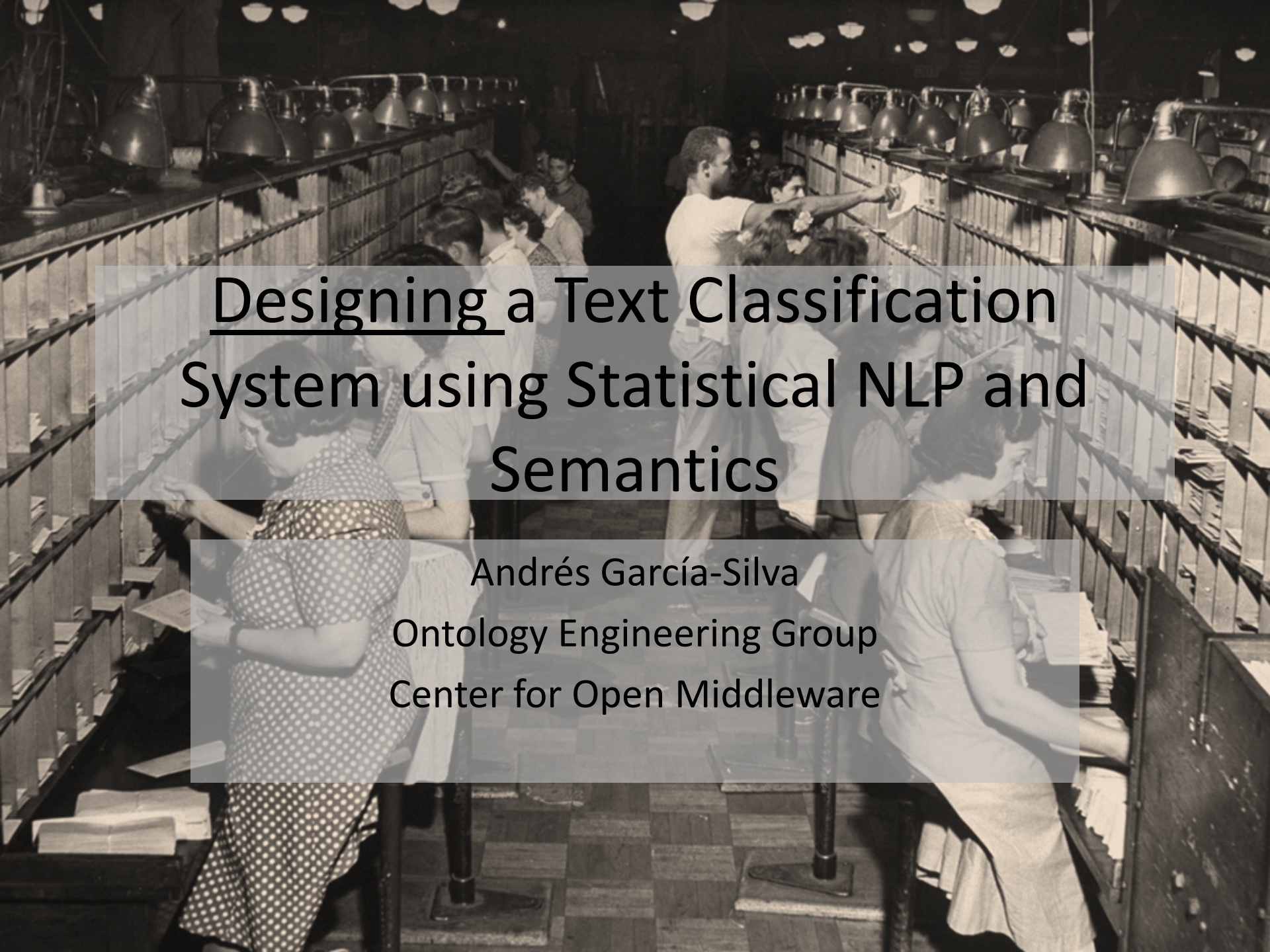


Designing a Text Classification System using Statistical NLP and Semantics

Andrés García-Silva
Ontology Engineering Group
Center for Open Middleware

Overview

- Introduction
- Text Classification
- Machine Learning approach
  - Generate Training Set
  - Document Representation
  - Learning Method
  - Evaluation
- Architecture

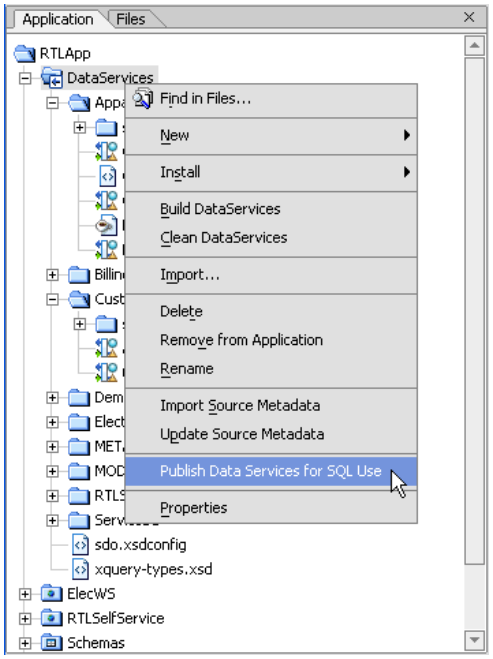
Introduction

Long training period

Difficult to keep up to date with new functionalities

Introduction

Quiero ver la información de los productos y las aplicaciones instalados en la producción del banco



Problem

- To identify the subset of system queries that might answer a user question written in Spanish.

Introduction

Problem Definition

S=(Q, U, R)

Q = System queries, U=User questions,

R = U x Q, relates user questions with system queries

We need to find:

$f(u \in U)=\{q \in Q : (u , q) \in R\}$

If we think of system queries as categories we can address this problem as a text classification problem

<Cuál es la información de los prod.., ?c>

Approaches for text classification

Rule-based

if (u.match(^<InterrogativePronoun> word^{+} <SystemObject>))
return "MapaFuncional"

Rules are written by specialists (e.g. Computational linguistic) in collaboration with Domain experts

Advantages

- Highly Precise
- Rules can be interpreted
- There exists some tools allowing rule processing:

GATE general architecture for text engineering

Disadvantages

- Does not scale
- Knowledge acquisition:
  - Writing rules is a tough work
  - Time-consuming task
- Managing rules is not a trivial task
- Does not have ability to learn

Approaches for text classification

Supervised Learning

- Our goal is to find $f: U \rightarrow C$
- We use a learning method T that takes as input a training set $D \subseteq R$ and returns the learned classifier f

Advantages

- Human resources are not needed to make rules
- Ability to adapt to a changing environment
- Many available learning methods and tools

Disadvantages

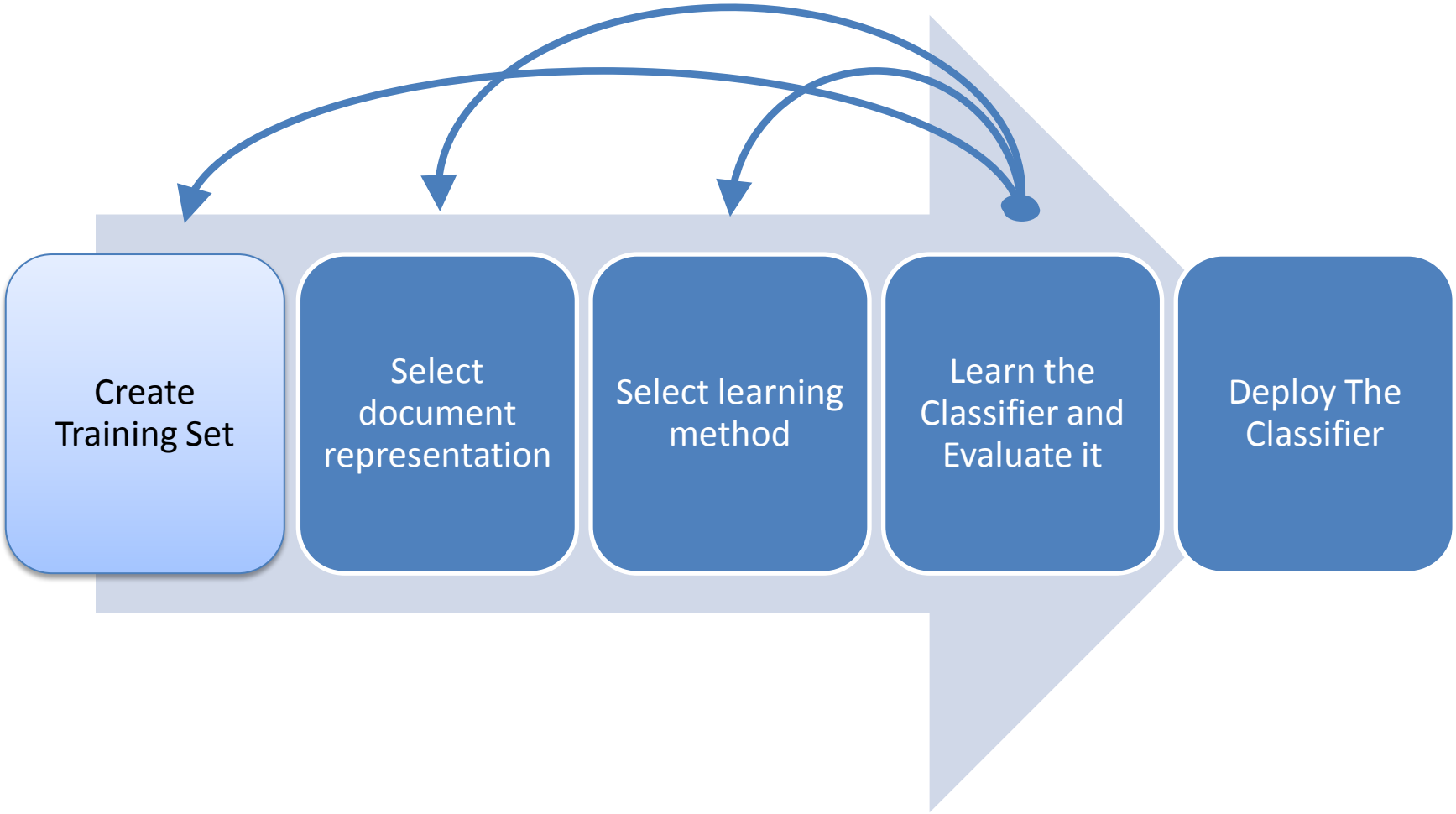
- Needs annotated data (Training data)
- Classifier performance depends on the training data quality
- Classification rules are not always easily interpreted.

