



Introducción al procesamiento del lenguaje natural



Organización del curso

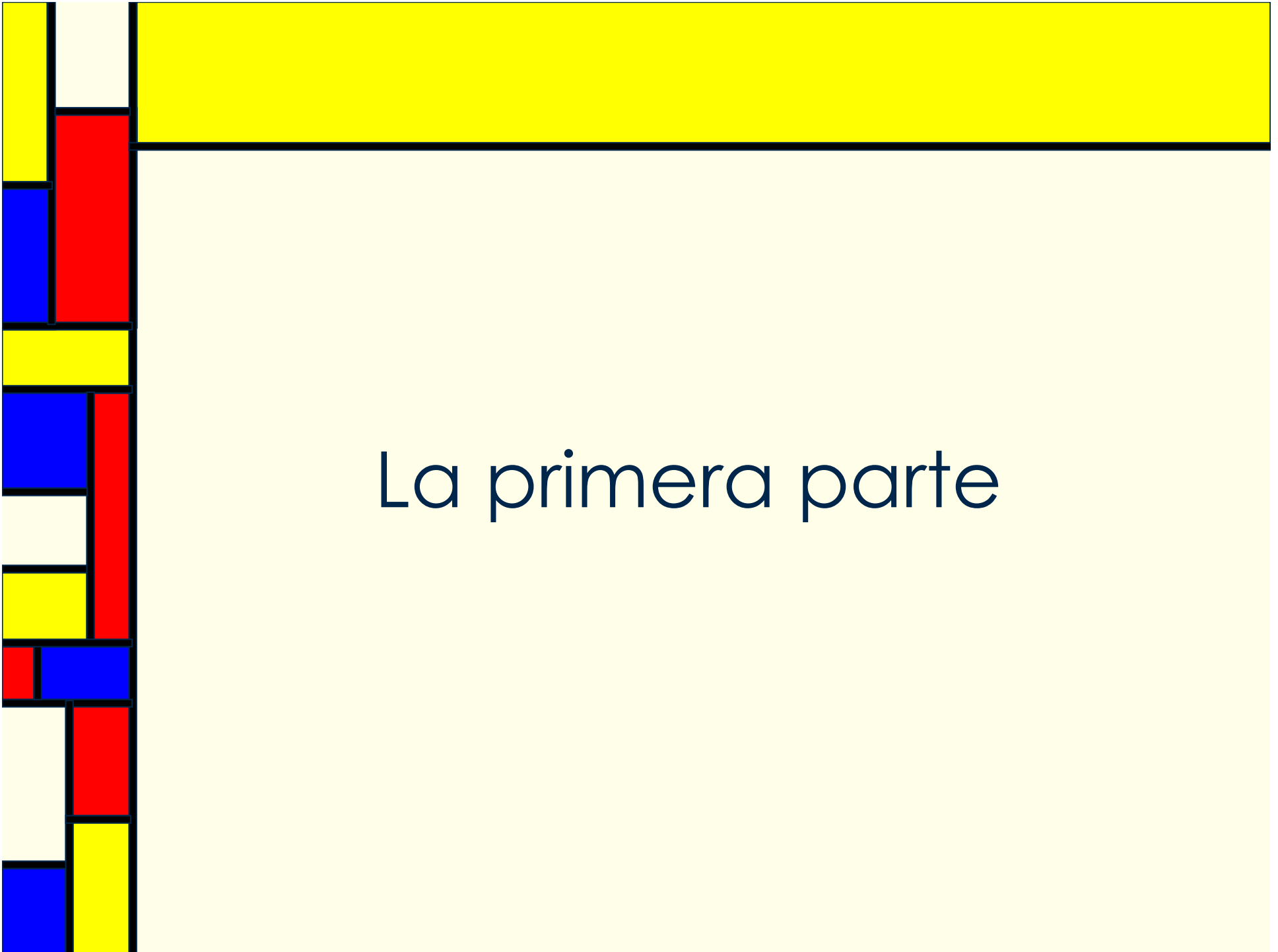
Parte I:

