

## PROJETO EM CIÊNCIA DE DADOS - BUSINESS CASE

SEMESTRE	2025/1
ORGANIZAÇÃO PARCEIRA	WEG Tintas Ltda e Vent Digital Ltda
CONTATO PRINCIPAL	Victor Yuudi Suzuki ( <a href="mailto:suzuki@weg.net">suzuki@weg.net</a> ) e Thomaz Borela ( <a href="mailto:thomaz@vent.digital">thomaz@vent.digital</a> )

### Informações sucintas sobre a organização e seus principais produtos e serviços

A WEG Tintas é a unidade de negócios do Grupo WEG, multinacional brasileira reconhecida pelos seus motores elétricos, que atua no mercado de tintas industriais para a proteção de ativos da degradação por corrosão.

A WEG Tintas é líder no mercado nacional de tintas em pó e vernizes isolantes elétricos, sendo também responsável pela produção de tintas líquidas, fornecendo produtos para a indústria, marinha/offshore, plásticos, repintura automóvel, eletrodomésticos e estruturas metálicas.

### Título do Projeto

Prospecção tecnológica: identificação e classificação de informações químicas em patentes

### Justificativa do Projeto

Explicitar a justificativa do projeto do ponto de vista de negócio.  
Que valor o projeto poderá agregar à organização/aos clientes?

A empresa enfrenta desafios na área de pesquisa e desenvolvimento (P&D) devido à dificuldade e ao tempo excessivo despendido na busca e análise de informações relevantes em patentes, em virtude do volume de novos dados gerados diariamente. Este gargalo no processo de P&D impacta diretamente a capacidade da empresa de inovar e acompanhar o estado da arte, principalmente em segmentos que lidam com o desenvolvimento de tecnologias químicas, formulações e novos materiais.

Atualmente, a equipe de P&D dispense cerca de 20% de tempo de trabalho na busca manual de informações em bancos de patentes, resultando em sobrecarga dos colaboradores, bem como em atrasos tanto na identificação de oportunidades quanto na tomada de decisões estratégicas. Estima-se que, em média, cada busca manual leve de duas a três semanas, desde o recebimento de demanda de busca tecnológica até a entrega do resultado para a área solicitante, podendo representar um custo de oportunidade significativo para a empresa.

Além disso, a análise manual de patentes é suscetível a erros humanos e pode não capturar todas as informações relevantes, especialmente em documentos complexos e extensos, nos quais há elevado volume de informações disponíveis. Isso pode levar a decisões menos assertivas bem como a não identificação de tendências e tecnologias relevantes.

A falta de agilidade na análise de patentes também dificulta a identificação de novas abordagens, metodologias e materiais inovadores que poderiam impulsionar a inovação e melhorar as propriedades e funcionalidades dos produtos atuais. Como consequência, a empresa pode ficar defasada em relação aos concorrentes, perdendo oportunidades de lançar produtos inovadores e conquistar novos mercados.

Em resumo, a ineficiência no processo de busca e análise de patentes representa um obstáculo para o crescimento e a competitividade da empresa no setor químico, principalmente frente às concorrentes multinacionais. A adoção de soluções inovadoras, como a inteligência artificial, é fundamental para superar esses desafios e garantir que a empresa se mantenha na vanguarda da inovação e continue a se destacar em seu segmento de atuação.

## Objetivos do Projeto

Quais os objetivos pretendidos, quais os resultados esperados para cada um?

1. Reduzir o lead time para identificação e classificação de componentes químicos, ensaios e parâmetros experimentais em patentes.  
Resultado esperado: agilidade na identificação de termos relevantes, sem a necessidade de ler o documento completo.
2. Aumentar a qualidade de informações extraídas de patentes.  
Resultado esperado: ter assertividade na classificação e possibilitar o fornecimento de informações de valor.
3. Identificar novas abordagens, metodologias e materiais inovadores.  
Resultado esperado: aumentar a identificação de temáticas e estudos inovadores que contribuam para o direcionamento da inovação tecnológica e levantamento de histórico de estudos relevantes.

## Conjunto de Dados

Se possível, descrever brevemente o conjunto de dados (dataset) que será fornecido.  
Volume de dados, características, localização temporal e regional, formato, etc.  
Preferencialmente, inclua uma pequena amostra de dados.

### Dataset

Os dados mapeados para treinamento inicial dos modelos serão baseados no estudo do artigo “ChEMU 2020: Natural Language Processing Methods Are Effective for Information Extraction From Chemical Patents” (<https://doi.org/10.3389/frma.2021.654438>)

Dataset para download: <https://data.mendeley.com/datasets/wy6745bjfj/2>

São 900 conjuntos de dados incluindo parte de um texto extraído da literatura científica em par com a classificação dos termos químicos de interesse.