WEKA The University of Waikato

GATE general architecture for text engineering

openNLP™

Machine Learning Approach



Training Set

Labeled dataset:

$$TrainingSet=\{<u_i, c_k>: <u_i, c_k> \epsilon R\}$$

<"Cuáles son todas las versiones que se ha instalado de el objeto OBJETO", HistObj>

– Similar distribution that unseen data.

Domain experts are involved
However they only provide a few examples ~12 questions per class

With the help of a NLP expert we identify patterns from the original questions and generate new questions

Training Set

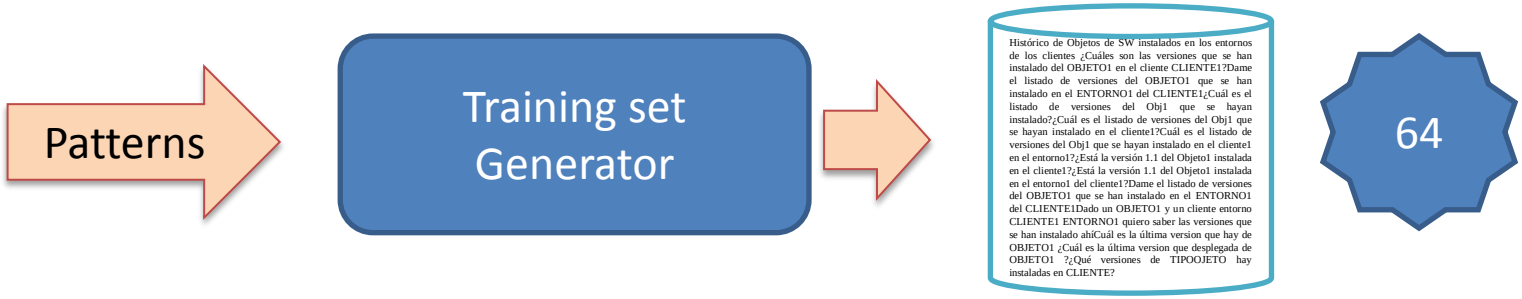
User Question:

<“Cuáles son todas las versiones que se ha instalado de el objeto OBJETO”, HistObj>

Pattern:

“Cuáles son [todas] las versiones [<que se ha instalado | que han instalado| que están | que están instaladas | que hay | que hay instaladas | que tenemos instaladas | que tenemos | que se tienen | que se tienen instaladas | que hemos instalado | que hayamos instalado | que se haya instalado| que se hayan instalado | instaladas>] de [el objeto] OBJETO”

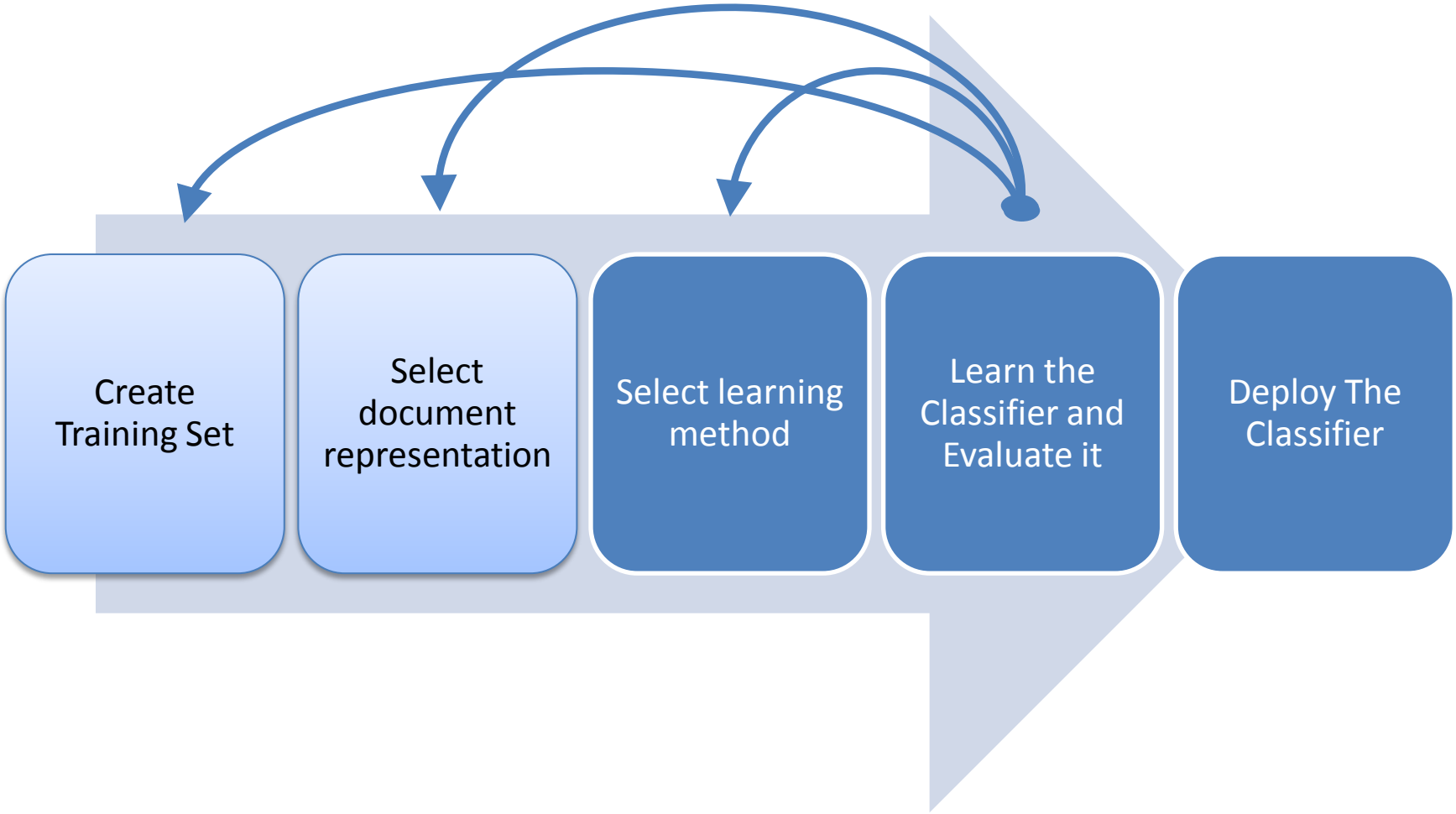
Notation: [optional] and <alternative | alternative >



Training Set

- For the first system version we included 7 system queries (classes) and generated **264K** user questions

Machine Learning Approach



Document Representation

Machine Learning methods require documents to be represented using feature vectors “d”

U: document space (the set o user queries)

V: term vocabulary in U

Bag of words: $d \in \mathbb{R}^n$ where $n = |V|$

- Assumption: Order is not important

- Decisions to take:
  - **Accents and diacritics**
  - **Case-folding**
  - **Stemming**
  - **Lemmatize**
  - **Stop words**

Document Representation

Bag of Words

- What about compound nouns or names?
  - entorno de cliente, agrupación técnica
  - Should we add n-grams to the bag?