- José Antonio Troyano
- Procesamiento a nivel de documentos
- Clasificación automática de textos

Parte II:

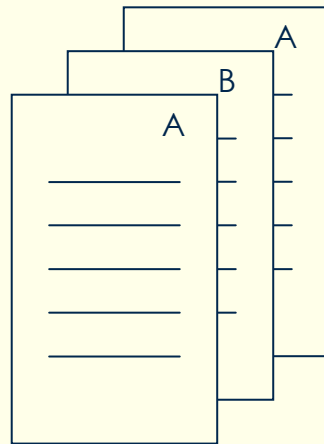
- Fernando Enríquez de Salamanca Ros
- Procesamiento a nivel de palabras
- Identificación de entidades con nombre

Material en: <http://www.lsi.us.es/>



La primera parte

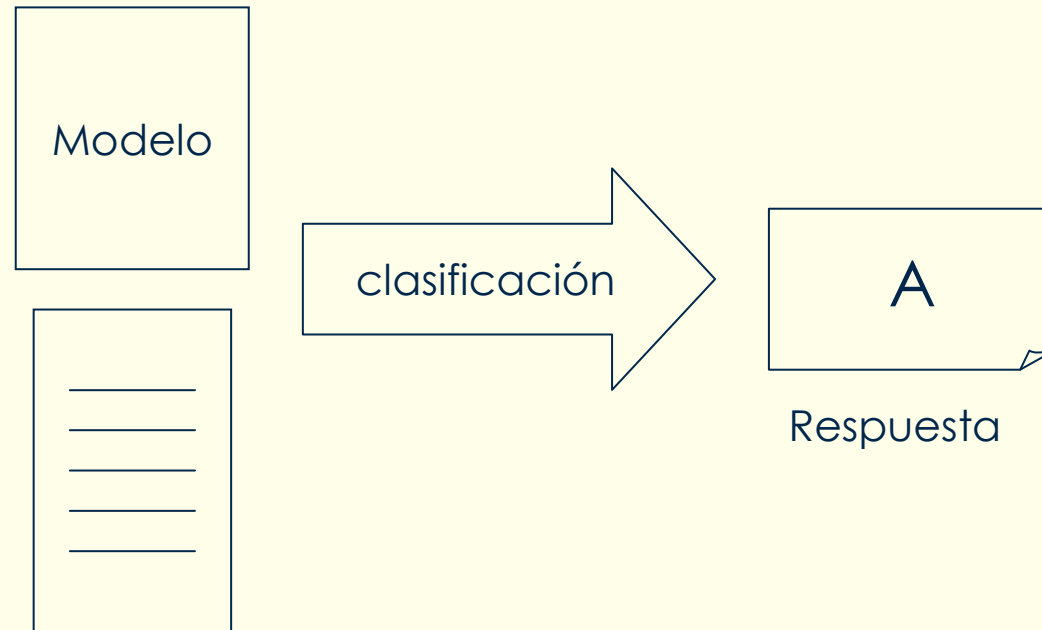
Clasificación de documentos



Corpus



Clasificación de documentos (II)