Os testes dos algoritmos deverão ser realizados com base de dados de patentes, seguindo as especificações definidas em “Informações Adicionais”.

## ChEMU - dataset

### Step 5. 4-Bromo-3-fluoro-5-methylbenzaldehyde

To a solution of 2-bromo-1-fluoro-3-methyl-5-vinylbenzene (5.46 g, 25.4 mmol) in acetone (46 mL) and water (4.6 mL) was sequentially added sodium periodate (21.7 g, 102 mmol) and a 4% aqueous solution of osmium tetroxide (8.07 mL, 1.27 mmol). The reaction was stirred at r.t. for 2 h. The reaction mixture was then filtered over a pad of celite, diluted with water, and extracted with ethyl acetate. The combined organic phases were washed with brine, dried over magnesium sulfate, and concentrated under reduced pressure. The crude product was purified by Biotage Isolera™ (3.22 g, 58%).

### Texto para análise

### Termos classificados

```
T0  EXAMPLE_LABEL 5 6 5
T1  TIME 326 329 2 h
T2  SOLVENT 230 237 aqueous
T3  OTHER_COMPOUND 431 444 ethyl acetate
T4  STARTING_MATERIAL 63 103 2-bromo-1-fluoro-3-methyl-5-vinylbenzene
T5  REAGENT_CATALYST 250 266 osmium tetroxide
T6  SOLVENT 147 152 water
T7  TEMPERATURE 317 321 r.t.
T8  REACTION_PRODUCT 579 586 product
T9  YIELD_OTHER 621 627 3.22 g
T10 OTHER_COMPOUND 491 496 brine
T11 OTHER_COMPOUND 384 390 celite
T12 REAGENT_CATALYST 185 201 sodium periodate
T13 OTHER_COMPOUND 405 410 water
T14 YIELD_PERCENT 629 632 58%
T15 REACTION_PRODUCT 8 45 4-Bromo-3-fluoro-5-methylbenzaldehyde
T16 SOLVENT 127 134 acetone
T17 OTHER_COMPOUND 509 526 magnesium sulfate
```

Figura 1. Exemplo de conjunto de dataset para treinamento.

### Reivindicação

1. A preparation method of an anticorrosive paint based on **benzoxazine resin** is characterized in that **benzoxazine monomers** and metal chloride are used as raw materials to carry out polymerization reaction to obtain the **benzoxazine resin**; and mixing the **benzoxazine resin** with **polyimide** to obtain the **benzoxazine resin**-based anticorrosive paint.
2. The preparation method of the **benzoxazine resin** based anticorrosive paint according to claim 1, wherein the **benzoxazine monomer** is dissolved in an organic solvent, and then a **metal chloride** solution is added dropwise for polymerization reaction at a temperature of 80-140 °C for 2-8 h.
3. The preparation method of the **benzoxazine resin** based anticorrosive paint according to claim 1, wherein the molar ratio of the **metal chloride** to the **benzoxazine monomer** is 1: 0.5-1: 8.
4. The method for preparing the **benzoxazine resin** based anticorrosive paint according to claim 1, wherein the **metal chloride** is FeCl<sub>3</sub>, AlCl<sub>3</sub> and CuCl<sub>2</sub>. At least one of (1).
5. The preparation method of the **benzoxazine resin** based anticorrosive coating according to claim 1, wherein a **phenolic** compound, a **primary amine** compound and a **formaldehyde** compound are used as raw materials to carry out a synthesis reaction to obtain the **benzoxazine monomer**, wherein the synthesis reaction is carried out at a temperature of 50-100 °C for 2-8 hours.
6. The method for preparing an anticorrosive coating based on **benzoxazine resin** according to claim 5, wherein the **phenolic compound** is at least one of **cardanol**, **urushiol**, **guaiaicol**, **eugenol**, **phenol** and **bisphenol A**, the **primary amine compound** is at least one of **butylamine**, **octylamine**, **dodecylamine**, **stearylamine**, **aniline**, **gamma-aminopropyltriethoxysilane**, **ethylenediamine** and **ethanolamine**, and the **formaldehyde** compound is **formaldehyde** or **paraformaldehyde**.
7. The method for preparing the benzoxazine resin based anticorrosive coating according to claim 5, wherein the **primary amine** compound: **phenolic** compounds: the molar ratio of the **formaldehyde** compound is 1: n:2n, wherein n is the number of **primary amines** in the **primary amine** compound.
8. The method for preparing the **benzoxazine resin** based anticorrosive paint according to claim 5, wherein the **benzoxazine monomer** is sequentially subjected to extraction, washing, filtering and rotary evaporation before the polymerization reaction.
9. The method for preparing the **benzoxazine resin** based anticorrosive paint according to claim 1, wherein the weight percentage of the **polyimide** is 0.1-10% of that of the **benzoxazine resin**, and the **polyimide** is a condensation type **aromatic polyimide**.
10. The preparation method of the **benzoxazine resin** based anticorrosive paint according to claim 1, wherein the **benzoxazine resin** and **polyimide** are mixed by an ultrasonic stirring method, and the ultrasonic stirring time is 10-60 min.