- And named entities?
  - Modulo FACT01

- What about comparison structures?
  - entre Banco Santander y el BBVA?

NLP Tasks

Part-Of-Speech Tags

NER and NEC

Chunker

Document Representation

- NLP Toolkits with Spanish support

Freeling is written in C++ but the library can be used in Java by wrapping the API using (Simplified Wrapper and Interface Generator) **SWIG**

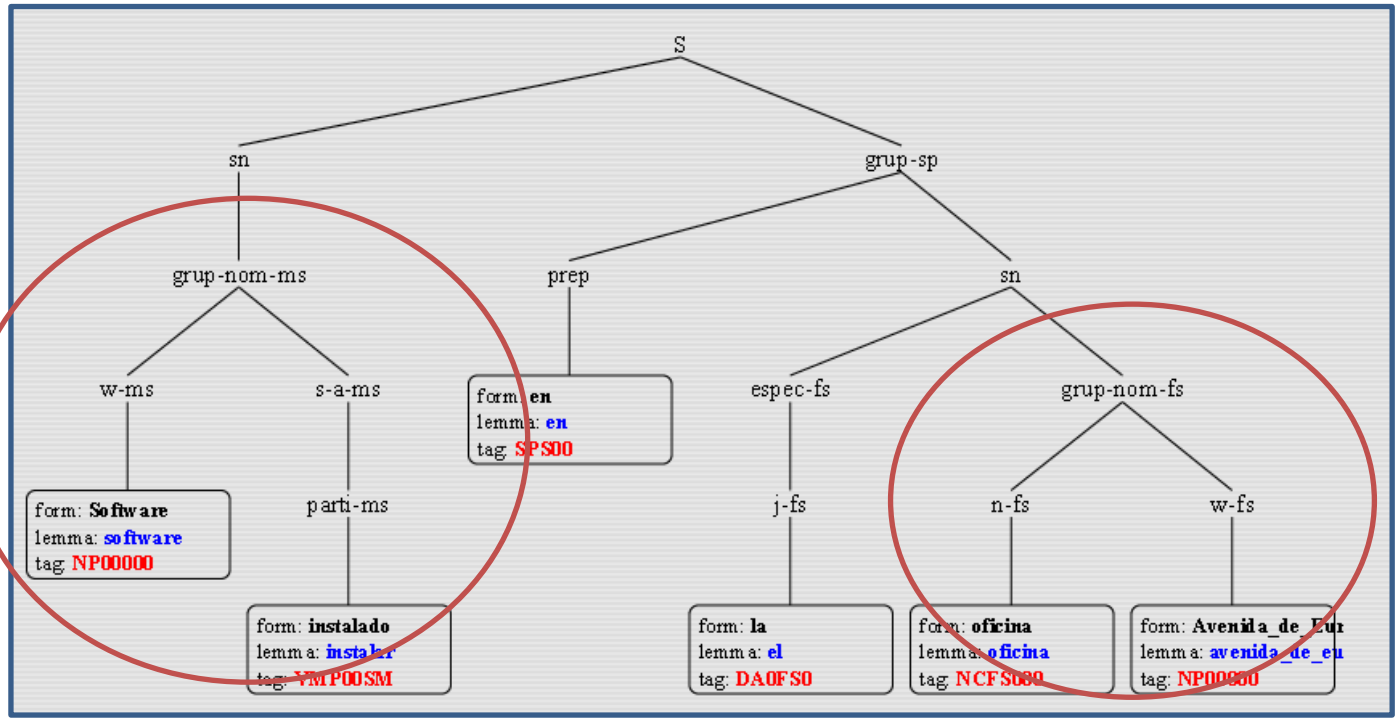
Document Representation

Freeling

POS-tagging & Named Entity detection

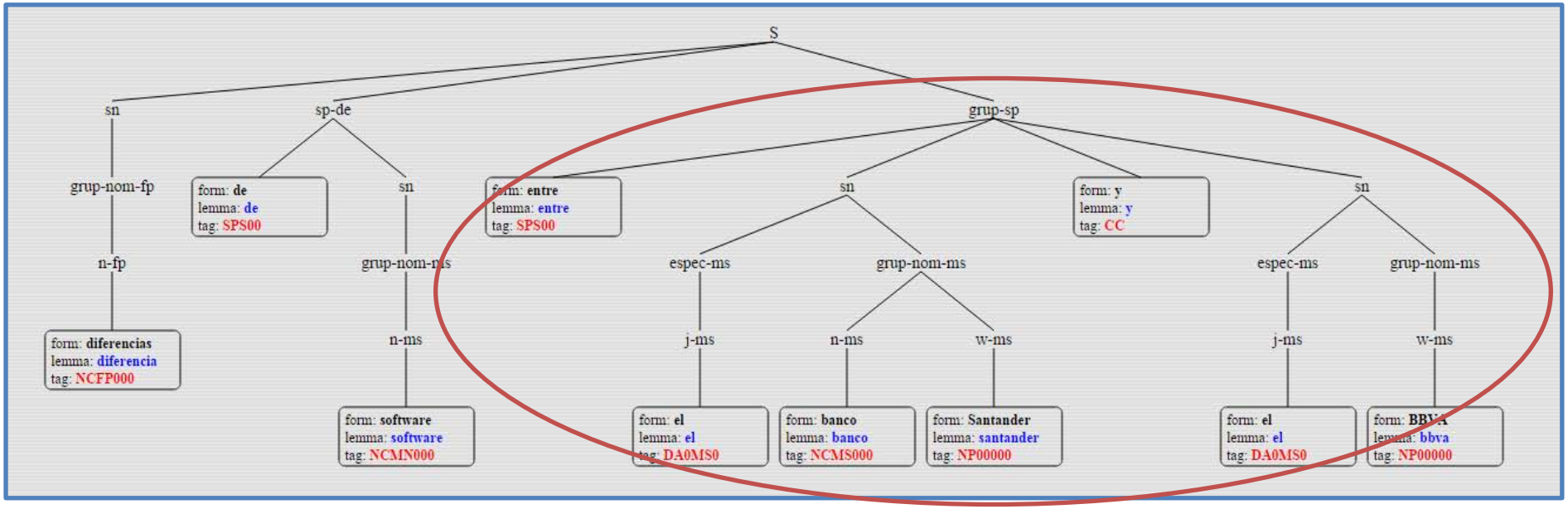
Software	instalado	en	la	oficina	Avenida_de_Europa
software	*instalar*	*en*	*el*	*oficina*	*avenida_de_europa*
NP00000	VMP00SM	SPS00	DA0FS0	NCFS000	NP00000

Shallow Parsing



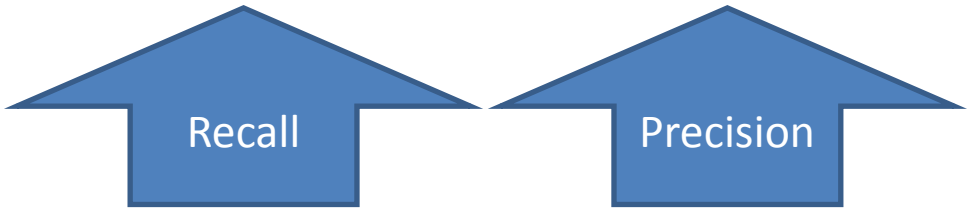
Document Representation

Freeling Shallow Parsing



Document Representation

- Semantic Features
  - Entity Types
    - Localización, Aplicación, Entorno
  - Hierarchical relations
    - Servicios -> Aplicaciones -> Subsistemas -> Sistemas -> Capas



Document Representation

- In summary we use:
  - Tokens
  - Named Entities: Freeling to spot candidates but not to classify
  - Chunks
    - grup-nom: noun, nominal chunk
    - sp-de: preposition, prepositional phrase “de”
    - grup-sp: preposition, prepositional chunk

