Universidad Simón Bolívar
Departamento de Computación y Tecnología de la Información
CI5438: Inteligencia Artificial II
Docente: Ivette Martínez

CLANews

Support Vector Machine

Alumnos: **[Grupo 06]**
Héctor Domínguez / 09-10241
Oskar González / 09-10351
Roberto Heligon / 09-10395

Agenda

- Introducción - ¿Qué es CLANews?
- Extracción de Datos
  - Pre-procesamiento
- Construcción de la máquina
- Resultados obtenidos
- Resultados esperados
- Conclusiones
- Referencias Bibliográficas

Introducción

- CLANews: Clasificador de titulares de noticias
- Idiomas: inglés y español

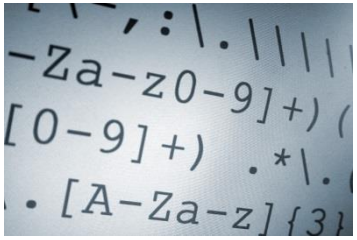
Extracción de Datos

TWITTER4J

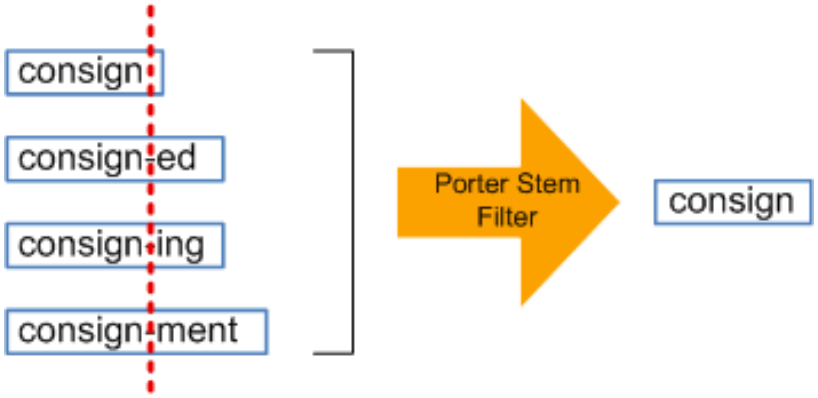
- Twitter
- Categorías de noticias consideradas:
  - Business: ReutersBiz, BBCBusiness, Forbes, nytimesbusiness / **WSJbusiness**
  - Entertainment: ReutersShowbiz, EW, Enews, MSN_Entertain / **CNNent**
  - Politics: ReutersPolitics, foxnewspolitics, ABCPolitics, BBCPolitics / **CNNPolitics**
  - Sport: ReutersSports, BBCSport, SportsCenter, espn / **CBSSports**
  - Technology: ReutersTech, engadget, BBCTech, CNET / **usatodaytech**

Extracción de Datos

- Pre-procesamiento de los datos:
  - Remover links
  - Remover números

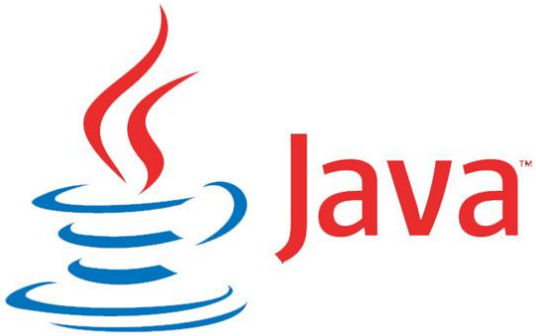


  - Tokenizing
  - Stemming
  - Remover stop-words



Construcción de la máquina

- Weka
- LibSVM: Chih-Chung Chang and Chih-Jen Lin – autores



- 2.500 atributos y 4.000 instancias (tweets) por categoría.
- Entrenamiento: 10-cross validation

Resultados Obtenidos

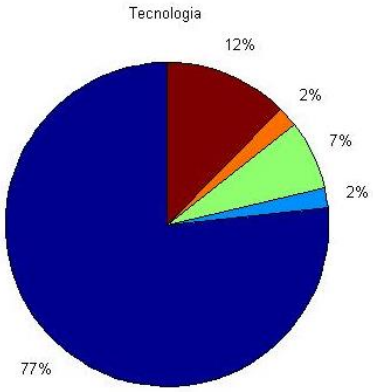
- Instancias correctamente clasificadas en entrenamiento:
  - 16.445 de 20.000 → 82.225 %

