

Integração Semântica

Produtos Alimentares

Realizado por:
César Ferreira
Jóni Batista

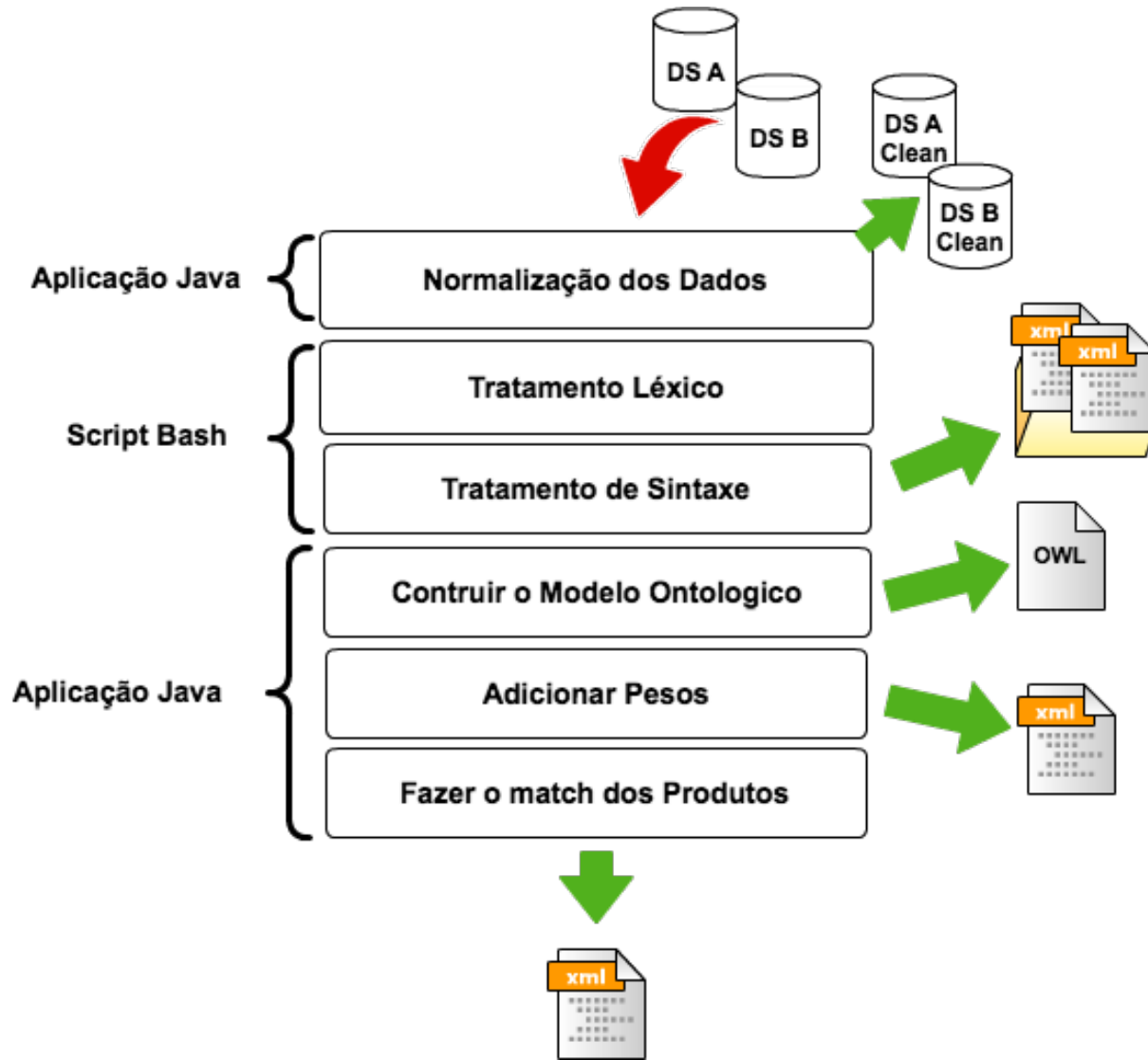
Problema

Fazer o mapeamento de informação de produtos alimentares entre dois sistemas heterogêneos, de forma a obter uma fonte de dados com maior granularidade de informação de cada produto.