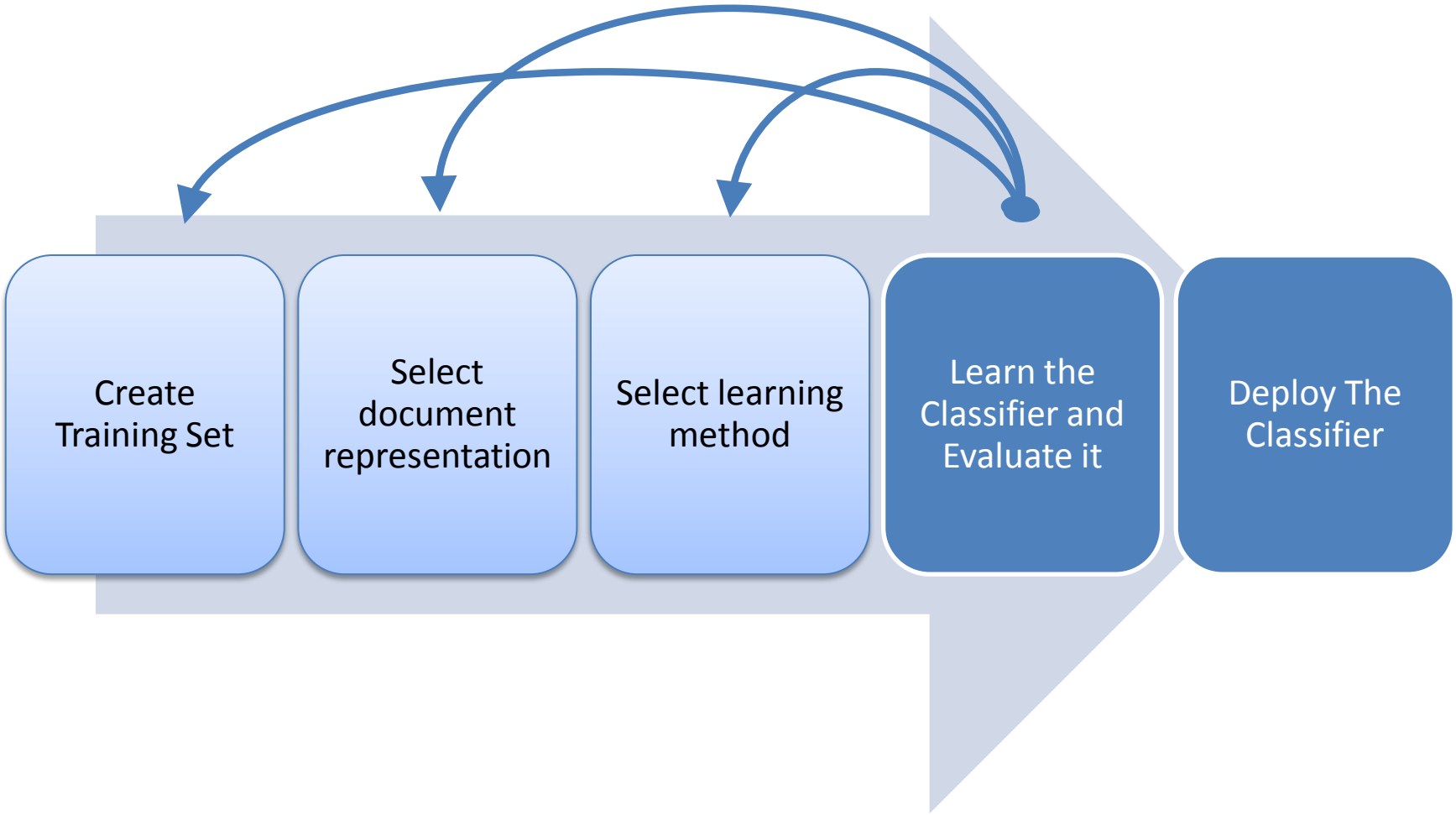
```
[token] objeto [162193][NCMS000]
[token] instalar[160024][VMP00SM]
[token] versión[147360][NCFS000]
[token] cliente[136134][NCCS000]
[token] haber[105036][VAIP3S0]
[token] entorno_de_cliente [26694][NP00000]
```

```
[chunk]versión instalar[1820]
[chunk]cliente distinto[1806]
[chunk]entorno distinto[1806]
[chunk]en entorno[80375]
[chunk]en el cliente CLIENTE[49381]
[chunk]de el objeto OBJETO[37384]
```

Document Representation

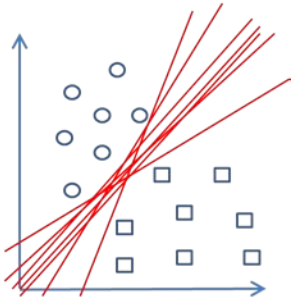
- Weighting Scheme
  - Term presence/absence
  - Term Frequency
  - Document Frequency
  - TF-IDF
  - Learn the weights?

Machine Learning Approach

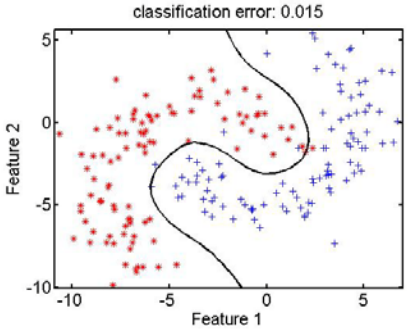


Learning Method

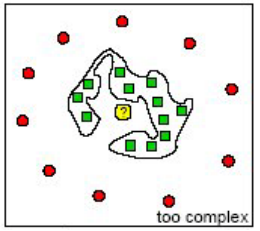
$$\vec{w}^T\vec{x} = b$$

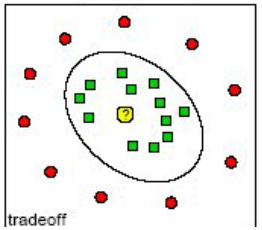


- **Linear**
  - Decides class membership by comparing a linear combination of the features to a threshold
  - E.g., Naive Bayes, Logistic regression, Support vector machine SVM.

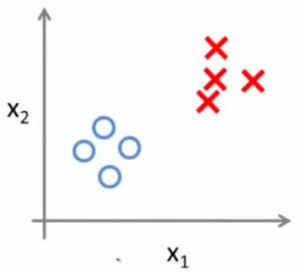


- **Non-linear**
  - More versatile
  - Might suffer of overfitting: Fails to Generalize
  - E.g., K nearest neighbor, SVM with kernel trick.

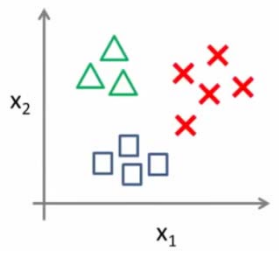




Learning Method



Binary Classification

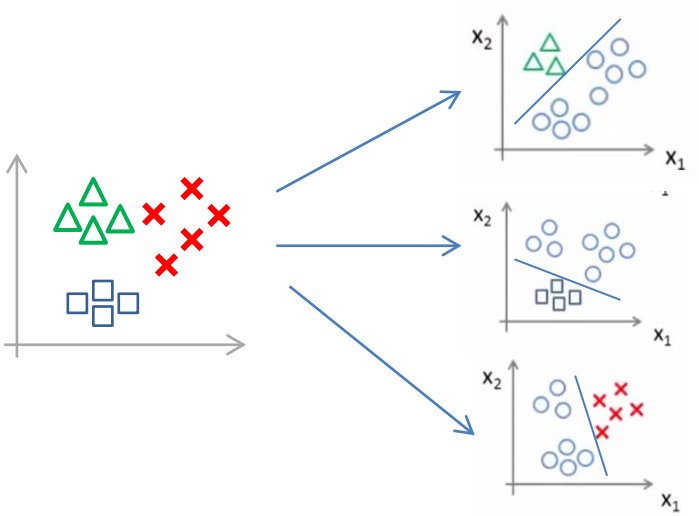


Multiclass Classification

In our system each query is a class

One vs All (One vs Rest)

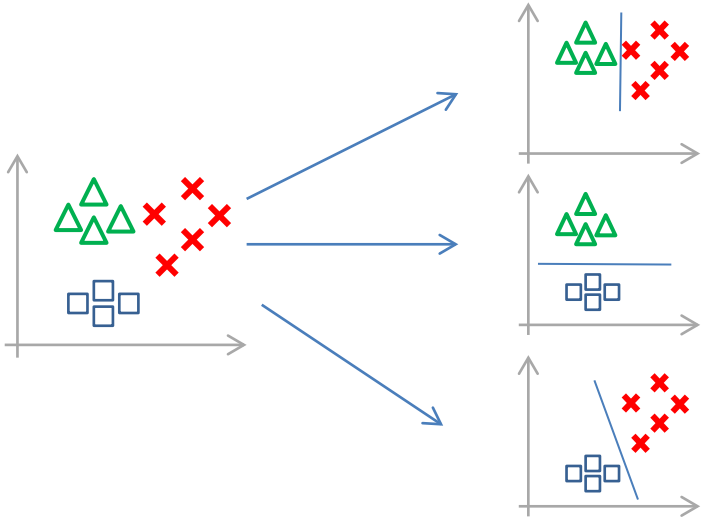
n classifiers



Pick the classifier with the highest confidence score

All vs All (one vs one)

