Ferramentas Java para Recuperação e Mineração de Informações

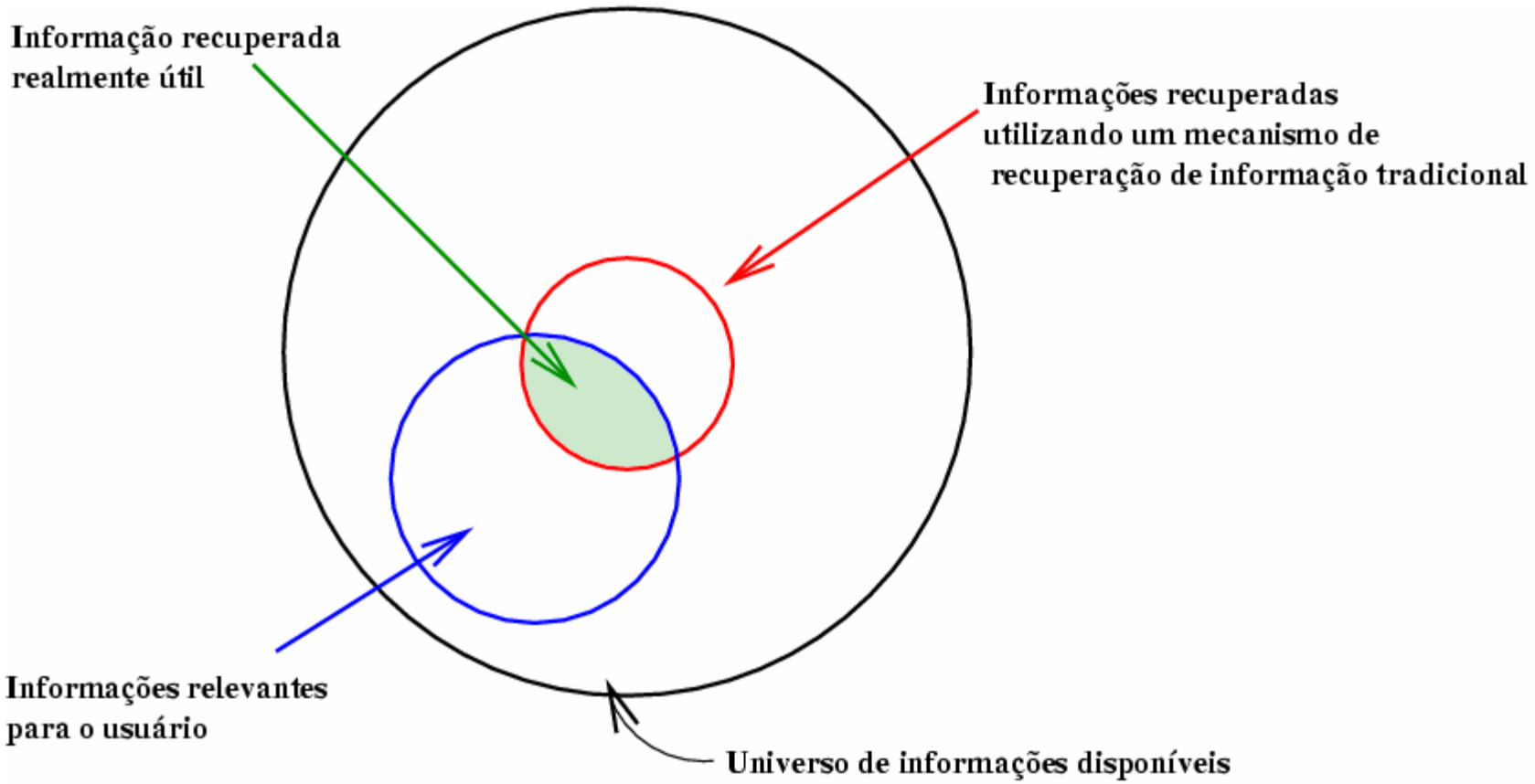
Fabrício J. Barth[1,2]

[1]Fundação Atech Tecnologias Críticas (fbarth@atech.br)
[2]Centro Universitário SENAC (fabricio.jbarth@sp.senac.br)

9 de setembro de 2008

Contexto: Enorme quantidade de dados que precisa ser processada

Problema 1: **recuperar** a informação



Problema 2: **tratar** a informação recuperada

Web Resultados **1** - **100** de aproximadam nte **174.000.000** para anga **(0,14** segundos)

Mangá - Wikipédia, a enciclopédia livre
O **mangá** ou **manga** (漫画, **Manga**?) é a palavra usada para designar as histórias em quadrinhos japonesas, o seu estilo próprio de desenho e o movimento ...
pt.wikipedia.org/wiki/**Mangá** - 63k - Em cache - Páginas Semelhantes - Anotar isso

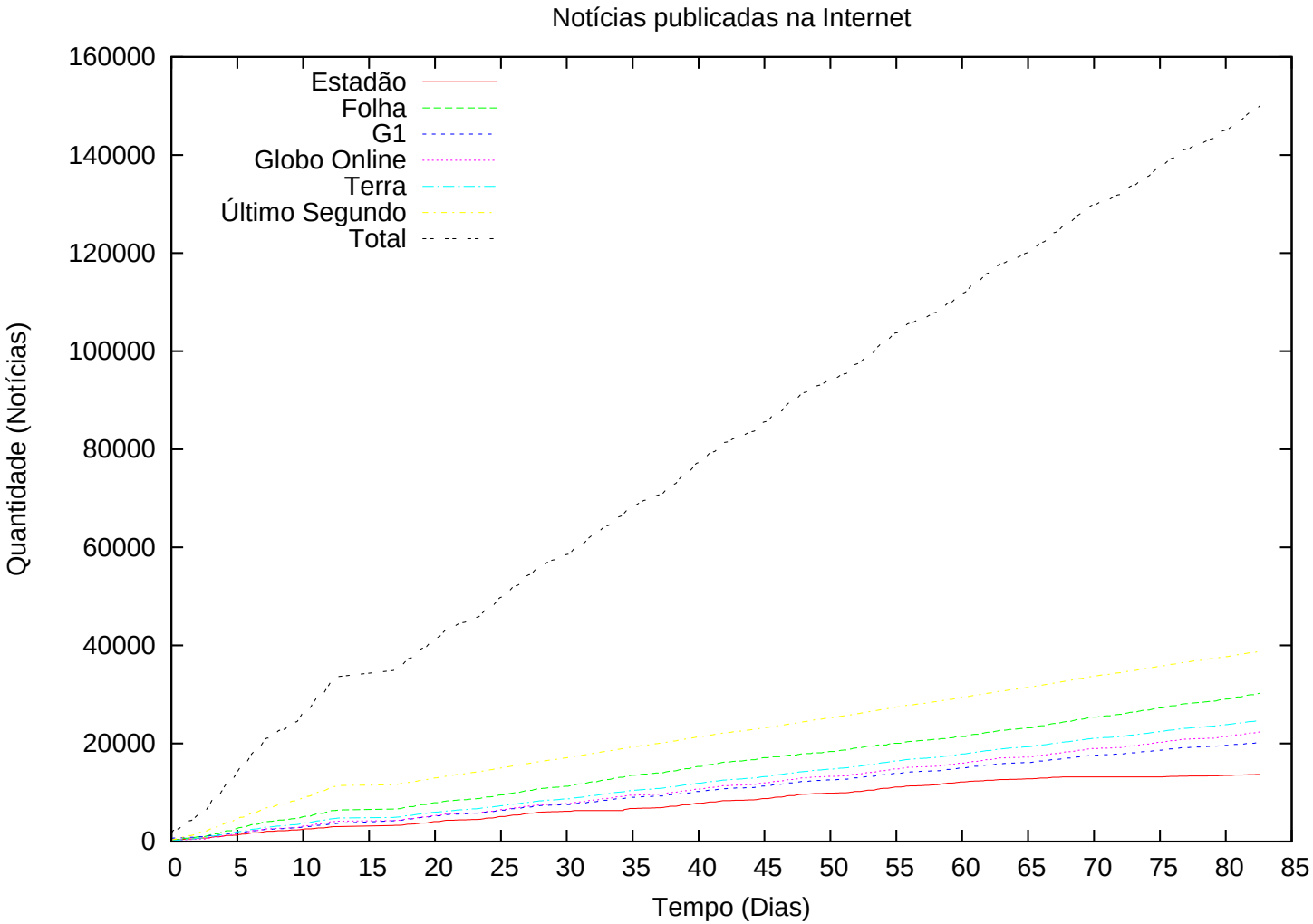
Manga - Wikipédia, a enciclopédia livre
Mangá - histórias em quadrinhos japonesas, grafadas **Manga** em português europeu. ... **Manga** - Haílton Corrêa de Arruda, jogador brasileiro de futebol . ...
pt.wikipedia.org/wiki/**Manga** - 18k - Em cache - Páginas Semelhantes - Anotar isso
Mais resultados de pt.wikipedia.org »