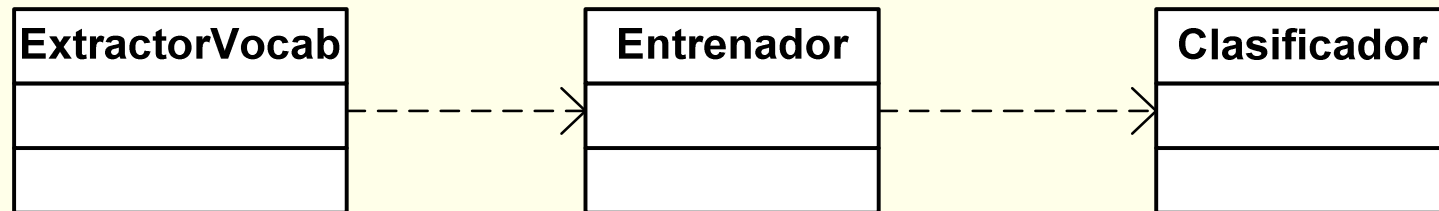
Sesiones previstas para la parte I

1. Colecciones de documentos (Corpus, Documento)
2. Identificación de palabras (Tokenizador)
3. Cálculo de frecuencias (Diccionario, ExtractorVocab)
4. Modelo vectorial, índice TF/IDF (VectorPesos, Entrenador)
5. Vecinos más cercanos, centroides (VectorRocchio)
6. Distancias y clasificación (Clasificador)
7. Precisión, cobertura y medida F (Evaluador)
8. ¿Cómo funciona Google?

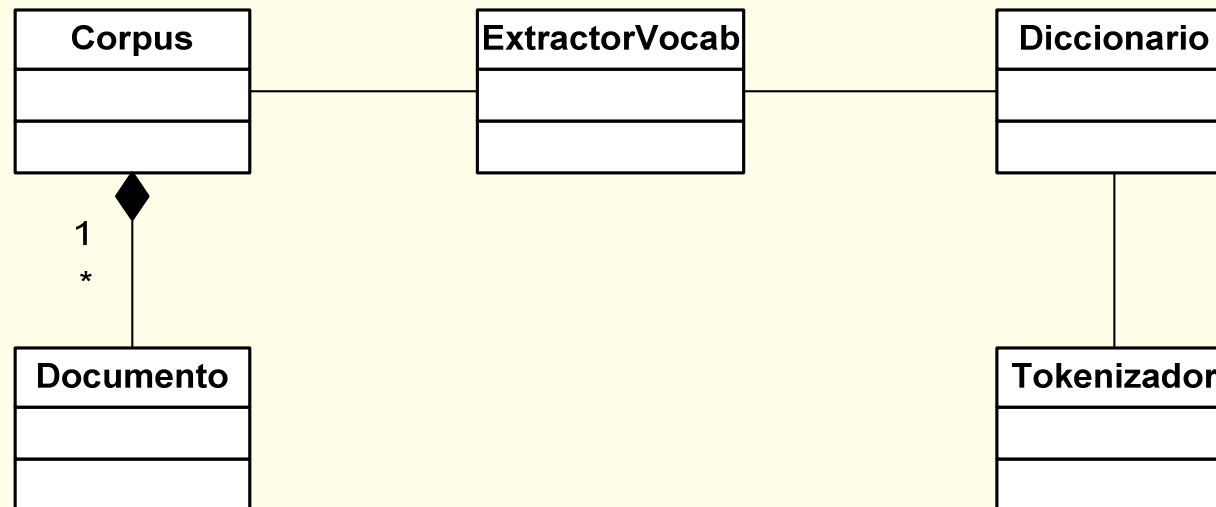
Referencias (disponibles en internet):

- Machine Learning in Automated Text Categorization (Fabrizio Sebastiani)
- Text Categorization and Prototypes (Alexander Bergo)