Comparar duas Strings!?!

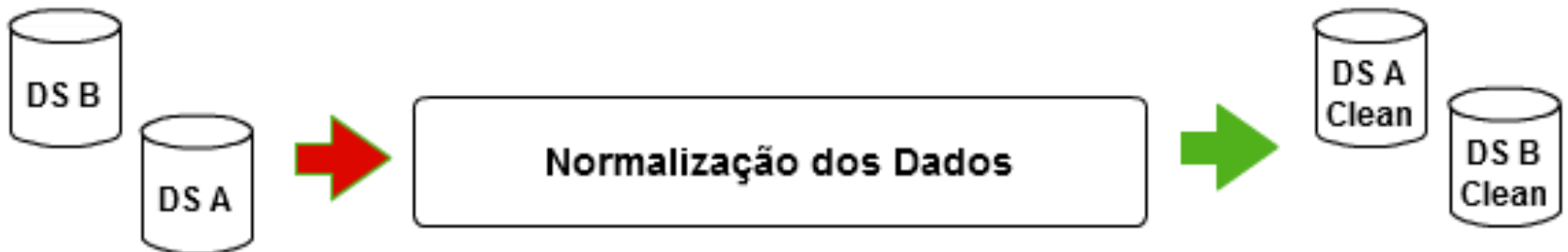
Sim, **parece** fácil...