Manga.com - [Traduzir esta página]
Official site of **Manga** Entertainment, publisher of anime titles such as Astro Boy, Ghost in the Shell, Ninja Scroll, Blood: The Last Vampire, and many more.
www.**manga**.com/ - 28k - Em cache - Páginas Semelhantes - Anotar isso

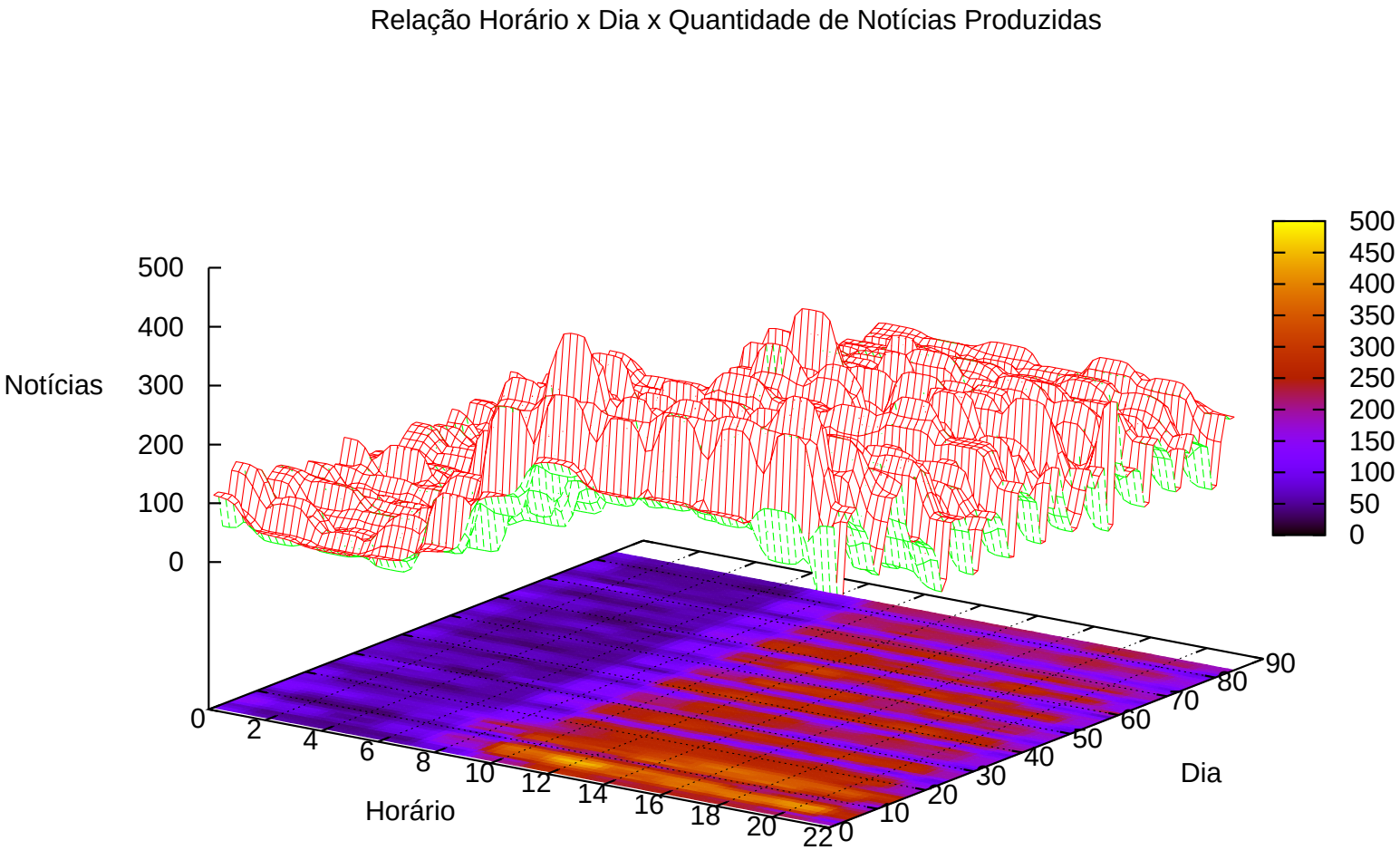
Manga - Wikipedia, the free encyclopedia - [Traduzir esta página]
In Japan, **manga** are widely read by people of all ages, [2] and include a broad range of subjects: action-adventure, romance, sports and games, ...
en.wikipedia.org/wiki/**Manga** - 140k - Em cache - Páginas Semelhantes - Anotar isso

Guia de **Mangá**
Conheça o que é o **mangá**, o gênero de quadrinhos japonês que deu origem à 99% dos animes. Fique por dentro dos diversos gêneros e aprenda a escolher o que ...
www.sobresites.com/**manga**/ - 21k - Em cache - Páginas Semelhantes - Anotar isso

Quantidade de notícias produzidas na Web?



Quantidade de notícias produzidas na Web?



Problema e Sugestões

Problema:

Para tirar proveito desta informação é necessário organizá-la de alguma forma...

Problema e Sugestões

Problema:

Para tirar proveito desta informação é necessário organizá-la de alguma forma...

Sugestões:

- Agrupamento de Notícias.
- Classificação, Recomendação e Filtragem de Notícias.

Por que agrupar notícias?

- Como agrupá-las?

Definições de Algoritmos de Agrupamento

- O objetivo dos algoritmos de agrupamento é colocar os objetos similares em um mesmo grupo e objetos não similares em grupos diferentes.
- Normalmente, objetos são descritos e agrupados usando um conjunto de atributos e valores.
- Não existe nenhuma informação sobre a classe ou categoria dos objetos.

Formato de um documento

... Esta disciplina tem como objetivo apresentar os principais conceitos da área de Inteligência Artificial, caracterizar as principais técnicas e métodos, e implementar alguns problemas clássicos desta área sob um ponto de vista introdutório.
A estratégia de trabalho, o conteúdo ministrado e a forma dependerão dos projetos selecionados pelos alunos. Inicialmente, os alunos deverão trazer os seus Projetos de Conclusão de Curso, identificar intersecções entre o projeto e a disciplina, e propor atividades para a disciplina. ...

Atributo/Valor usando vetores

Como representar os documentos?

Atributo/Valor usando vetores

Como representar os documentos?

$$\overrightarrow{d_i} = (p_{i1}, p_{i2}, \cdots, pin) \quad (1)$$

- Os atributos são as palavras que aparecem nos documentos.

Atributo/Valor usando vetores

Como representar os documentos?

$$\overrightarrow{d_i} = (p_{i1}, p_{i2}, \cdots, pin) \quad (1)$$

- Os atributos são as palavras que aparecem nos documentos.
- Se todas as palavras que aparecem nos documentos forem utilizadas, o vetor não ficará muito grande?

Diminuindo a dimensionalidade do vetor

- Como filtrar as palavras que devem ser usadas como atributos?
- Em todos os idiomas existem átomos (palavras) que não significam muito. Stop-words

Esta disciplina tem como objetivo apresentar os principais conceitos da área de Inteligência Artificial, caracterizar as principais técnicas e métodos, e implementar alguns problemas clássicos desta área sob um ponto de vista introdutório.
...

Diminuindo ainda mais a dimensionalidade do vetor

- Algumas palavras podem aparecer no texto de diversas maneiras: técnica, técnicas, implementar, implementação...
- Stemming - encontrar o radical da palavra e usar apenas o radical.