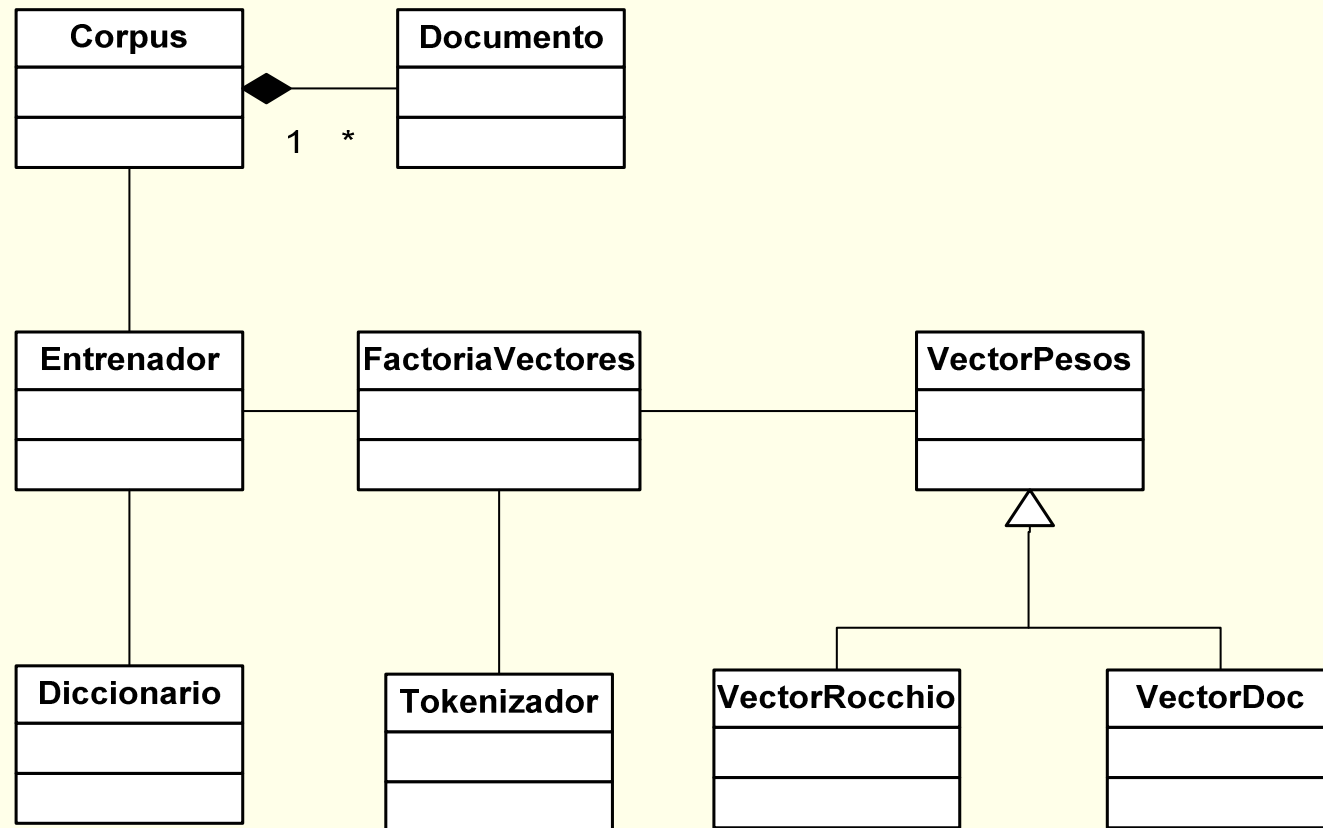
Clasificación: Diagrama principal



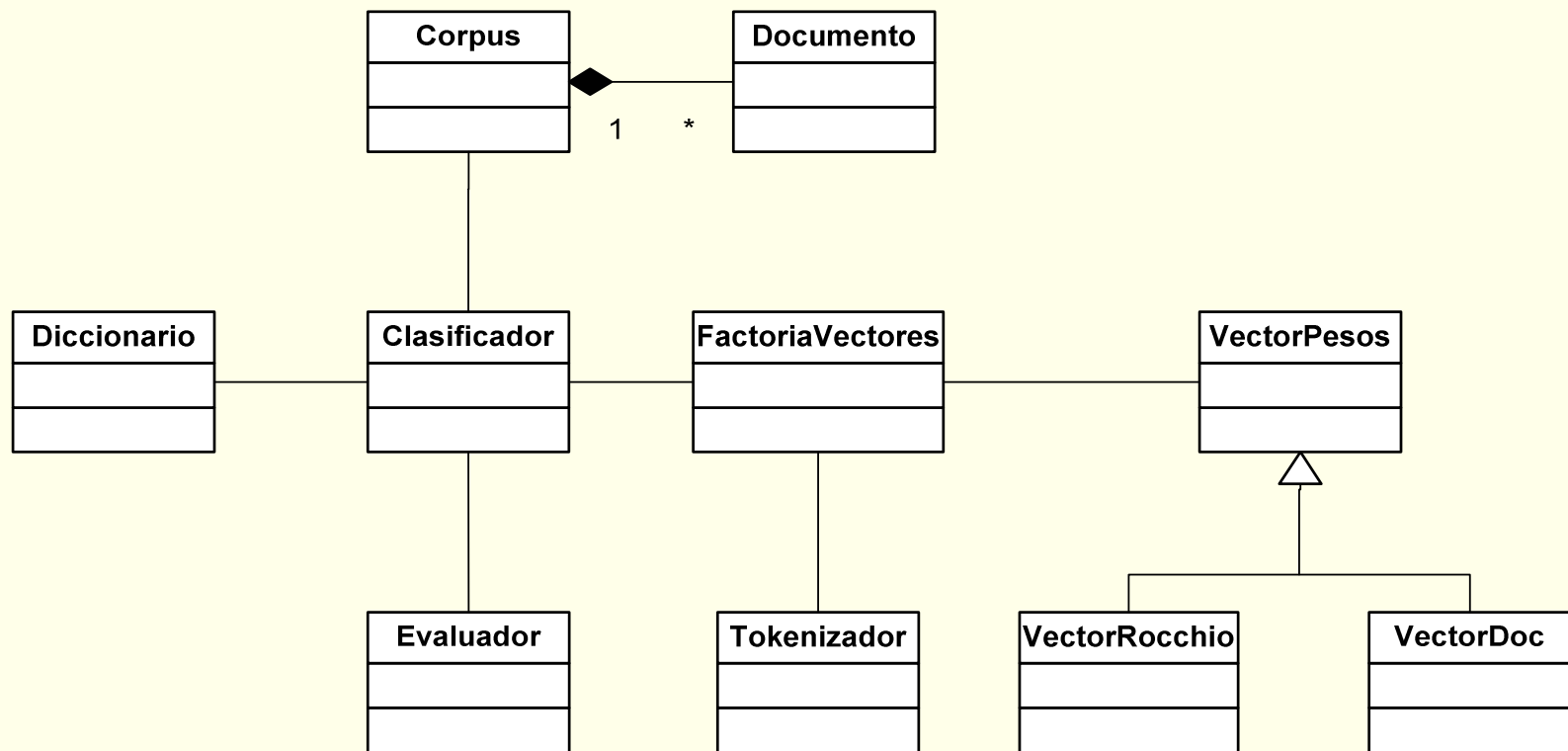
Extracción de vocabulario

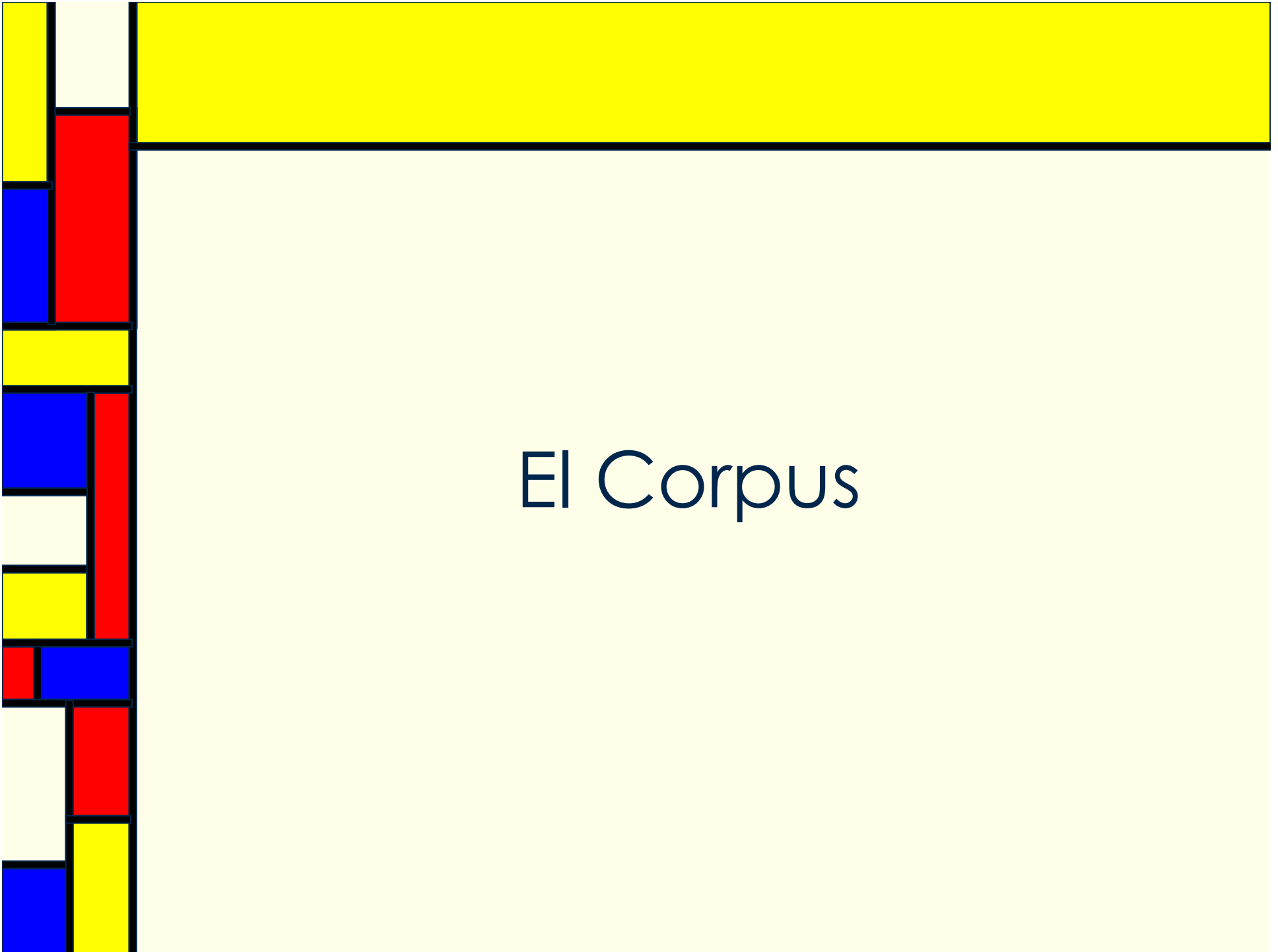


Entrenamiento del clasificador



Clasificación y cálculo de resultados





El Corpus

Un ejemplo

```
<corpus>
```

```
<documento>
```

```
<id> 1001 </id>
```

```
<clase> deporte </clase>
```

```
<texto>
```