n(n-1)/2 classifiers



For each class aggregate confidence scores and select the class with the maximum

Learning Method

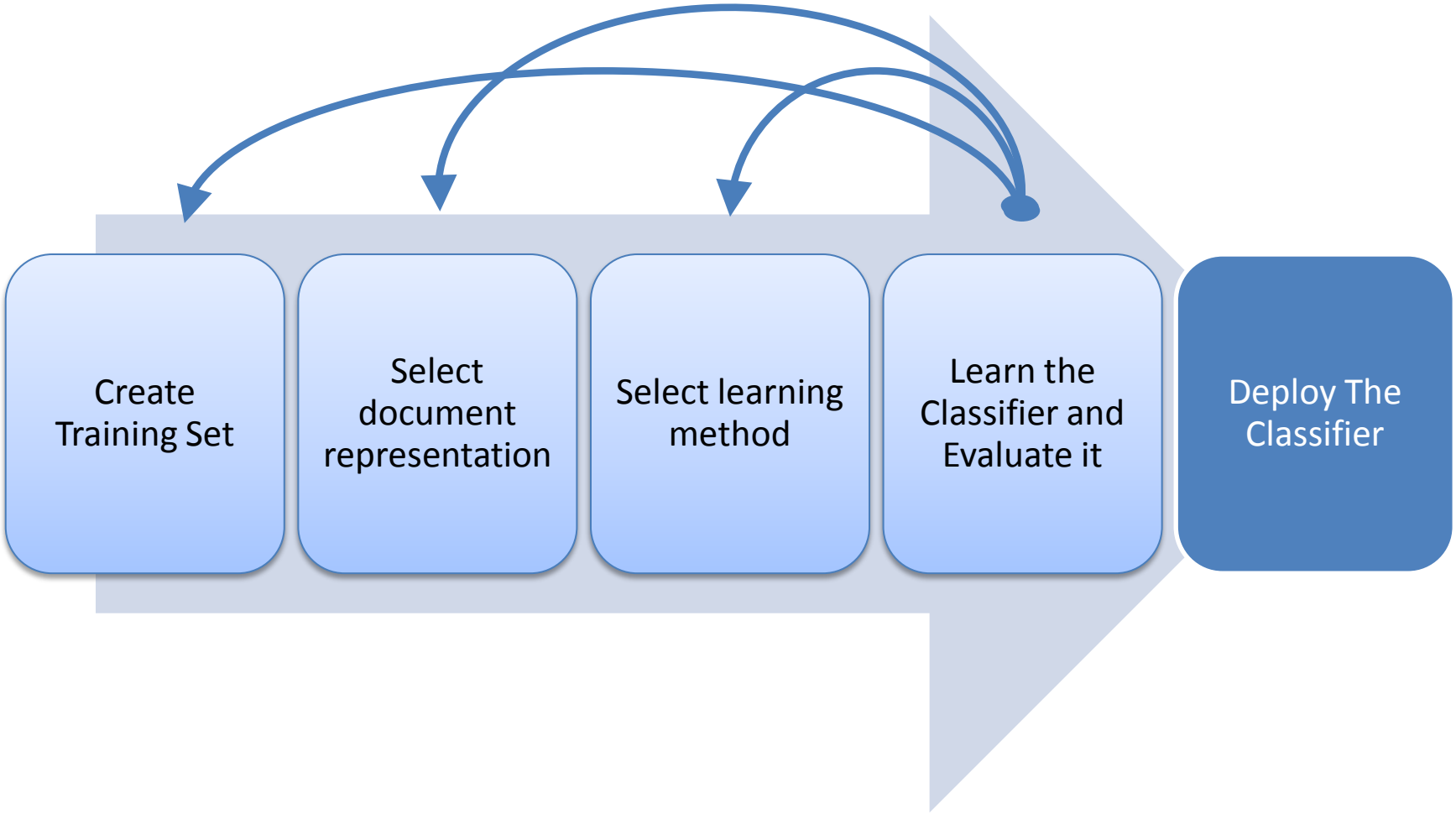
- **How to select one?**
  - Are the classes linearly separable?
  - Use well known linear classifiers and check the error rate

- **Selected Learning Method:**
  - Naive Bayes is used as Baseline
  - Support Vector Machine
    - Can produce linear and non-linear classifiers
    - There's a method to select the parameters

- **Library**

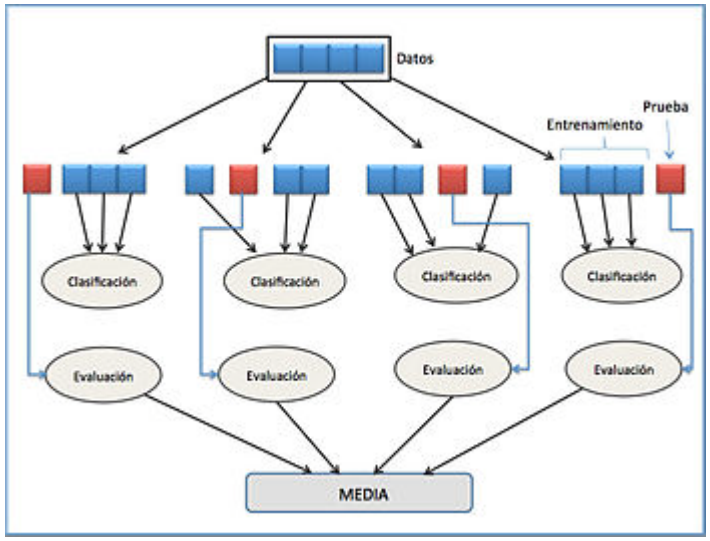
Java API

Machine Learning Approach

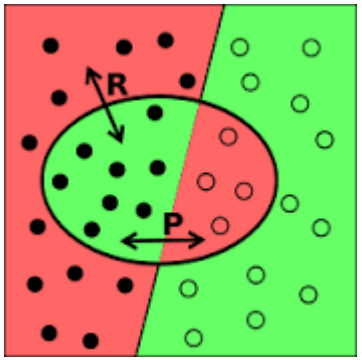


Learn and Evaluate the Classifier

- K-fold Cross Validation



k-fold cross validation, con k=4
http://es.wikipedia.org/wiki/Validaci%C3%B3n_cruzada