Atributo/Valor usando vetores

- Já conhecemos os atributos.
- E os valores?

Atributo/Valor usando vetores

- Já conhecemos os atributos.
- E os valores?
  - Booleana - se a palavra aparece ou não no documento (1 ou 0)
  - Por freqüência do termo - a freqüência com que a palavra aparece no documento (normalizada ou não)
  - Ponderação tf-idf - o peso é proporcional ao número de ocorrências do termo no documento e inversamente proporcional ao número de documentos onde o termo aparece.

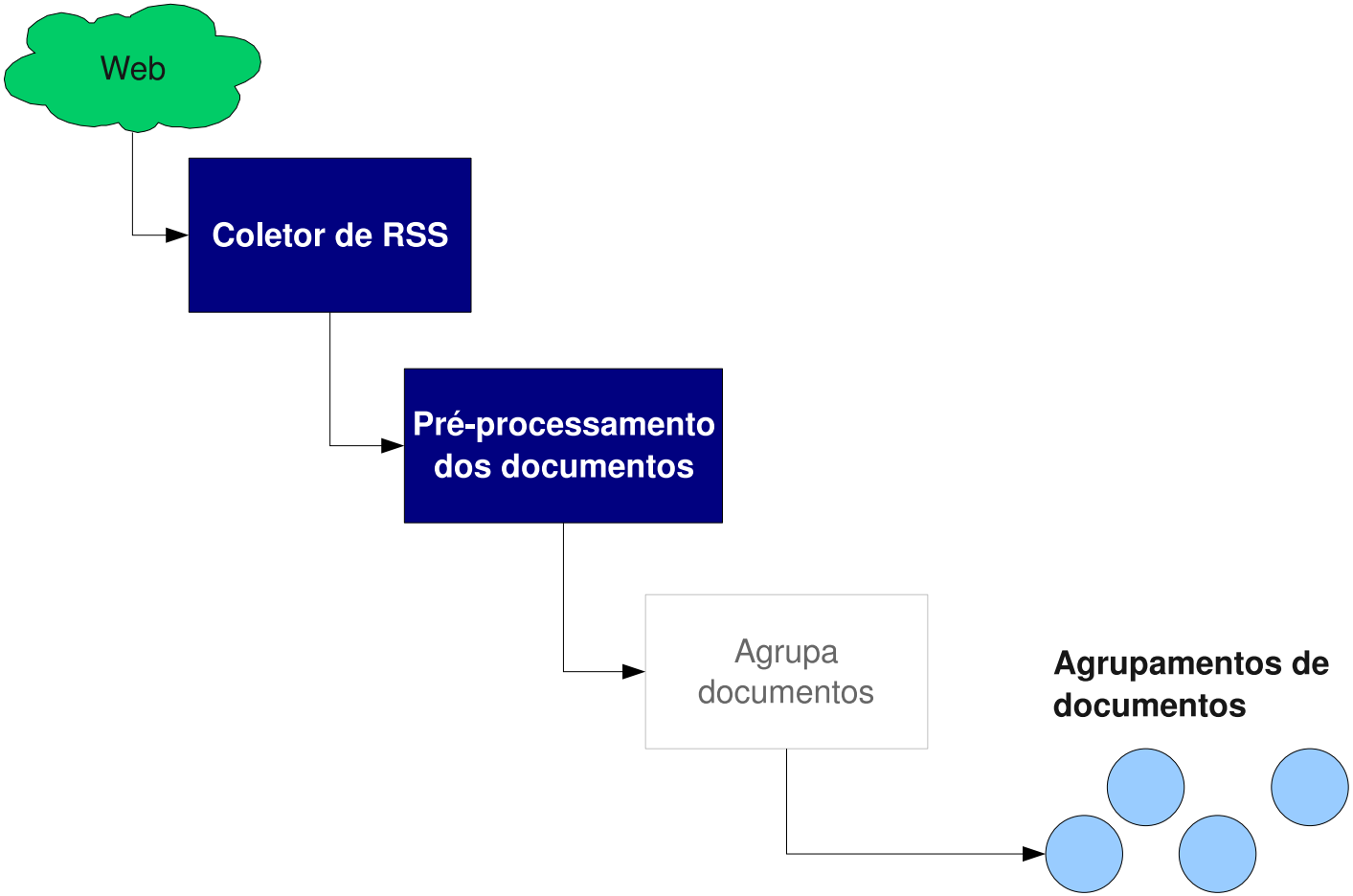
Por freqüência do termo

(apresent,0.33) (form,0.33) (tecnic,0.33) (caracteriz,0.33) (projet,1.0) (introdutori,0.33) (objet,0.33) (inteligente,0.33) (conclusa,0.33) (selecion,0.33) (intersecco,0.33) (classic,0.33) (identific,0.33) (conceit,0.33) (trabalh,0.33) (disciplin,1.0) (traz,0.33)

Conjunto de Exemplos - Atributo/Valor

Doc.	apresent	form	tecnic	caracteriz	...
d_1	0.33	0.33	0.33	0.33	...
d_2	0	0.5	0.2	0.33	...
d_3	1	0.6	0	0	...
d_4	0.4	0.3	0.33	0.4	...
d_5	1	0.4	0.1	0.1	...
d_n	...	...	...	...	...