La Agencia Tributaria ha decretado el embargo preventivo de los ingresos televisivos y publicitarios del Atlético de Madrid y 7 equipos de Segunda, entre ellos el recién ascendido Albacete. La cantidad que se embargaría asciende a 43,5 millones de euros debido al "riesgo de impago" de la deuda que tienen contraídos dichos clubes con el Ministerio de Hacienda.

```
</texto>
```

```
</documento>
```

```
<documento>
```

```
<id> 1005 </id>
```

```
<clase> nacional </clase>
```

```
<texto>
```

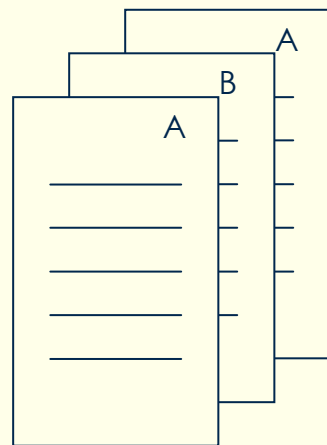
El presidente de la Generalitat, Jordi Pujol, ha asegurado que el pasado sábado, fecha en la que se constituyeron los ayuntamientos tras las municipales del 25-M, obligó al candidato de CiU de El Vendrell a no presentarse para evitar ser reelegido con el voto de la plataforma xenófoba.

```
</texto>
```