Learn and Evaluate the Classifier

- What if the text classifier does not perform as expected

Change classifier parameters

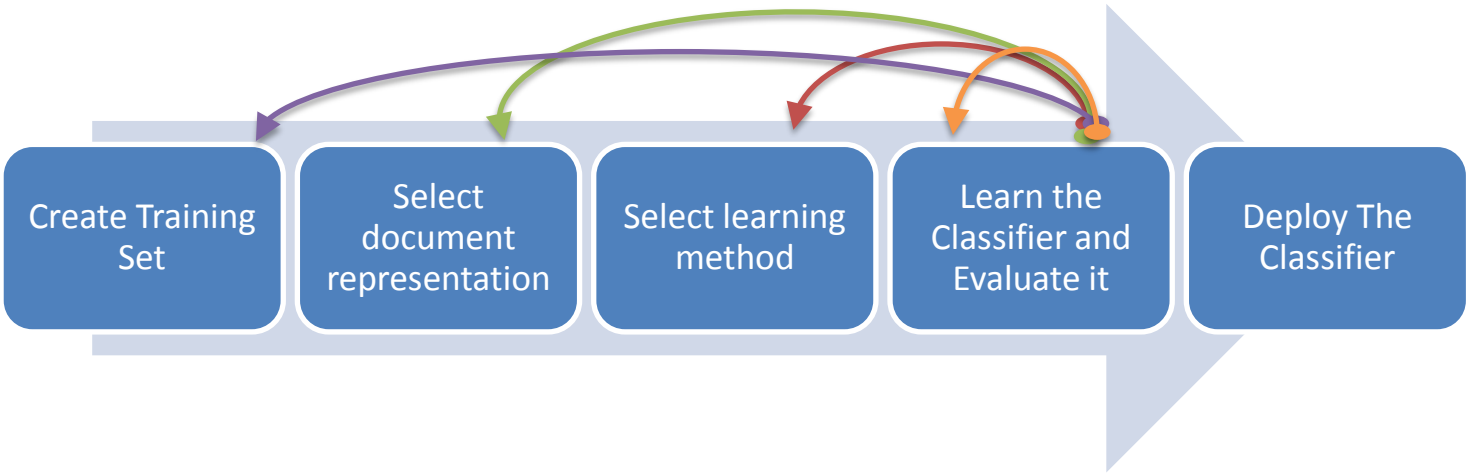
Use other classifiers

Feature Selection

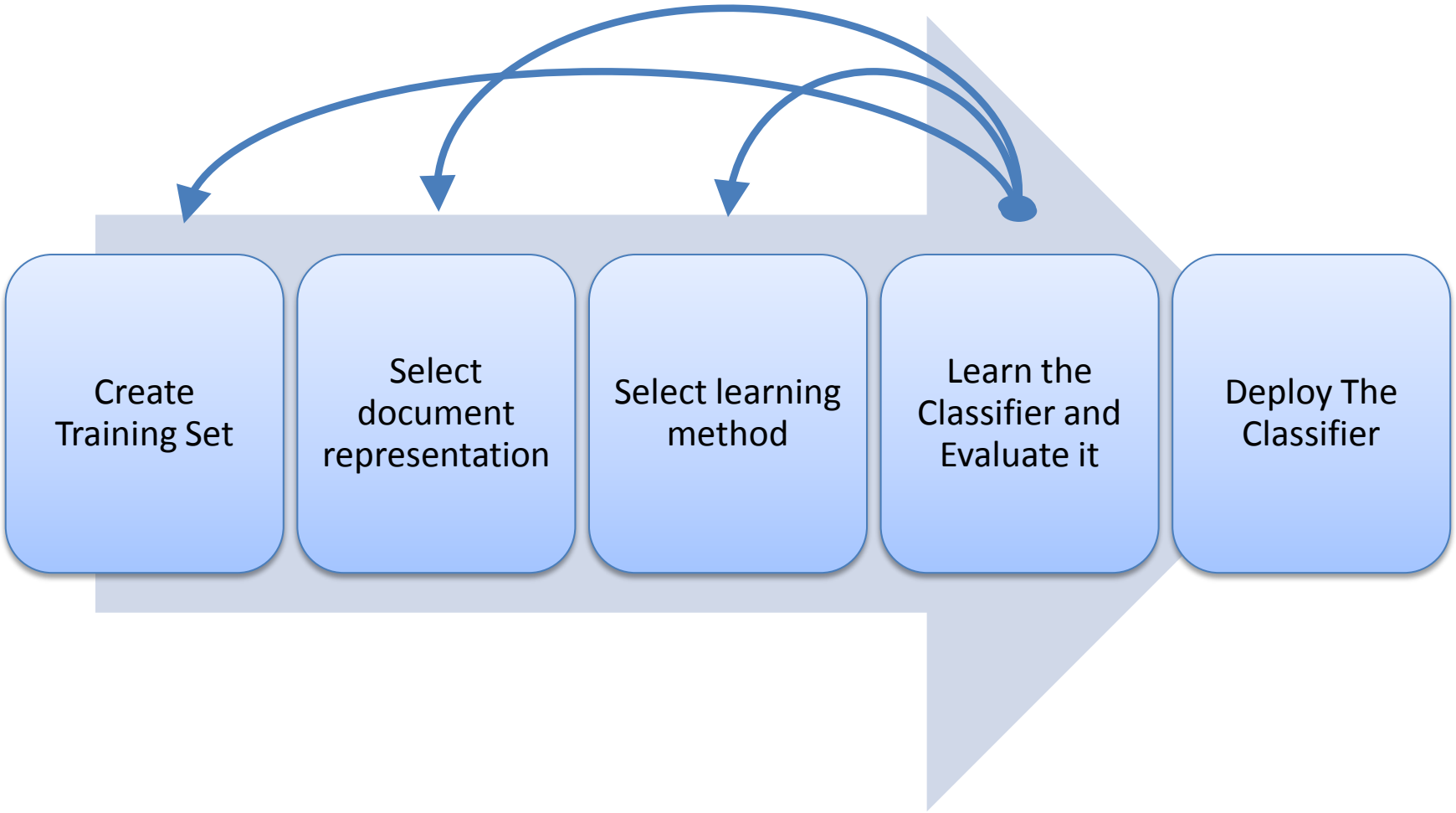
Mutual Information