Componentes da Solução

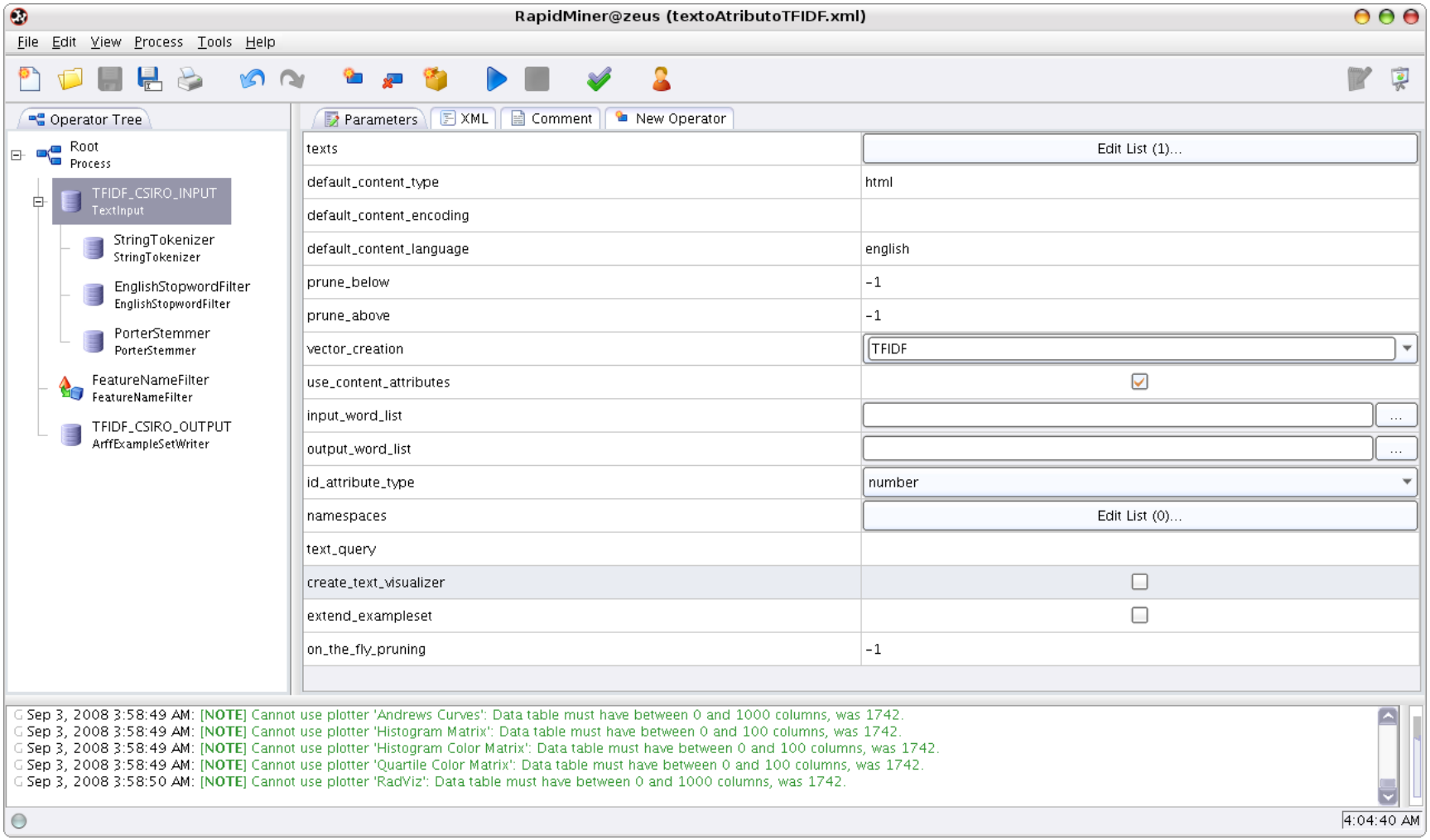


Coletor de RSS e Pré-Processamento

function coletorRSS(Lista de RSS): tabela

i $\leftarrow$ 0;
for cada rss_i em RSS **do**
  for cada $noticia_j$ em rss_i **do**
    $d_i \leftarrow d_i + rss_i.noticia_j$.TITLE;
    $d_i \leftarrow d_i + rss_i.noticia_j$.DESCRIPTION;
    $d_i \leftarrow$ eliminaStopWords(d_i);
    $d_i \leftarrow$ stemming(d_i);
    i $\leftarrow$ i + 1;
  end for
end for
return criaTabelaExemplos(d,TF-IDF);

Pré-processamento dos documentos - *RapidMiner*



Características e Funcionalidades do *RapidMiner*

- O usuário define um processo de tratamento dos dados.
- Os operadores podem ser divididos nas seguintes categorias:
  - *IO*
  - *Learner (Supervised / Unsupervised)*
  - *OLAP (On-line Analytical Processing)*
  - *Postprocessing*
  - *Preprocessing*
  - *Validation*
  - *Visualization*
- Cada operador pode ser devidamente configurado.
- Existem ambientes para: definição do processo e execução do processo.

Pré-processamento dos documentos - *RapidMiner*

RapidMiner@zeus (textoAtributoTFIDF.xml)

File Edit View Process Tools Help

ExampleSet

Meta Data View | Data View | Plot View

ExampleSet (10 examples, 0 special attributes, 1742 regular attributes) View Filter (10 / 10): all

row no.	network	februari	member	networkcsi...	wait	pmb	osmond	ph	mob	fax	email	elizabeth	heij	aufeatur	thought	love
1	0.131	0.004	0.018	0.004	0.004	0.004	0.004	0.004	0.004	0.002	0.016	0.018	0.014	0.004	0.014	0.007
2	0.131	0.004	0.018	0.004	0.004	0.004	0.004	0.004	0.004	0.002	0.016	0.018	0.014	0.004	0.014	0.007
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0.047	0.047	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0.015	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0.024	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0.014	0.014	0	0	0	0	0

Save...

Sep 3, 2008 3:58:49 AM: [NOTE] Cannot use plotter 'Andrews Curves': Data table must have between 0 and 1000 columns, was 1742.
Sep 3, 2008 3:58:49 AM: [NOTE] Cannot use plotter 'Histogram Matrix': Data table must have between 0 and 100 columns, was 1742.
Sep 3, 2008 3:58:49 AM: [NOTE] Cannot use plotter 'Histogram Color Matrix': Data table must have between 0 and 100 columns, was 1742.
Sep 3, 2008 3:58:49 AM: [NOTE] Cannot use plotter 'Quartile Color Matrix': Data table must have between 0 and 100 columns, was 1742.
Sep 3, 2008 3:58:50 AM: [NOTE] Cannot use plotter 'RadViz': Data table must have between 0 and 1000 columns, was 1742.

4:05:19 AM

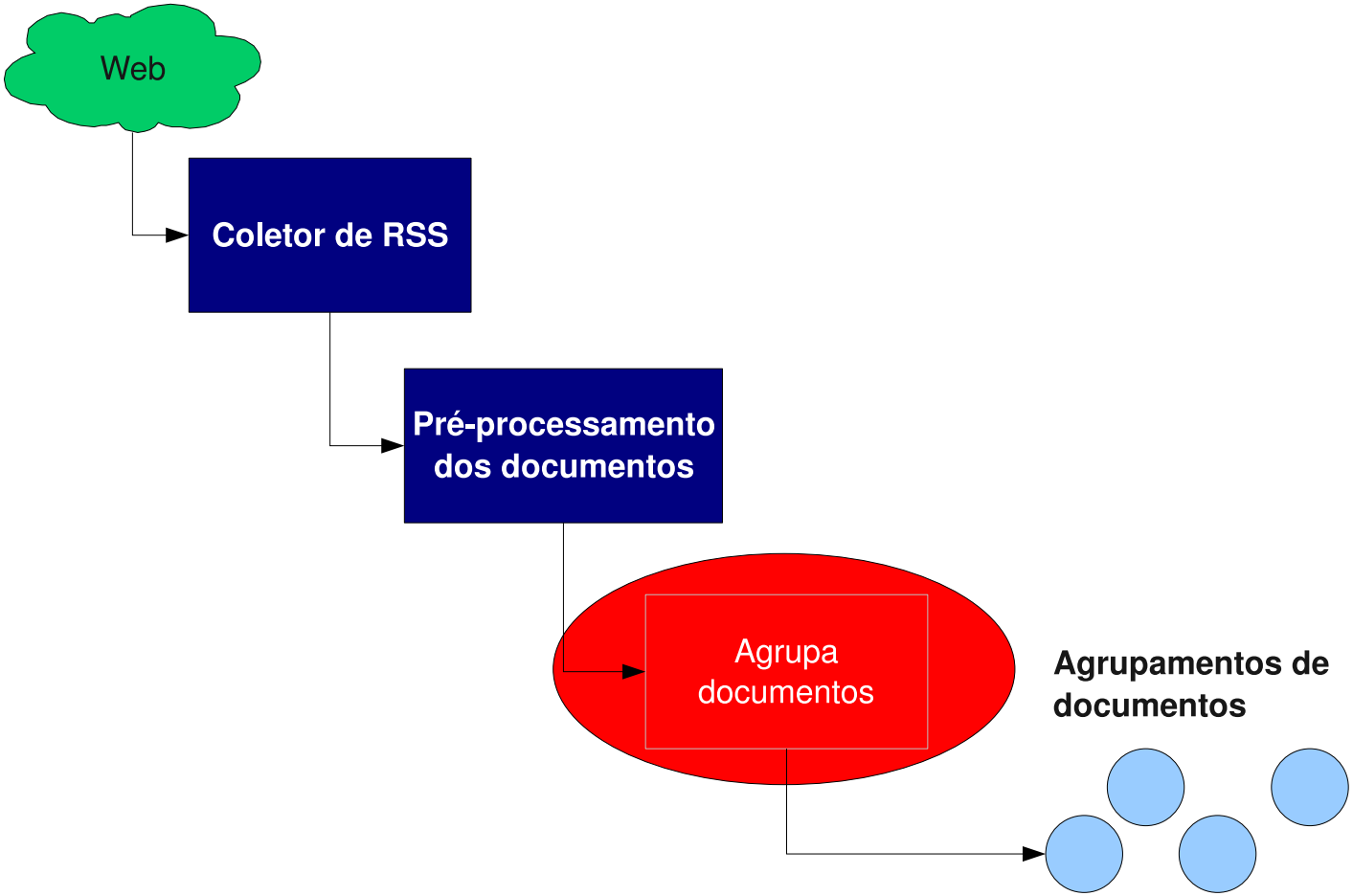
Pré-processamento dos documentos - Código

```
import edu.udo.cs.wvtool.generic.stemmer.PorterStemmerWrapper;
import edu.udo.cs.wvtool.generic.tokenizer.SimpleTokenizer;
import edu.udo.cs.wvtool.generic.wordfilter.StopWordsWrapper;

...

public String manipulaTextoComStemming(String nomeArquivo){
    try{
        WVTDocumentInfo documentInfo = new WVTDocumentInfo
            (null,"html",null,"english");
        SimpleTokenizer tokenizer = new SimpleTokenizer();
        TokenEnumeration tokens = tokenizer.tokenize(
                new InputStreamReader(
                        new FileInputStream(nomeArquivo)), documentInfo);
        PorterStemmerWrapper stemmer = new PorterStemmerWrapper();
        StopWordsWrapper stopWords = new StopWordsWrapper();
        TokenEnumeration tokenSemStopWord = stopWords.filter(
                                                tokens, documentInfo);
        String retorno = "";
        while(tokenSemStopWord.hasMoreTokens())
            retorno = retorno +
                stemmer.getBase(tokenSemStopWord.nextToken())+" ";
        return retorno;
    ...
```