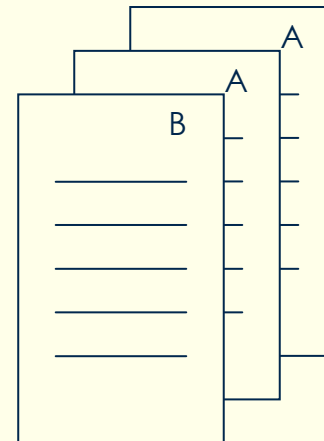
```
</documento>
```

```
</corpus>
```

Hacen falta dos



Corpus de
entrenamiento



Corpus de test

Manejo de documentos XML

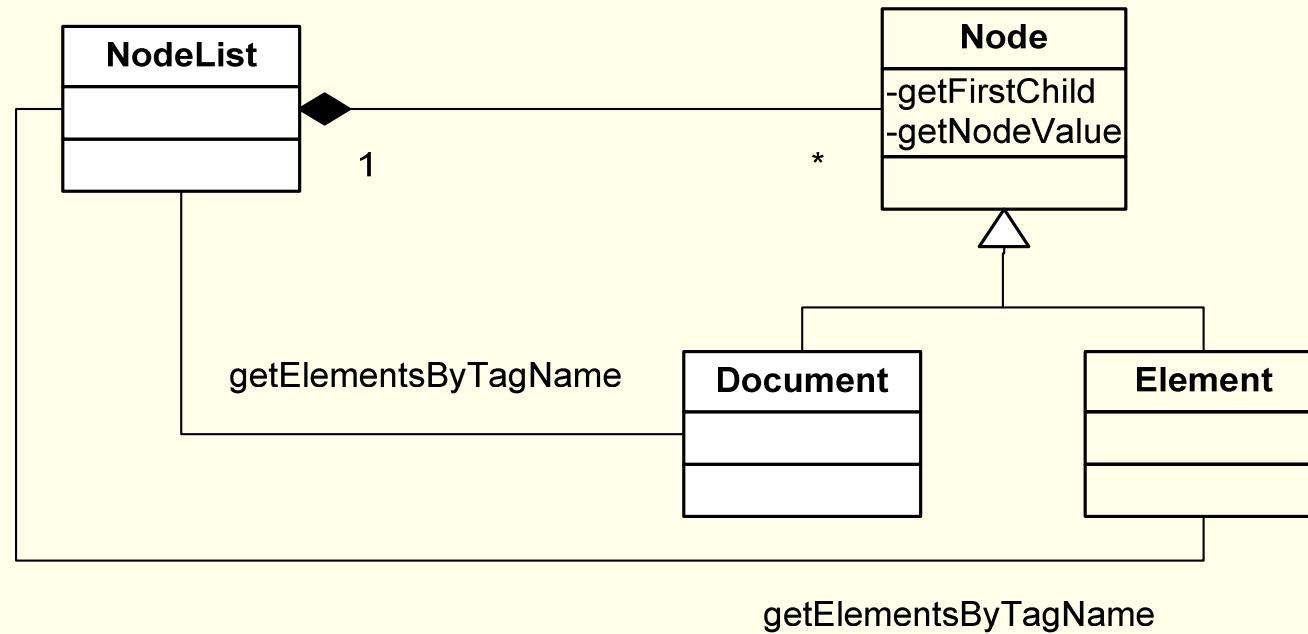
- javax.xml.parsers.*: Clases abstractas para parsers SAX y DOM
- org.xml.sax.*: conexión entre fichero de entrada y parser SAX

```
DocumentBuilderFactory factory =
DocumentBuilderFactory.newInstance();
DocumentBuilder builder = factory.newDocumentBuilder();
// Para los caracteres acentuados
// hay que usar InputSource y FileReader en el parsing
Document arbolDOM = builder.parse(new InputSource(
                                new FileReader(
                                new File(fichero))));
```