### Classificação de termos

STARTING\_MATERIAL benzoxazine monomer  
OTHER\_COMPOUND polyimide  
OTHER\_COMPOUND metal chloride  
REACTION\_PRODUCT benzoxazine resin  
STARTING\_MATERIAL benzoxazine monomer  
REACTION\_PRODUCT benzoxazine resin  
TEMPERATURE 80-140 °C  
TIME 2-8 h  
STARTING\_MATERIAL benzoxazine monomer  
REACTION\_PRODUCT benzoxazine resin  
OTHER\_COMPOUND metal chloride  
REACTION\_PRODUCT benzoxazine resin  
OTHER\_COMPOUND metal chloride  
REAGENT\_CATALYST FeCl<sub>3</sub>  
REAGENT\_CATALYST AlCl<sub>3</sub>  
REAGENT\_CATALYST CuCl<sub>2</sub>  
STARTING\_MATERIAL benzoxazine monomer  
REACTION\_PRODUCT benzoxazine resin  
STARTING\_MATERIAL phenolic  
STARTING\_MATERIAL primary amine  
STARTING\_MATERIAL formaldehyde  
TEMPERATURE 50-100 °C  
TIME 2-8 hours  
STARTING\_MATERIAL phenolic  
STARTING\_MATERIAL cardanol  
STARTING\_MATERIAL urushiol  
STARTING\_MATERIAL guaiaicol  
STARTING\_MATERIAL eugenol  
STARTING\_MATERIAL phenol  
STARTING\_MATERIAL bisphenol A  
STARTING\_MATERIAL primary amine  
STARTING\_MATERIAL butylamine  
STARTING\_MATERIAL octylamine  
STARTING\_MATERIAL dodecylamine  
STARTING\_MATERIAL stearylamine  
STARTING\_MATERIAL aniline  
STARTING\_MATERIAL gamma-aminopropyltriethoxysilane  
STARTING\_MATERIAL ethylenediamine  
STARTING\_MATERIAL ethanolamine  
STARTING\_MATERIAL formaldehyde  
STARTING\_MATERIAL paraformaldehyde  
STARTING\_MATERIAL primary amine  
STARTING\_MATERIAL phenolic  
STARTING\_MATERIAL formaldehyde  
STARTING\_MATERIAL benzoxazine monomer  
REACTION\_PRODUCT benzoxazine resin  
OTHER\_COMPOUND polyimide  
REACTION\_PRODUCT benzoxazine resin  
OTHER\_COMPOUND aromatic polyimide  
REACTION\_PRODUCT benzoxazine resin  
OTHER\_COMPOUND polyimide  
TIME 10-60 min

## Competências/habilidades trabalhadas

Marcar com um [X] os itens que se esperam que sejam trabalhados/desenvolvidos:

- ☒ Programação Java/Python
- ☒ Técnicas de para coleta, preparação e análise de dados
- ☐ Probabilidade e Estatística
- ☐ Álgebra linear e Cálculo
- ☐ Análise multivariada
- ☒ Bibliotecas de análise de dados
- ☐ Modelagem de Dados
- ☒ Bancos de Dados relacionais e não-relacionais
- ☒ Inteligência Artificial
- ☒ Aprendizado supervisionado
- ☒ Aprendizado não-supervisionado
- ☒ Aprendizado por reforço
- ☐ Sistemas de Recomendação
- ☒ Aprendizado Profundo (deep learning)
- ☒ Processamento de Linguagem Natural (PLN)
- ☐ Previsão de Séries Temporais
- ☒ Visualização de Dados
- ☐ Visão computacional
- ☐ Outras: \_\_\_\_\_

## Categorização do Problema

Escreva a categoria do problema. Exemplos: extração de informações, predição de custos, categorização de textos, detecção de fraudes, reconhecimento de padrões e objetos, comportamento de compras, etc.

Extração de informações, categorização de textos.

## Informações Adicionais

Utilize esse espaço para quaisquer informações adicionais que julgar necessário.

### Regras de Negócio

Para a validação utilizar preferencialmente a Classificação de Patentes ([IPC](#)): C09, C23C, C08F, C08G, C08K, C08L e H01B. Estes são os mais coerentes com as tecnologias que utilizamos na WEG Tintas.