Conjunto de treinamento - Arquivo ARFF

```
@RELATION RapidMinerData

@ATTRIBUTE 'network' real
@ATTRIBUTE 'februari' real
@ATTRIBUTE 'member' real
@ATTRIBUTE 'wait' real
...
...

@DATA
0.1313298612447743,0.004041576682790196,0.01774727854659112,0.003549455709318225
0.1313298612447743,0.004041576682790196,0.01774727854659112,0.003549455709318,0
0.0,0.0,0.0,0.0
0.0,0.0,0.0,0.0
0.0,0.014857582309589007,0.0,0.0024848992203904758
...
...
```

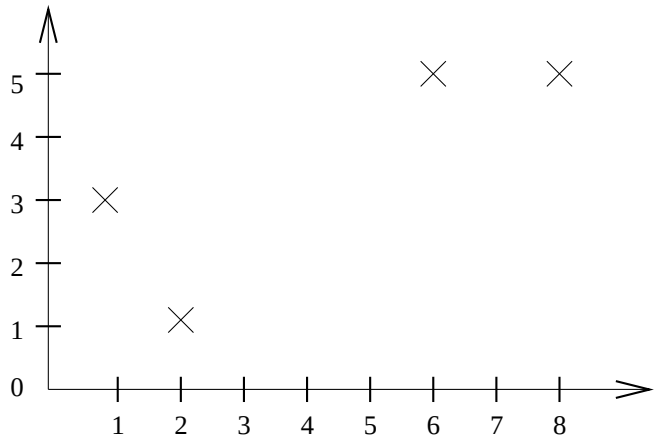
Que algoritmo de agrupamento utilizar?



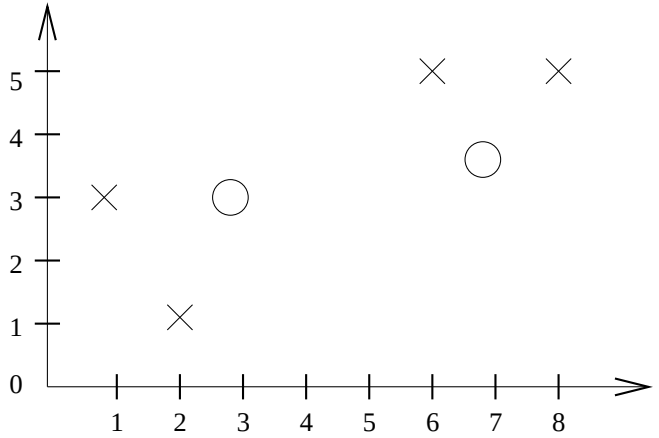
Algoritmos para Agrupamento - *K-means*

- K significa o número de agrupamentos (que deve ser informado à priori).
- Sequência de ações iterativas.
- A parada é baseada em algum critério de qualidade dos agrupamentos (por exemplo, similaridade média).