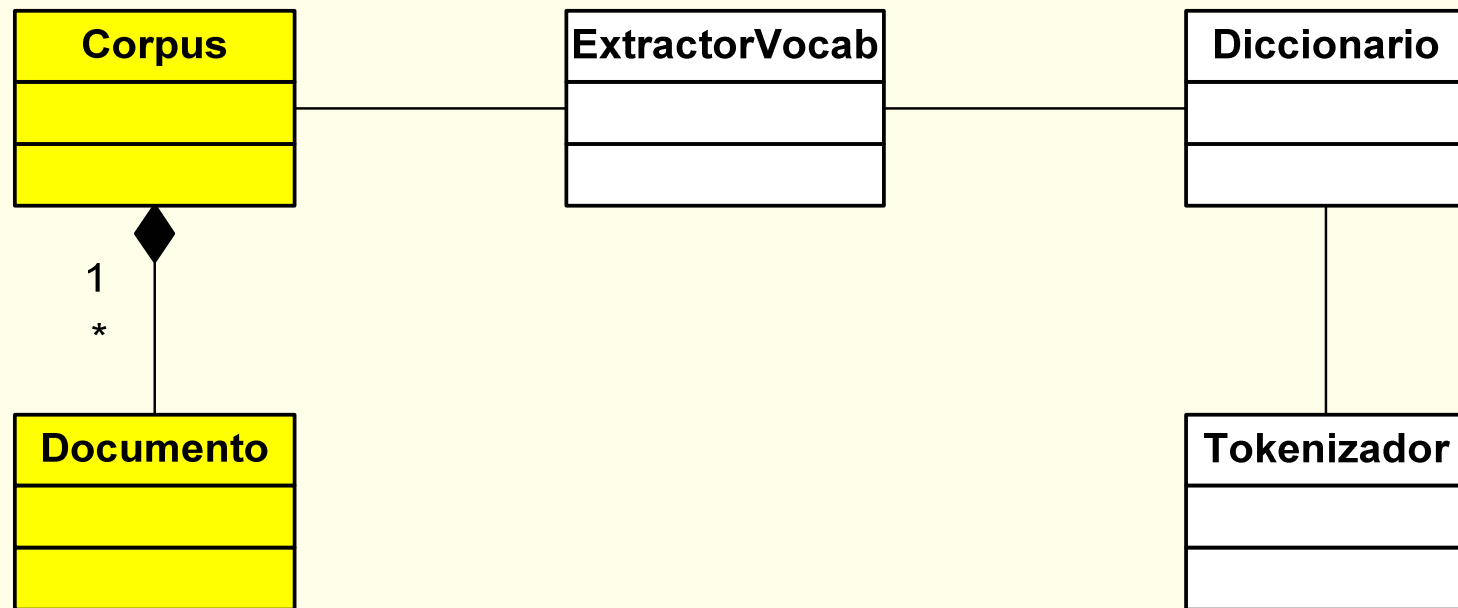
- org.w3c.dom: Interfaces para el manejo del árbol DOM

```
Element
Document
Node
NodeList
```

Interfaces para árbol DOM



Clases a desarrollar



La clase Corpus

```
import java.io.*;
import java.util.*;
import javax.xml.parsers.*;
import org.xml.sax.*;
import org.w3c.dom.*;

/*
 * Número aproximado de líneas de código: 60
 */
public class Corpus {
    ...
    public Corpus(String fichero) {...}
    public int numeroDocumentos(){...}
    public Set categorias(){...}
    public boolean hayMasDocumentos(){...}
    public Documento siguienteDocumento(){...}
    ...
}
```

Más información en: [Corpus.html](#)

La clase Documento

```
/*  
 * Número aproximado de líneas de código: 20  
 */  
public class Documento {  
    public Documento(String ident, String categoría,  
                    String texto) {...}  
    public String ident(){...}  
    public String categoría(){...}  
    public String texto(){...}  
}
```

Más información en: [Documento.html](#)