### Patentes

As patentes podem ser acessadas de fontes públicas como: Google Patents, Lens.org entre outros. Este tipo de documento possui uma estrutura padrão constituída basicamente de:

1. Título
2. Resumo
3. Relatório descritivo

4. Reivindicações (parte mais importante em que se encontram informações de detalhes inovadores a serem protegidos)
5. Desenhos

### Sinônimos

Para lidar com os sinônimos dos termos químicos, que frequentemente podem aparecer em distintos estudos com uma nomenclatura diferente, mas se referindo ao mesmo composto, podem ser usados base de dados como do PubChem e/ou do ChemSpider:

<https://github.com/CalebBell/chemicals/>

<https://medium.com/@nshangguan/access-chemical-database-with-python-9cddfcc12477>

## ANEXO A – Disciplinas de Projeto em Ciência de Dados

O curso de Bacharelado em Ciência de Dados e Inteligência Artificial contempla as seguintes disciplinas de caráter integrador, a saber: Projeto em Ciência de Dados I, Projeto em Ciência de Dados II e Projeto em Ciência de Dados III.

O objetivo destas disciplinas é fazer com que o aluno tenha a oportunidade de desenvolver projetos integrados, utilizando conceitos trabalhados em disciplinas previamente ministradas na matriz curricular, além do compartilhamento de experiências obtidas por meio desta integração. O trabalho desenvolvido nestas disciplinas busca, além do caráter formativo, preparar os alunos para as situações reais que o mercado de trabalho na área de Ciência de Dados apresenta.

Na disciplina Projeto em Ciência de Dados I espera-se ao final que o aluno tenha desenvolvido uma solução que envolva um processo completo de descoberta de conhecimento. Neste sentido, esta disciplina tem como pré-requisitos as disciplinas de Banco de Dados II, Coleta, Preparação e Análise de Dados, e Aprendizado Supervisionado.

Na disciplina Projeto em Ciência de Dados II espera-se que o aluno, além de desenvolver um projeto que empregue um processo completo de descoberta de conhecimento, também seja capaz de explorar outras formas de aprendizado (não supervisionado, por reforço e profundo) para projetos avançados. Neste sentido, esta disciplina tem como pré-requisito a disciplina Projeto em Ciência de Dados I.

Por fim, na disciplina de Projeto em Ciência de Dados III, o aluno deve consolidar em um trabalho todos os conhecimentos e habilidades adquiridas ao longo do curso. Esta disciplina tem como pré-requisito a disciplina Projeto em Ciência de Dados II.

Nas três disciplinas os alunos serão distribuídos em times, na forma de ilhas de trabalho, de acordo com os tipos de problemas abordados e as técnicas empregadas. Além disso, dinâmicas de interação serão utilizadas para o compartilhamento de lições aprendidas ao longo dos projetos desenvolvidos e a revisão de conteúdos, sempre que necessário, para o melhor desenvolvimento dos projetos.

### EMENTA

Aplicação das competências construídas durante os semestres anteriores de forma integrada, através da identificação de uma oportunidade de projeto em ciência de dados, e da modelagem e implementação do mesmo. Uso de dinâmicas de interação para compartilhamento de lições aprendidas nos projetos desenvolvidos.

### OBJETIVOS

O cumprimento da disciplina busca dar ao aluno condições de elaborar soluções utilizando conceitos, metodologias, técnicas e ferramentas estudadas até o momento no curso, desenvolvendo suas habilidades de um ponto de vista prático como formação complementar.

## ANEXO B – Cronograma de Referência

AULA	CONTEÚDO
1	Apresentação da Disciplina; Coleta de dados sobre o perfil do grupo.
2	Avaliação do perfil grupo (disciplinas cursadas, competências, habilidades); Organização dos grupos de trabalho. Nivelamento intergrupos e intragrupos.
3 (*)	Apresentação do(s) projeto(s) do semestre (organização parceira)
4	Planejamento do Projeto
5	Planejamento do Projeto
6	R1 (CRISP: Compreensão do Negócio + Compreensão dos Dados)
7 (*)	Apresentação das propostas dos grupos à empresa parceira
8	Execução do Projeto
9	R2 (CRISP: Preparação dos Dados)
10 (*)	Apresentação de andamento dos grupos à empresa parceira
11	Execução do Projeto
12	Execução do Projeto
13	Execução do Projeto
14	R3 (CRISP: Modelagem + Avaliação)
15 (*)	Entrega dos projetos e dos relatórios finais dos grupos (RF). Apresentações finais dos grupos à empresa parceira.
16	Reflexões sobre a execução dos projetos
17	Entrega dos relatórios individuais

(\*) momentos que envolvem a participação da organização parceira

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL  
GRADUAÇÃO EM CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL

DISCIPLINA DE PROJETO EM CIÊNCIA DE DADOS  
Coord. Prof. Dr. Daniel Callegari