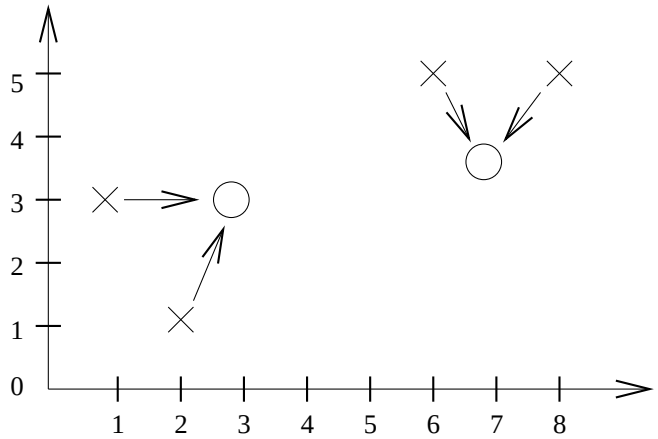
Algoritmo para Agrupamento - *K-means*



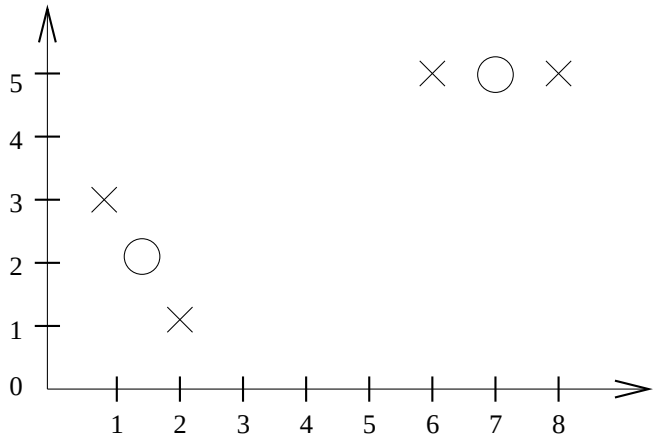
(1) Objetos que devem ser agrupados



(2) Sorteio dos pontos centrais dos agrupamentos

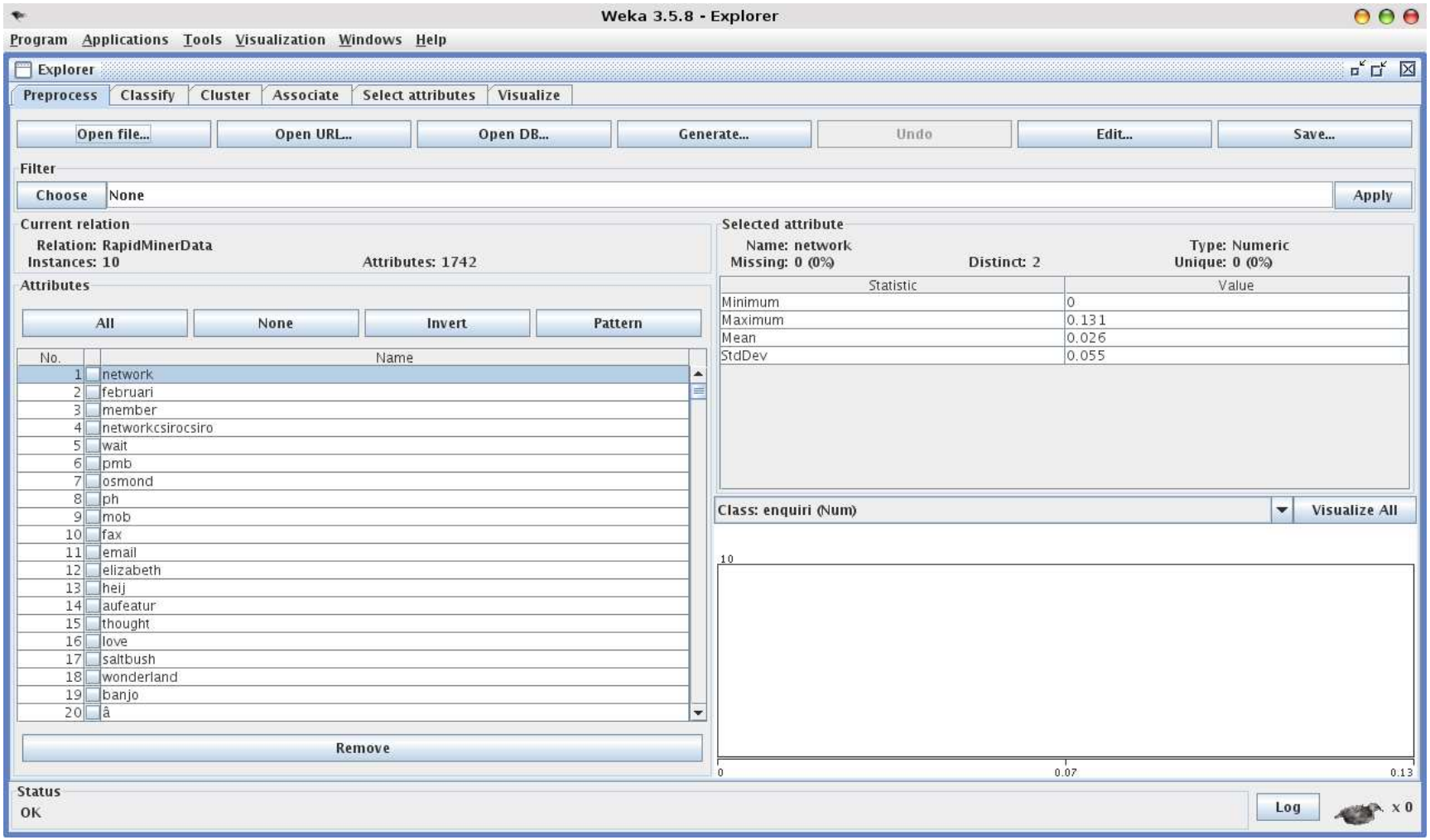


(3) Atribuição dos objetos aos agrupamentos



(4) Definição do centro do agrupamento

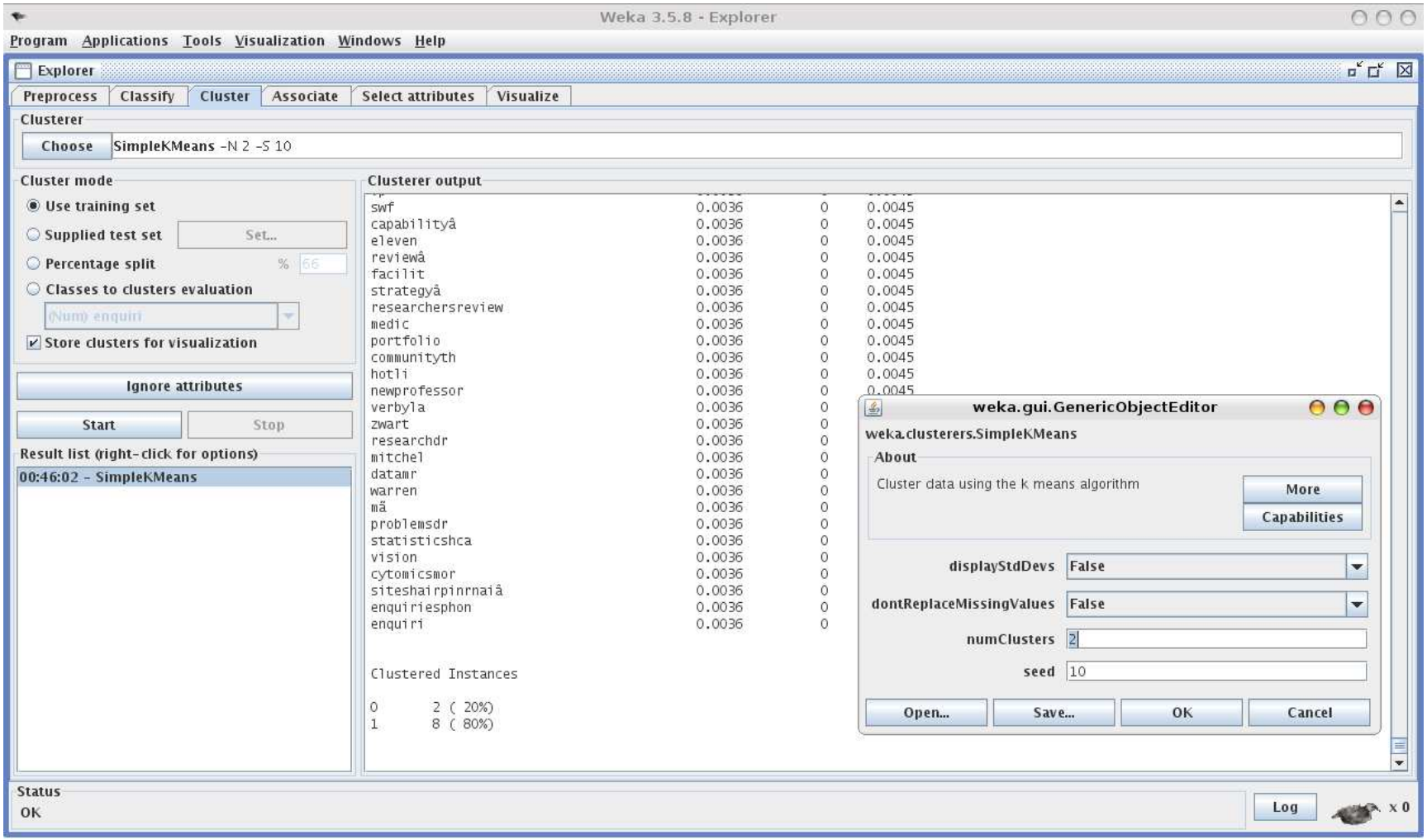
Algoritmo para agrupamento dos documentos - WEKA



Características e Funcionalidades do *Weka*

- Mais simples que o *RapidMiner*.
- Com menos funcionalidades.
- Os operadores podem ser divididos nas seguintes categorias:
  - Entrada e pré-processamento.
  - Classificação.
  - Agrupamento.
  - Associação.
  - Visualização.

Algoritmo para agrupamento dos documentos - WEKA



Algoritmo para agrupamento dos documentos - Código

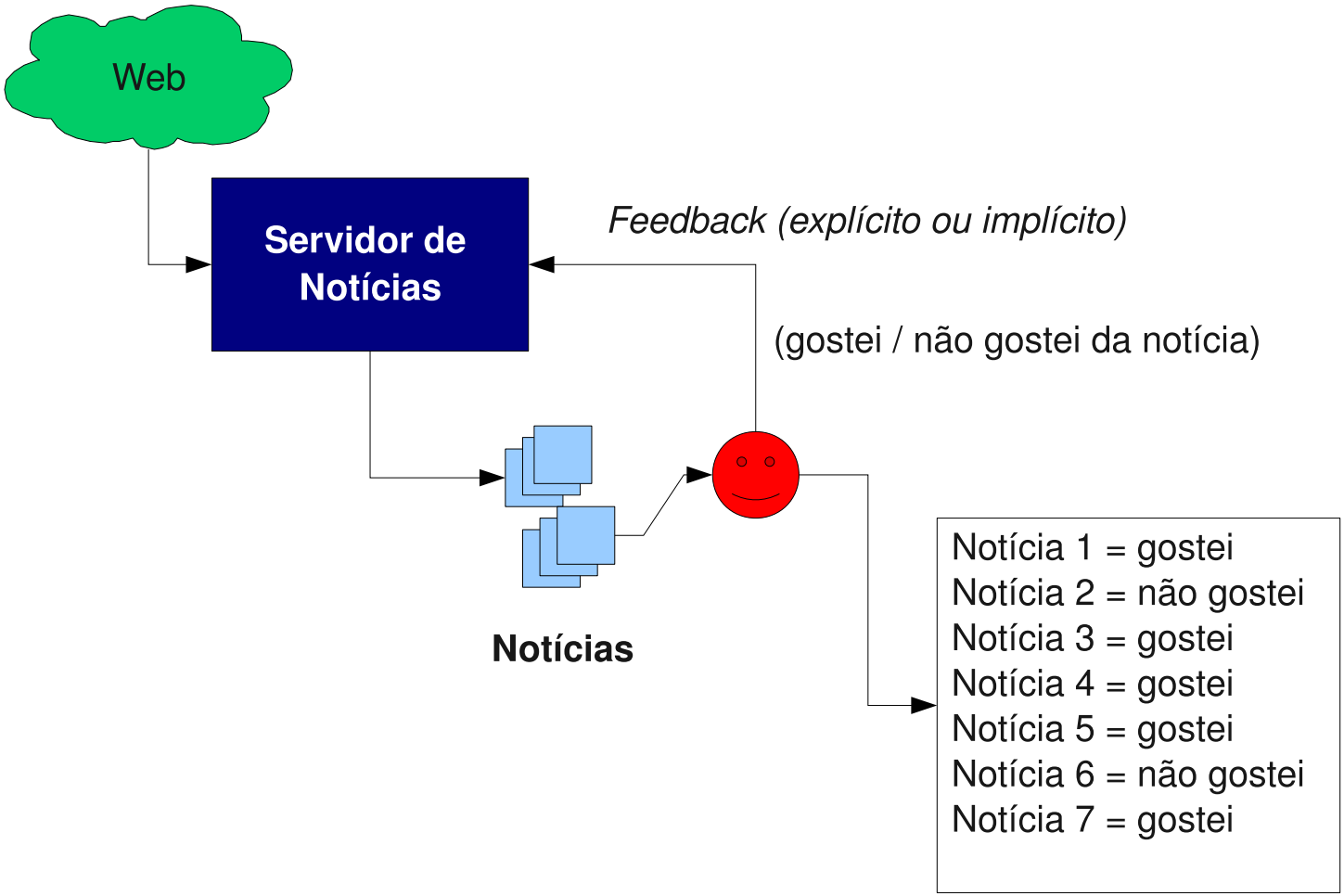
```
public AgrupamentoComKMeans(String arquivo){
    try{
        Instances instances = new Instances(new FileReader(arquivo));
        /*
         * Para visualizar os dados do arquivo arff
         */
        System.out.println("Dataset: ");
        System.out.println(instances);

        /*
         * Utilização do KMeans
         */
        SimpleKMeans kmeans = new SimpleKMeans();
        kmeans.buildClusterer(instances);
        /*
         * Imprimindo informação sobre instância - agrupamento
         */
        for(int i=0; i<instances.numInstances(); i++){
            System.out.println("A instância "+
                               instances.instance(i).toString()+
                               " estah no cluster "+
                               kmeans.clusterInstance(instances.instance(i)));
        }
    }catch(Exception e){
        System.out.println(e);
    }
}
```

