Descoberta de Conhecimento: GitHub

```
SEMESTRE  =  "2025/1"
DISCIPLINA  =  "Mineração e Análise de Redes Sociais"
ESTUDANTES  = ["Gabriel de Paula", "Wasterman Apolinário"]
PROFESSOR  =  "Vinicius Vieira"
```

1. Introdução

Essa atividade tem como foco principal praticar a metodologia de Descoberta de Conhecimento em Bases de Dados (KDD - Knowledge Discovery in Databases). É necessário, portanto, seguir bem os passos do processo: coletar dados, estruturar a base, aplicar técnicas e refletir sobre os resultados. Ou seja, o objetivo é entender como cada etapa funciona na prática, experimentando o uso das técnicas vistas em sala.

Para isso foi escolhido o GitHub como rede social real, já utilizado nas outras etapas da disciplina, e, a partir disso, houve a criação de uma base estruturada com atributos bem definidos. O foco foi transformar os dados brutos em algo organizado, que permitisse a aplicação das etapas do KDD.

2. Construção da base de dados

Para a construção da base de dados, por se tratar de uma rede social voltada para desenvolvedores, é possível relacionar usuários não apenas por se seguirem, mas considerando também aspectos dos repositórios de cada um. O foco foi extrair dados relacionados aos usuários e seus relacionamentos dentro da plataforma, de modo a representar a rede de conexões e a atividade colaborativa presente ali.

2.1. Atributos selecionados

Alguns dados são particularmente interessantes para estabelecer conexões e identificar padrões de comportamento entre diferentes usuários, pois permitem compreender como determinados perfis se relacionam, compartilham interesses ou apresentam trajetórias semelhantes dentro de um sistema.

Dado coletado	**Tipo**
Nome do usuário	String
Quantidade de seguidores	Número
Quantidade de "seguindo"	Número
Quantidade de repositórios públicos	Número
Nome da Empresa	String
Linguagens mais utilizadas	String []
Quantidade média de estrelas	Número

2.2. Tratamentos nos dados

Para possibilitar associações mais precisas entre os usuários, foi necessário realizar um tratamento prévio dos dados. Esse processo envolveu a padronização de formatos e a remoção de duplicidades, garantindo maior integridade e confiabilidade dos registros. Além disso, a transformação de certos atributos em categorias mais significativas facilitou a identificação de padrões e a comparação entre perfis distintos.

Dado coletado	**Tipo**
Nome do usuário	String
Quantidade de seguidores	Número
Quantidade de "seguindo"	Número
Quantidade de repositórios públicos	Número
Está em uma empresa	Booleano
Linguagens mais utilizadas	{ nome: String, quantidade: Número} []
Quantidade média de estrelas	Número

Com os dados devidamente tratados, tornou-se possível aplicar técnicas de análise mais eficazes, permitindo relações mais robustas e representativas entre os usuários.

3. Análise e Resultados

4. Discussão