Workflow

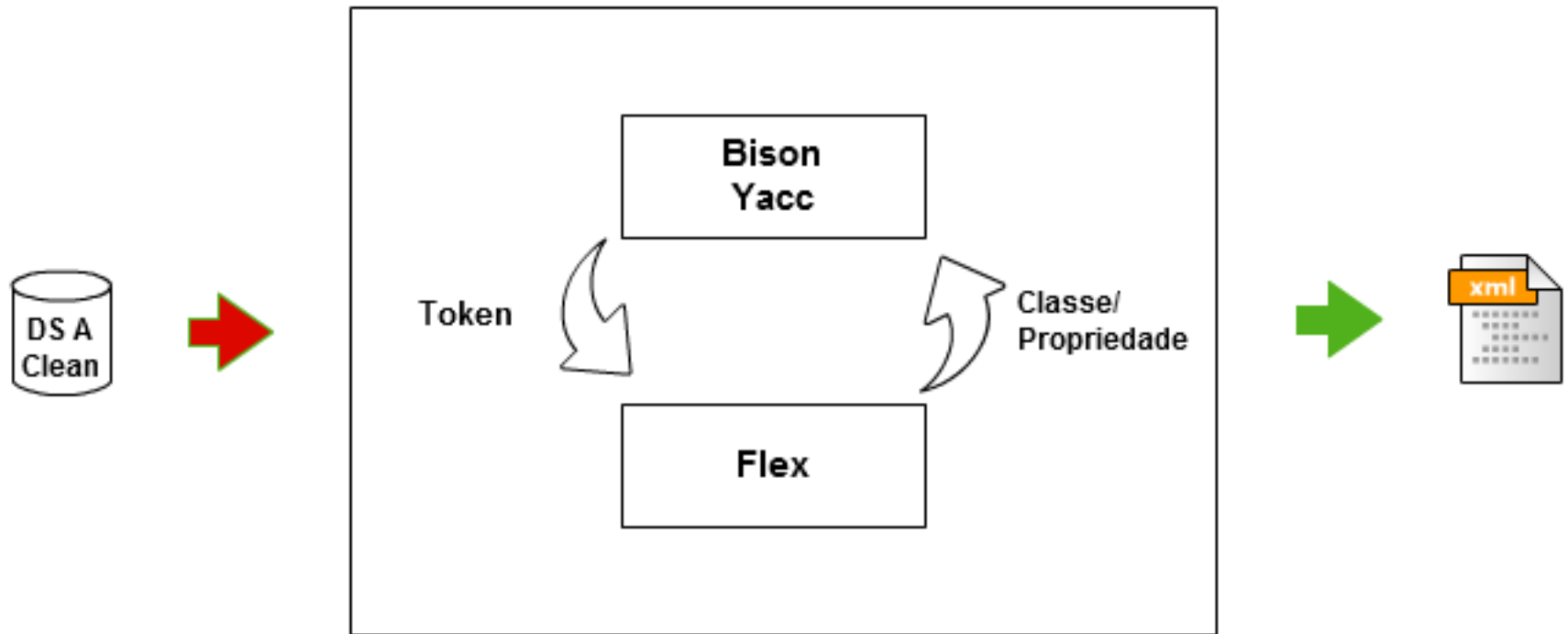


Normalização

- Remover acentos dos vocabulos;
- Converter strings para minúsculas;
- Remover caracteres especiais que não sejam relevantes;
- Aplicação em Java.



Tratamento Léxico e Gramatical

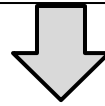


Tratamento Léxico e Gramatical

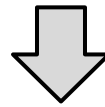
- Flex para definir o vocabulário domínio;
- Bison-Yacc para a sintaxe/gramática;
- Script em bash para enviar strings de produtos para dentro do Yacc;
- Este por sua vez envia token a token para o Flex;
- O Flex devolve a definição do vocábulo ao Yacc;
- O Yacc escreve a sintaxe para um ficheiro xml.