Algoritmo para agrupamento dos documentos - Resultados

```
A instância 0.1,0.1,0.1,0.1,0.1 está no cluster 1
A instância 0.1,0.2,0.3,0.1,0.8 está no cluster 1
A instância 0.3,0.4,0.5,0.8,0.9 está no cluster 0
A instância 0.3,0.1,0.1,0.1,0.1 está no cluster 1
A instância 0.3,0.1,0.1,0.1,0.1 está no cluster 1
A instância 0.8,0.7,0.8,0.8,0.8 está no cluster 0
A instância 0.1,0.1,0.1,0.1,0.1 está no cluster 1
A instância 0.1,0.1,0.1,0.1,0.1 está no cluster 1
A instância 0.1,0.1,0.1,0.1,0.1 está no cluster 1
A instância 0.6,0.5,0.6,0.6,0.6 está no cluster 0
A instância 0.6,0.5,0.6,0.6,0.6 está no cluster 0
A instância 0.1,0.1,0.1,0.1,0.1 está no cluster 1
A instância 0.2,0.8,0.8,0.7,0.9 está no cluster 0
A instância 0.1,0.1,0.1,0.1,0.1 está no cluster 1
```

Classificação e Filtragem de Notícias

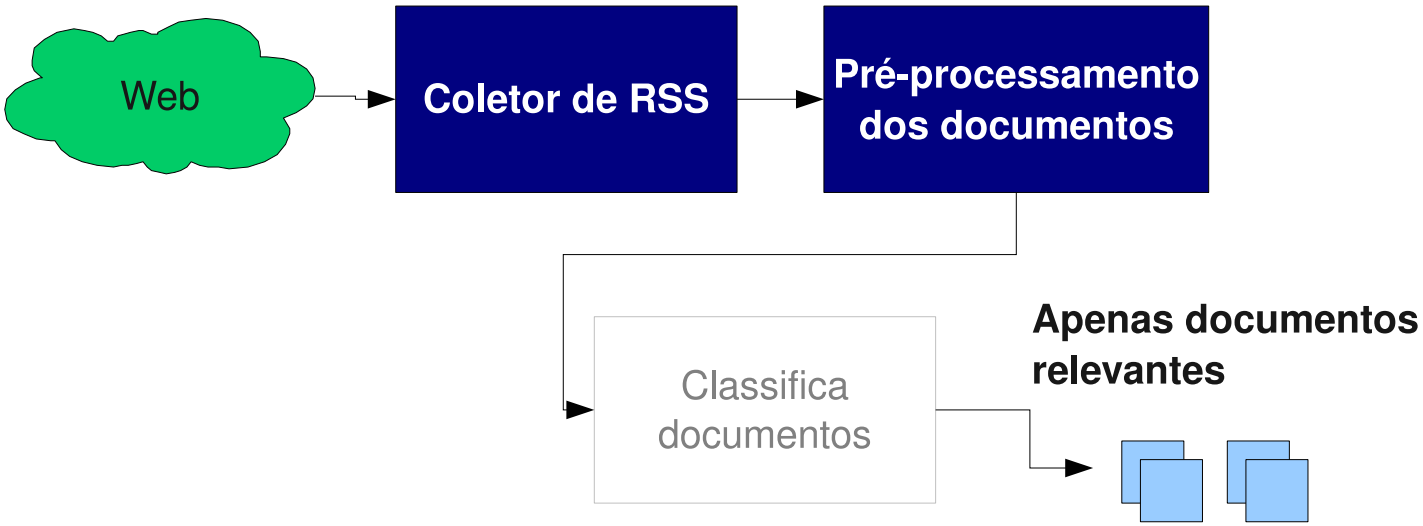
Classificação e Filtragem de Notícias



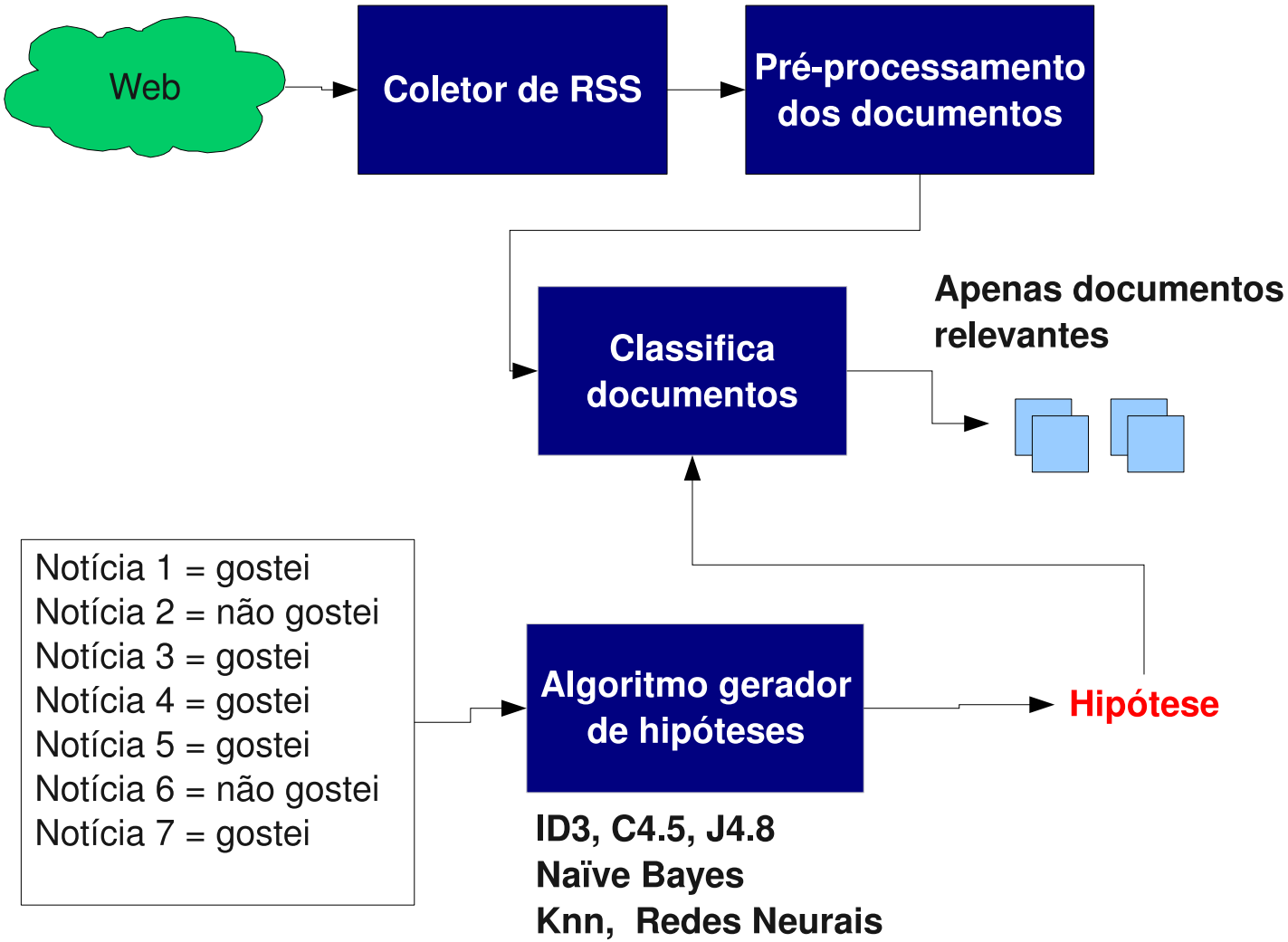
Conjunto de Exemplos - Atributo/Valor e Classe