Add new features that help to discriminate between classes

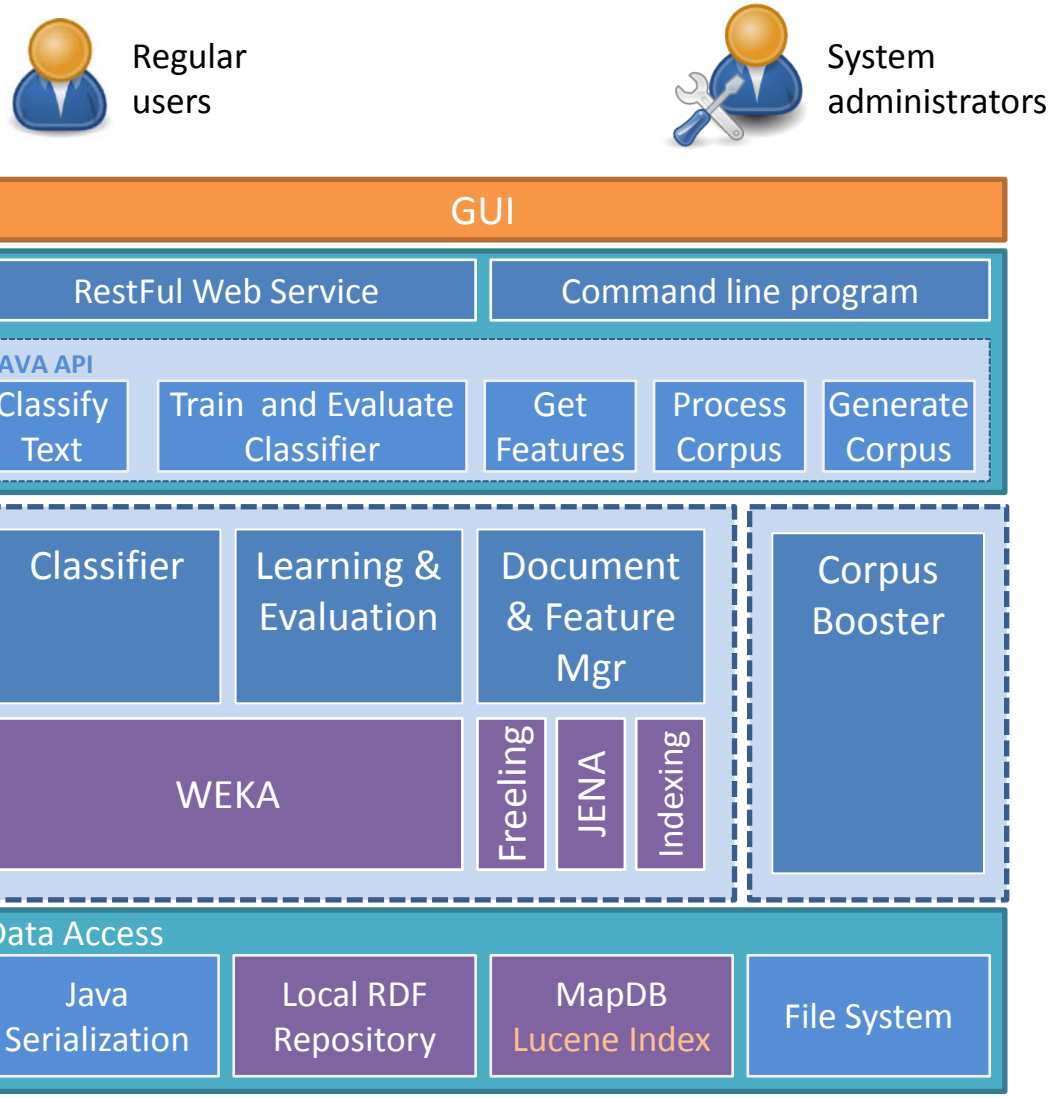
Modify your training set



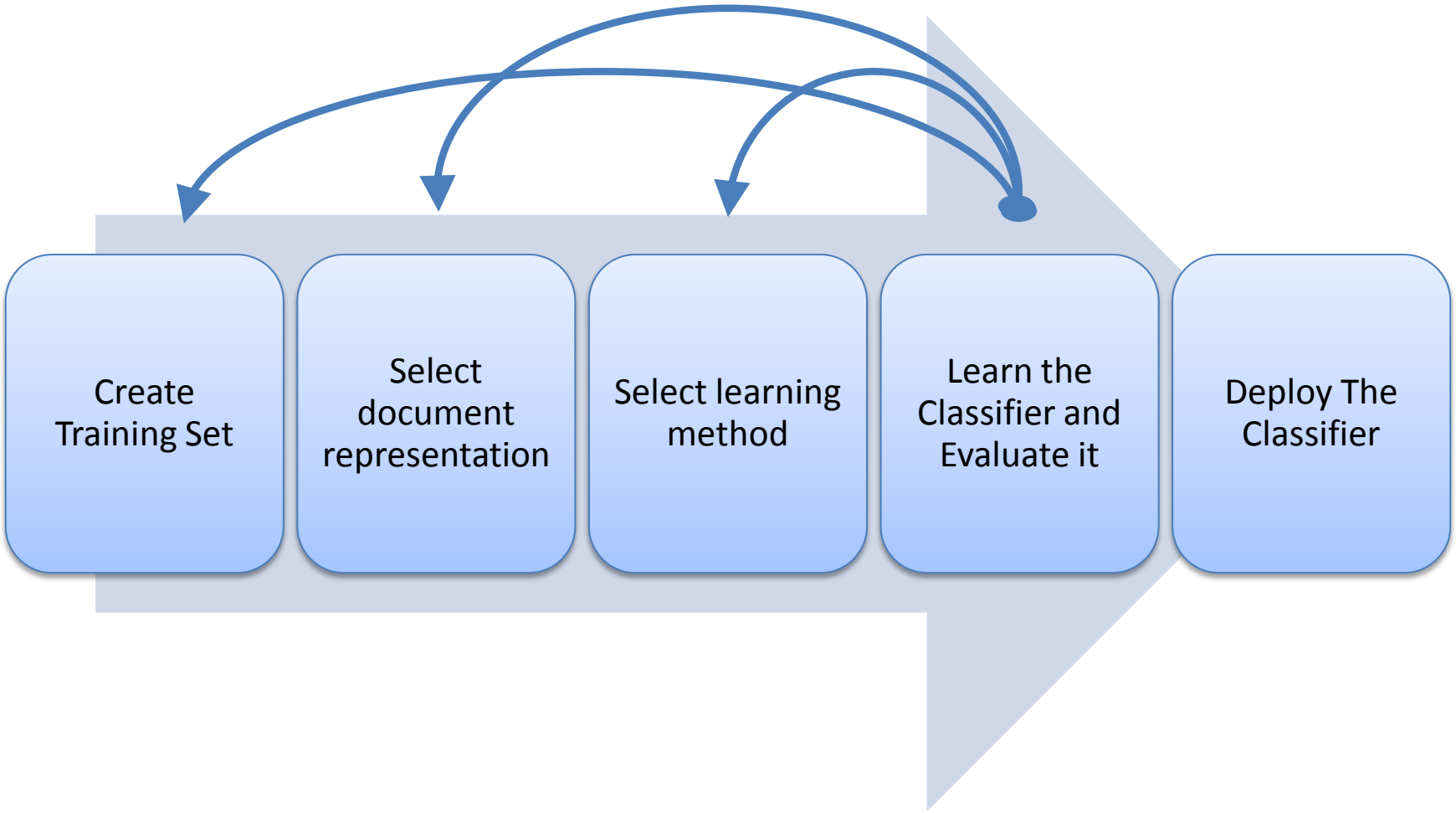
Machine Learning Approach

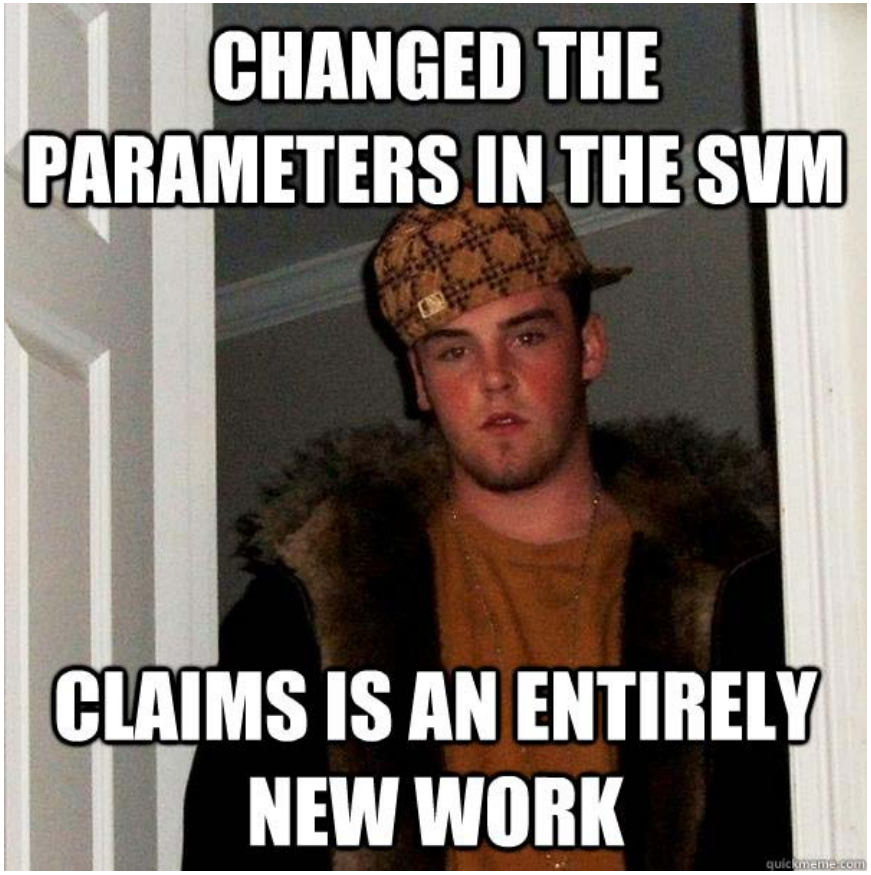


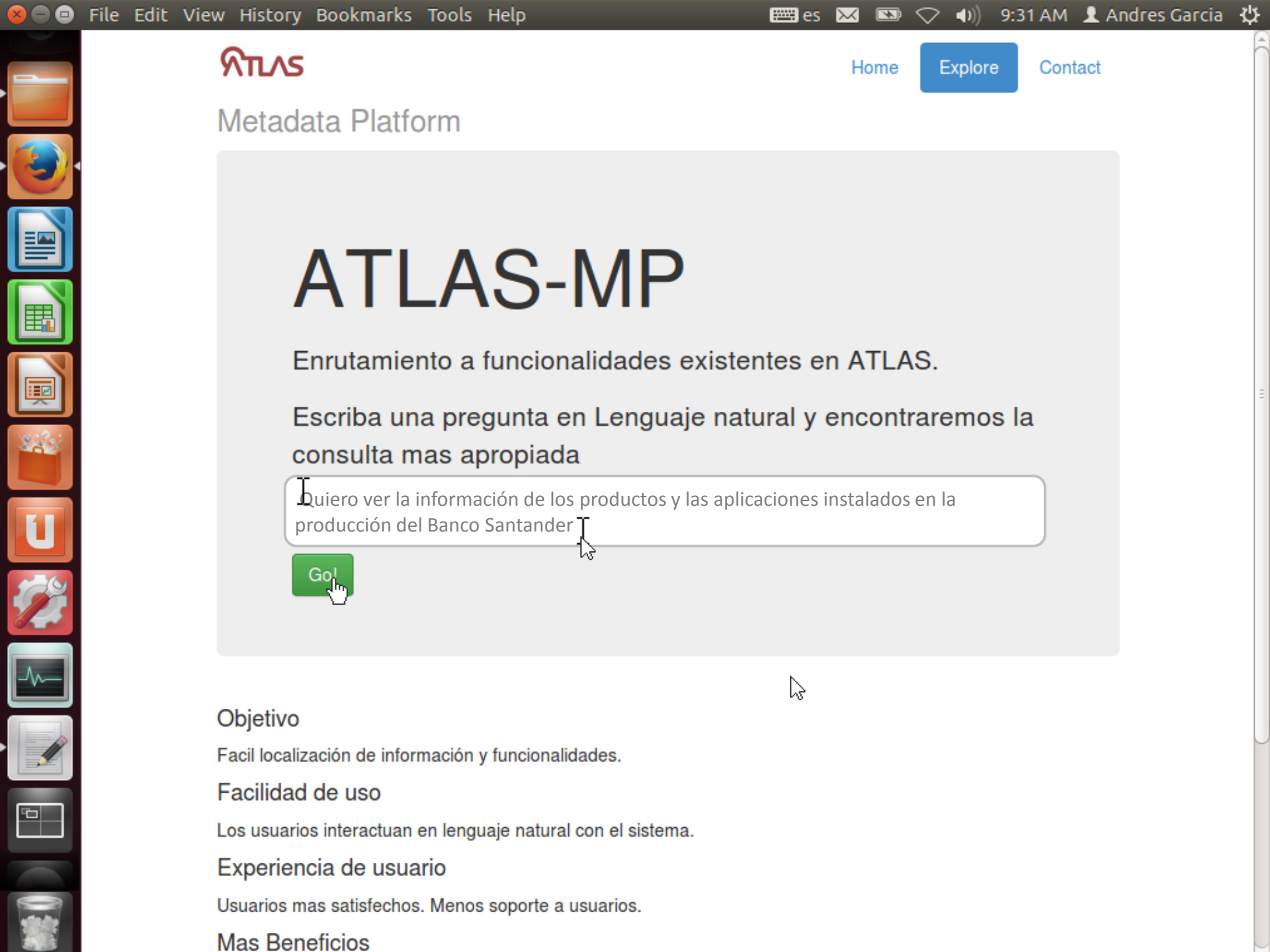
Architecture of a Text Classification System