Tratamento Léxico e Gramatical

Vinho tinto borba 0,75cl

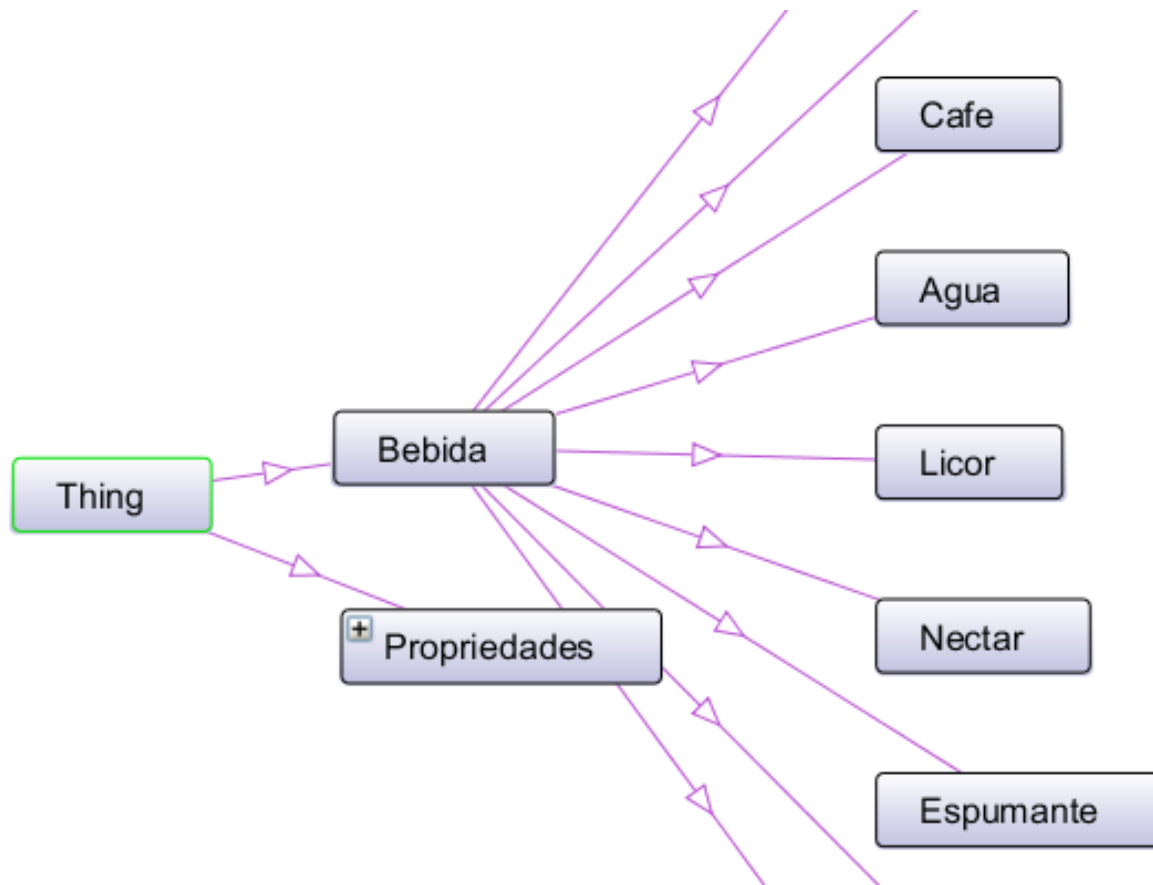


Tratamento Léxico e Gramatical



```
- <xml version="1.0" encoding="utf-8">
- <linha>
  <produto familia="bebida">vinho</produto>
  <propriedade atributo="marca">borba</propriedade>
  <propriedade atributo="quantidade">0.75</propriedade>
  <propriedade atributo="grandeza">cl</propriedade>
  <propriedade atributo="cor">tinto</propriedade>
</linha>
..
</xml>
```

Ontologia



Pesos

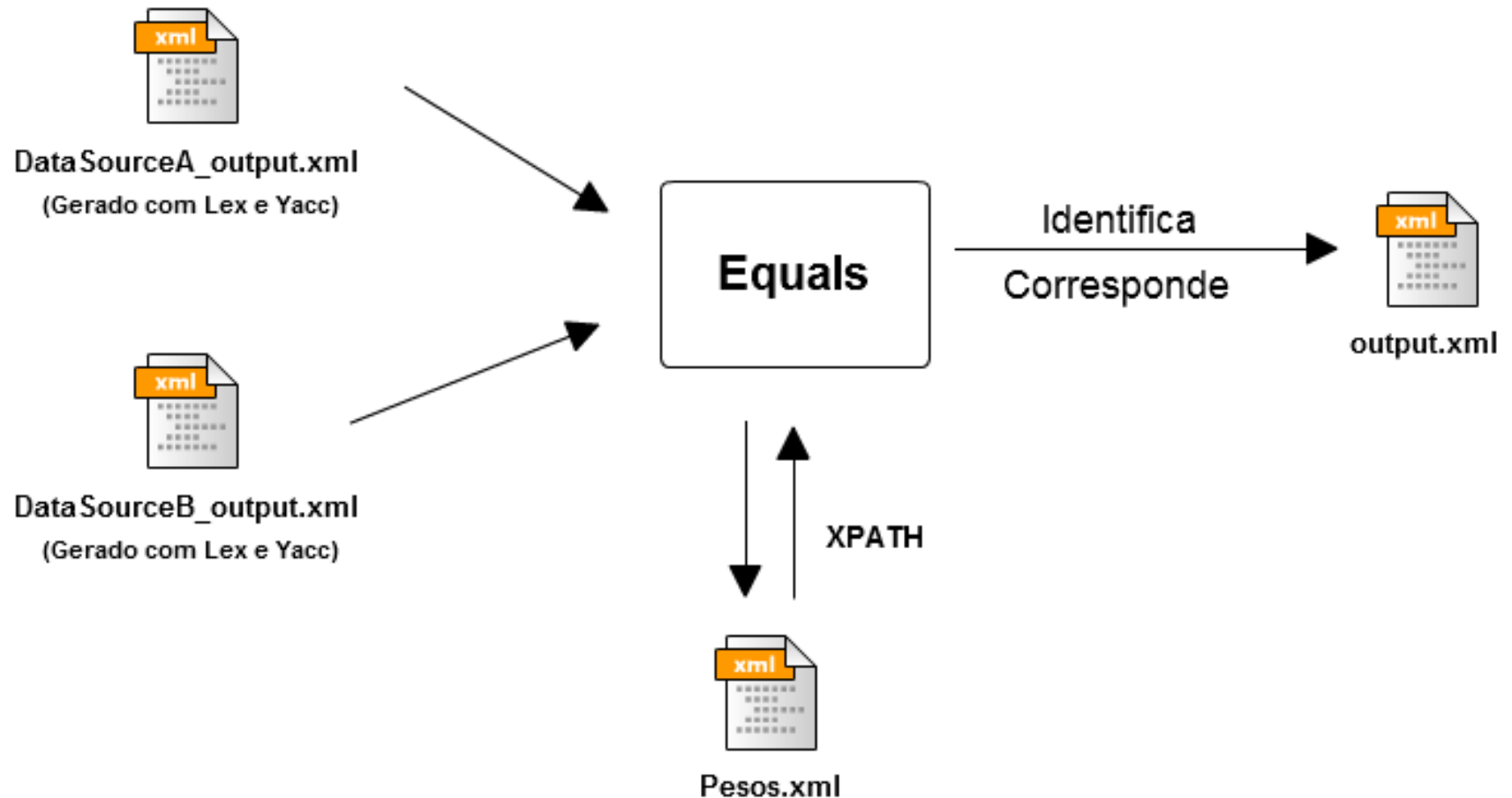
- Mecanismo utilizado para identificar se dois produtos são iguais;
- Cada característica do produto terá um peso;
- O produto do DS B correspondente ao produto em causa será o que tiver o sumatório dos pesos mais elevado;
- Só serão sumados os pesos das