Doc.	apresent	form	tecnic	caracteriz	...	Relevante
d_1	0.33	0.33	0.33	0.33	...	1
d_2	0	0.5	0.2	0.33	...	0
d_3	1	0.6	0	0	...	1
d_4	0.4	0.3	0.33	0.4	...	1
d_5	1	0.4	0.1	0.1	...	1
d_n	...	...	...	...	...	...

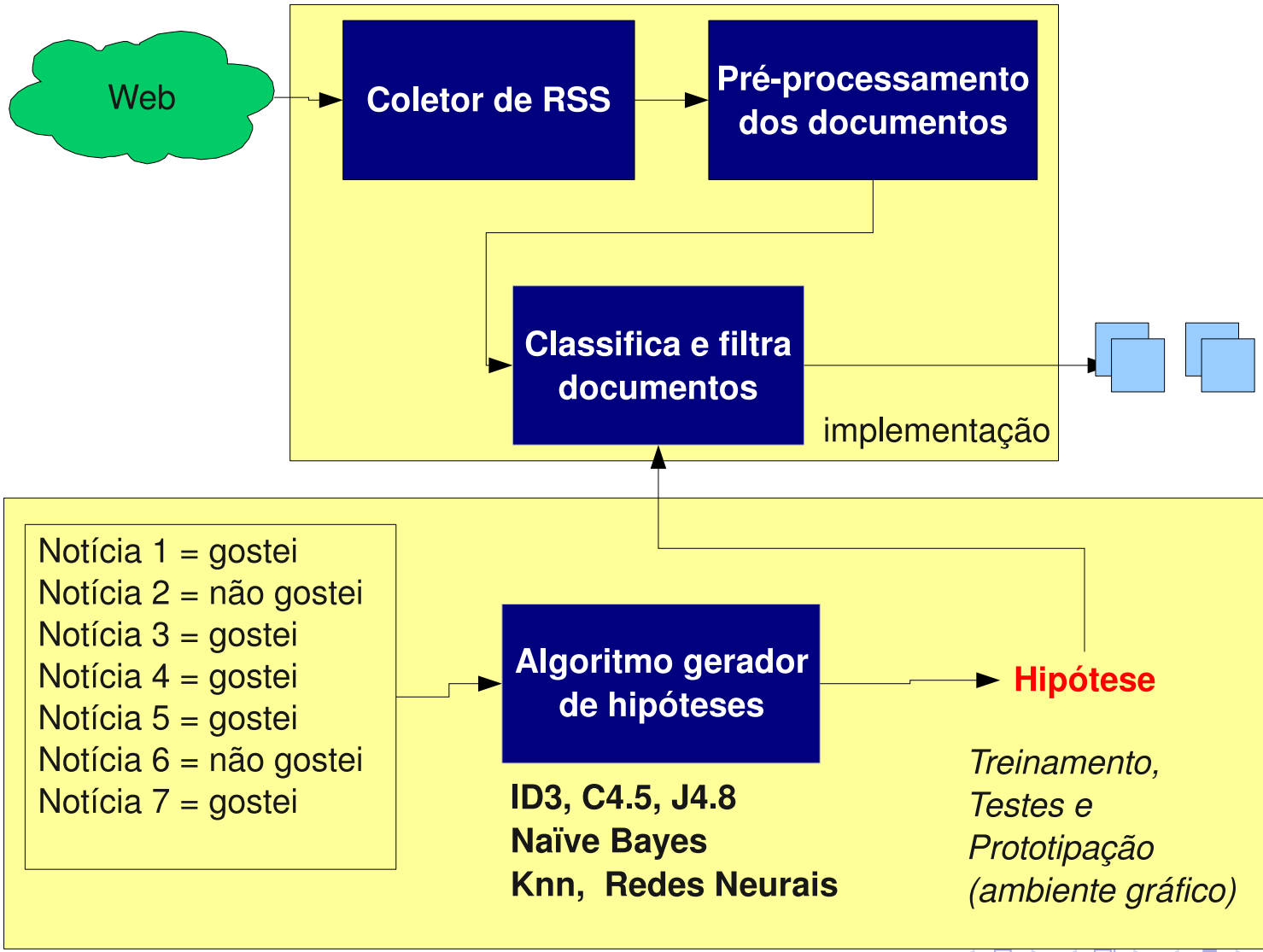
Qual é o problema?



Uma solução...


Web
Coletor de RSS
Pré-processamento dos documentos
Classifica documentos
Apenas documentos relevantes
Notícia 1 = gostei
Notícia 2 = não gostei
Notícia 3 = gostei
Notícia 4 = gostei
Notícia 5 = gostei
Notícia 6 = não gostei
Notícia 7 = gostei
Algoritmo gerador de hipóteses
Hipótese
ID3, C4.5, J4.8
Naïve Bayes
Knn, Redes Neurais

Processo de trabalho



Recuperação de Informação

- Como construir sistemas de *"busca"* sob medida?

Recuperação de Informação

- Como construir sistemas de *"busca"* sob medida?

- Lucene: biblioteca para Recuperação de Informação escrita em Java e código aberto. Possui: indexador e search engine.
- Outras funcionalidades (*web crawler* e *parsing* de páginas HTML) são implementados por outras ferramentas baseadas no Lucene, i.e, Nutch.
- Mantido pela *Apache Software Foundation*.

Recuperação de Informação

- Utilizando o Lucene, o desenvolvedor pode alterar a search engine para ter um mecanismo de busca apropriado para o domínio da sua aplicação.
- Exemplos de alterações possíveis:
  - Expansão da consulta usando uma taxonomia da empresa.
  - Alteração da função de ordenação (PageRank, BM25, Booleano).
  - Reordenação dos resultados utilizando informações do perfil do usuário.

Considerações

- Todas as fases de um sistema ou componente para tratamento de informações podem ser implementadas com as ferramentas vistas nesta apresentação:
  - Indexação.
  - Recuperação.
  - Mineração (determinação de padrões).
- Com o RapidMiner e Weka é possível:
  - Reutilizar diversos algoritmos necessários.
  - Prototipar (criar e validar) uma solução rapidamente.
  - Integrar a solução criada em outras aplicações.
- Com o Lucene é possível:
  - Desenvolver um mecanismo de *"busca"* sob medida.

Referências (1/2)

- Ian H. Witten, Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques (Second Edition), 2005.
- Weka 3: Data Mining Software in Java (http://www.cs.waikato.ac.nz/m̃l/weka/index.html).
- RapidMiner Community Edition (http://rapid-i.com/).
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. (http://www-csli.stanford.edu/h̃inrich/information-retrieval-book.html).
- Apache Lucene (http://lucene.apache.org/java/docs/).

Referências (2/2)

Extra: Processamento de Linguagem Natural

- GATE, A General Architecture for Text Engineering (http://gate.ac.uk/).
- UIMA - Unstructured Information Management Architecture (www.research.ibm.com/UIMA/).