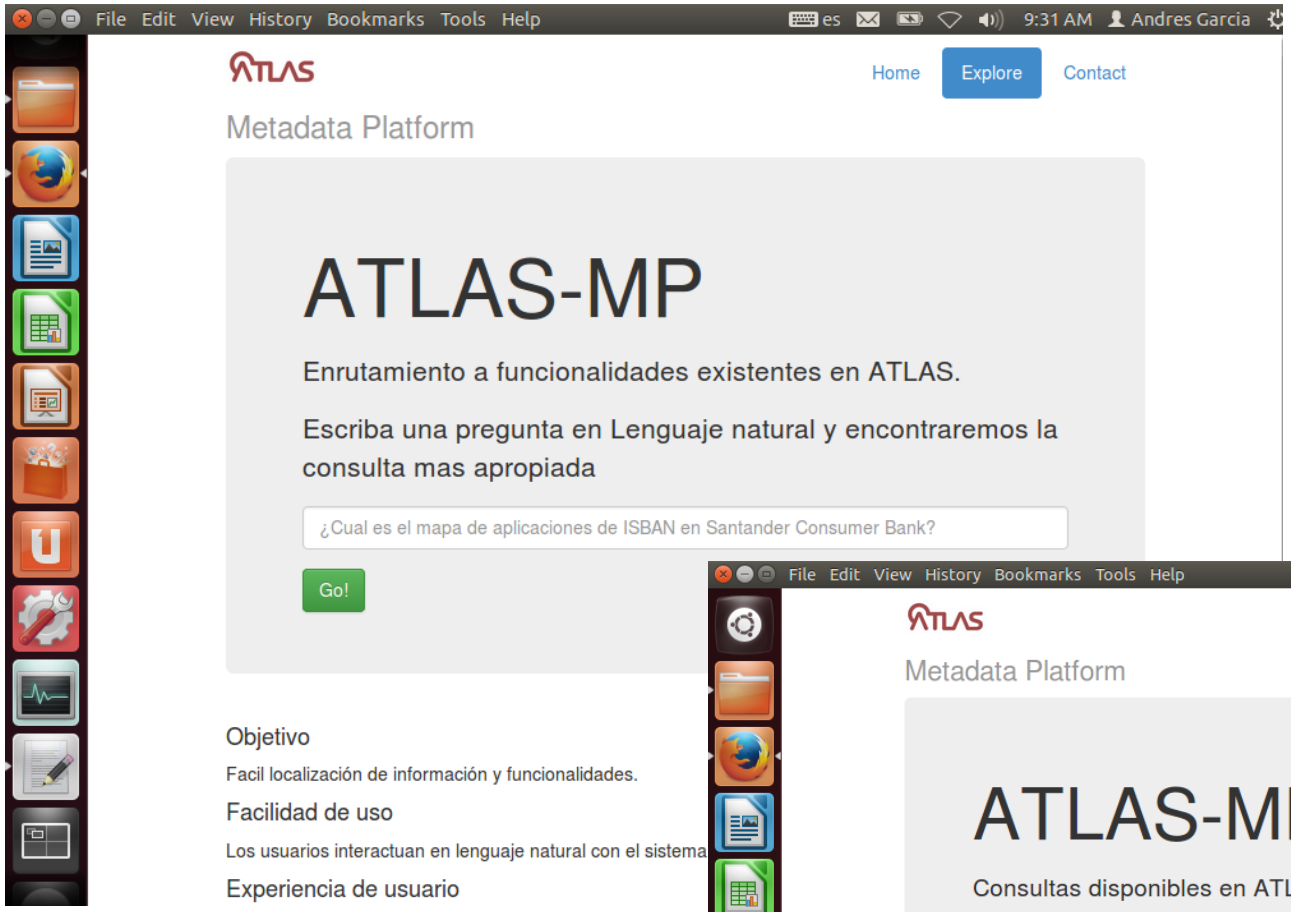


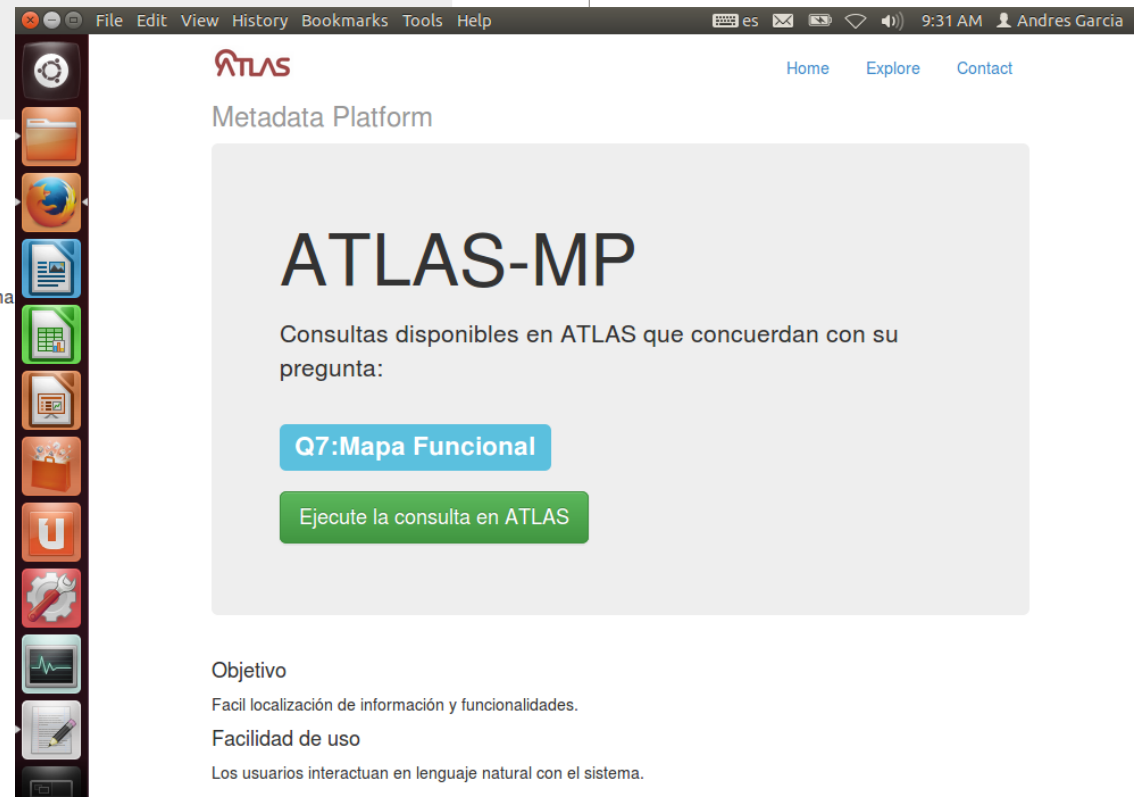
Machine Learning Approach




CHANGED THE PARAMETERS IN THE SVM
CLAIMS IS AN ENTIRELY NEW WORK
quickmeme.com


File Edit View History Bookmarks Tools Help
es 9:31 AM Andres Garcia
ATLAS
Home
Explore
Contact
Metadata Platform
ATLAS-MP
Enrutamiento a funcionalidades existentes en ATLAS.
Escriba una pregunta en Lenguaje natural y encontraremos la consulta mas apropiada
Quiero ver la información de los productos y las aplicaciones instalados en la producción del Banco Santander
Go!
Objetivo
Facil localización de información y funcionalidades.
Facilidad de uso
Los usuarios interactuan en lenguaje natural con el sistema.
Experiencia de usuario
Usuarios mas satisfechos. Menos soporte a usuarios.
Mas Beneficios


File Edit View History Bookmarks Tools Help
es 9:31 AM Andres Garcia
ATLAS
Home
Explore
Contact
Metadata Platform
ATLAS-MP
Enrutamiento a funcionalidades existentes en ATLAS.
Escriba una pregunta en Lenguaje natural y encontraremos la consulta mas apropiada
¿Cual es el mapa de aplicaciones de ISBAN en Santander Consumer Bank?
Go!
Objetivo
Facil localización de información y funcionalidades.
Facilidad de uso
Los usuarios interactuan en lenguaje natural con el sistema
Experiencia de usuario


File Edit View History Bookmarks Tools Help
es 9:31 AM Andres Garcia
ATLAS
Home
Explore
Contact
Metadata Platform
ATLAS-MP
Consultas disponibles en ATLAS que concuerdan con su pregunta:
Q7:Mapa Funcional
Ejecute la consulta en ATLAS
Objetivo
Facil localización de información y funcionalidades.
Facilidad de uso
Los usuarios interactuan en lenguaje natural con el sistema.