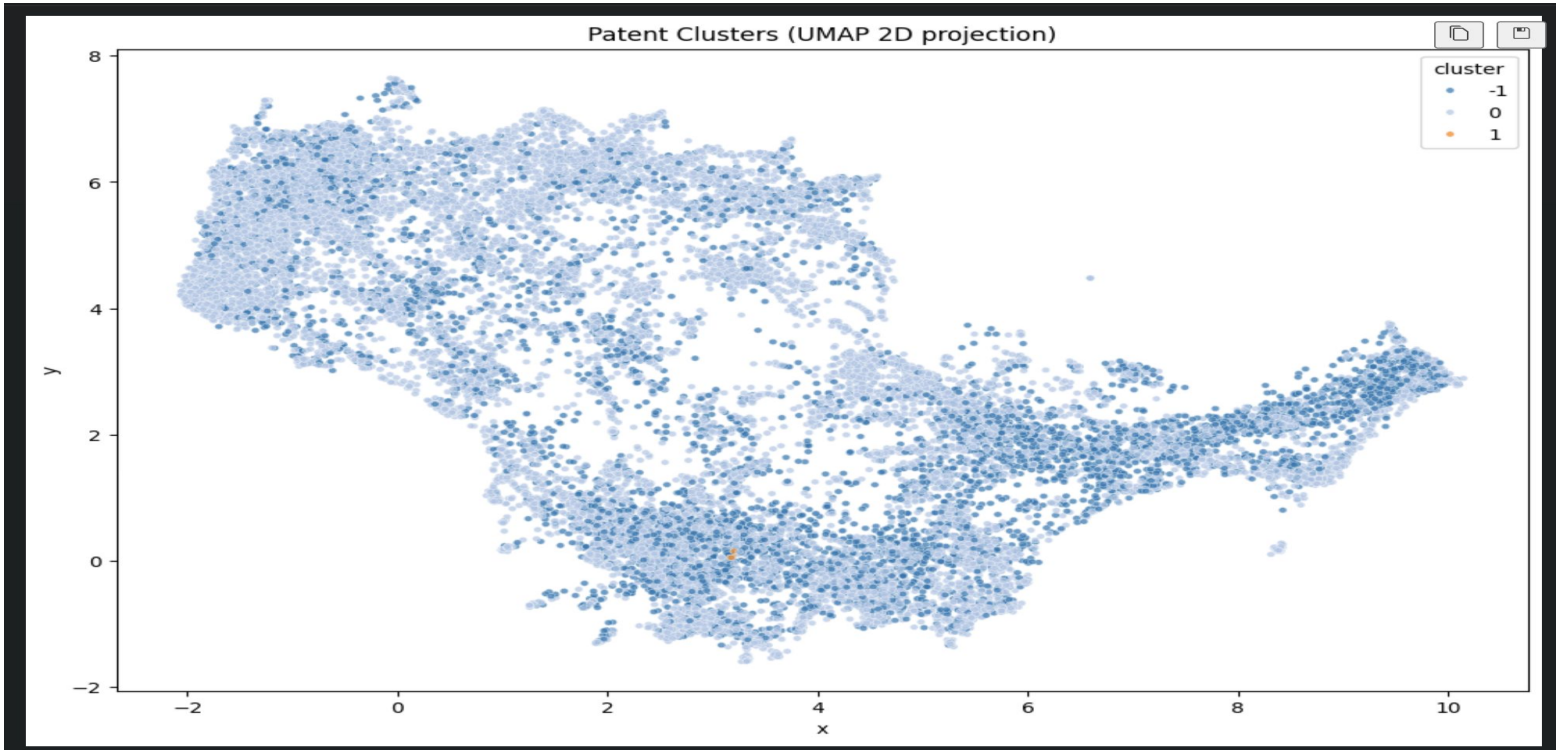
- A partir dos vetores, aplicar um algoritmo de clusterização (HDBSCAN) a fim de notar outliers.

Resultados da Primeira Abordagem

- Os tempos de execução dos algoritmos para uma entrada com cerca de 28 mil documentos de patentes foram demorados e geraram diversos problemas de implementação.

- O algoritmo gerou poucos clusters e, no geral, não conseguiu separar as patentes em diversos grupos distintos. Além disso, o treinamento precisaria ser feito com um conjunto de dados maior.

Resultados da Primeira Abordagem



Considerações da Primeira Abordagem

- Não gerou bons resultados.

- Precisa ser avaliada e discutida com o professor.

Ideia da Segunda Abordagem

A fim de criar um modelo que estima se uma patente vai ser aceita ou rejeitada, foram feitas as seguintes etapas de desenvolvimento:

- Limpeza nos dados;

- Utilização dos campos ‘título’ e ‘resumo’ das patentes

- Transformação dos textos em vetores numéricos por meio da técnica: Term Frequency-Inverse Document Frequency

- Separação dos dados para treino/teste (80/20)

- Aplicação do algoritmo RandomForest para a criação de várias árvores de decisão independentes (patente aceita ou rejeitada)

Resultados da Segunda Abordagem

Classe	Precisão	Recall	F1-Score	Suporte
REJECTED	0.28	0.50	0.36	4865
ACCEPTED	0.87	0.73	0.80	23096
Acurácia			0.69	27961
Média Macro	0.58	0.61	0.58	27961
Média Ponderada	0.77	0.69	0.72	27961

Resultados da Segunda Abordagem

Considerações da Segunda Abordagem

- Principal desafio: desbalanceamento das classes.
- Métodos testados não resolveram completamente o problema.
- Origem do problema incerta: pode estar na vetorização dos textos ou no modelo.
- Possível integração de um algoritmo de Doc2Vec

Obrigado!