Pesos

Extracto dos pesos.

```
<xml version="1.0" encoding="utf-8">  
  <classe familia="bebidas">80</classe>  
  <propriedade atributo="geral">5</propriedade>  
  <propriedade atributo="grandeza">20</propriedade>  
  <propriedade atributo="quantidade">20</propriedade>  
  <propriedade atributo="cor">40</propriedade>  
  <propriedade atributo="ano">5</propriedade>  
  <propriedade atributo="origem">10</propriedade>  
  <propriedade atributo="tipoprocessamento">20</propriedade>
```

Matching dos Produtos



Matching dos Produtos

- Para cada Produto do DS A vai procurar o mais semelhante em B;
- O produto mais semelhante é o que tiver a soma dos pesos mais elevada;
- No final é gravado o output num ficheiro xml;
- Este ficheiro contém as características do DS A e do seu correspondente em B.

Matching dos Produtos

Extracto do output final

```
- <xml version="1.0" encoding="utf-8">
  - <linha valor="vinho borba 0.75 cl tinto">
    <produto familia="bebida">vinho</produto>
    <propriedade atributo="marca">borba</propriedade>
    <propriedade atributo="quantidade">0.75</propriedade>
    <propriedade atributo="grandeza">cl</propriedade>
    <propriedade atributo="cor">tinto</propriedade>
    <propriedade atributo="tipobebida">alcoolicas</propriedade>
    <propriedade atributo="tipoprocessamento">fermentadas</propriedade>
    <propriedade atributo="classificacao">maduro</propriedade>
  </linha>
  ...
</xml>
```

Trabalhos Futuros

- Fazer stemming aos tokens;
- Completar a Ontologia;
- Optimização dos pesos;
- Aplicar aos restantes produtos.

Conclusões/Dificuldades

- Elevado número de produtos;
- Muitas ferramentas para um curto espaço de tempo;
- Utilização de ferramentas muito antigas;
- Poderíamos ter usado ferramentas de Ontologia para fazer a parte gramatical e lexic.

Questões