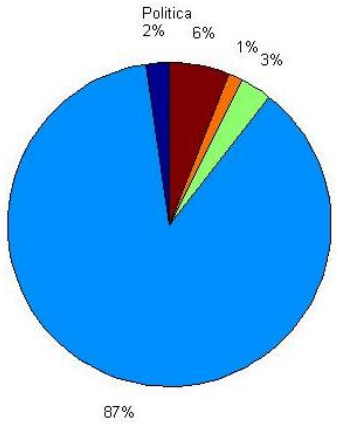
- Matriz de Confusión:

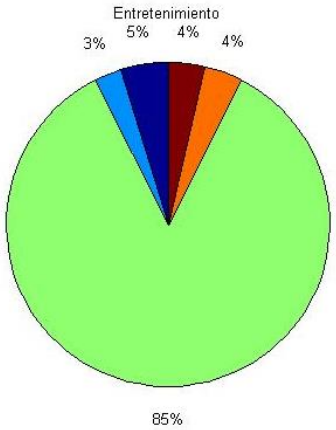
Clasificada como / Verdadera clase	A	B	C	D	E
A: technology	3066	78	279	81	496
B: politics	94	3483	127	55	241
C: entertainment	187	110	3406	151	146
D: sport	80	65	220	3509	126
E: business	533	176	186	124	2981

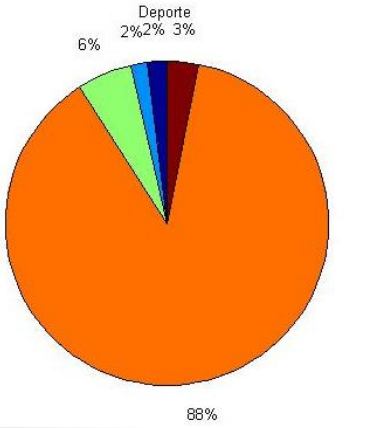
Resultados Obtenidos

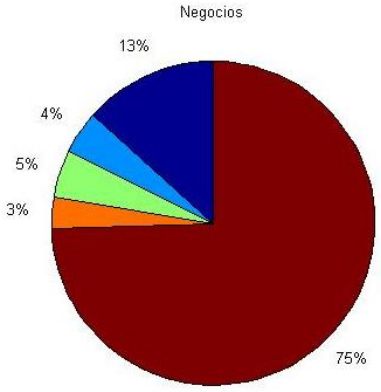
- Entrenamiento

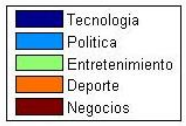






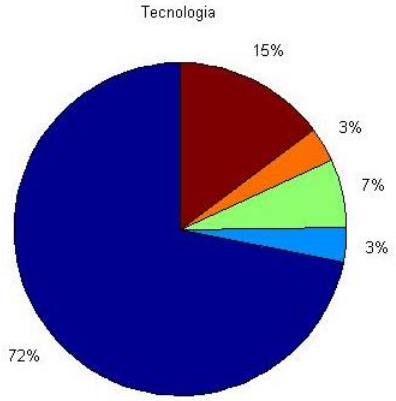


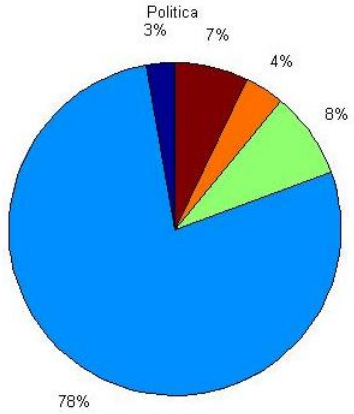


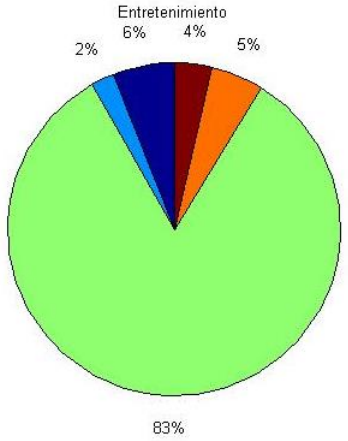


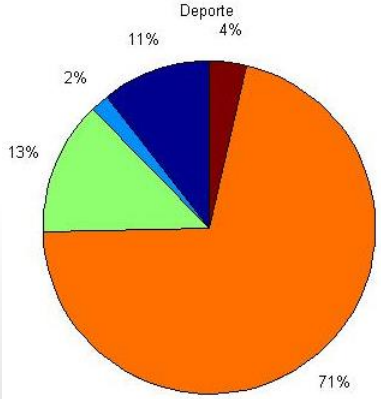
Resultados Obtenidos

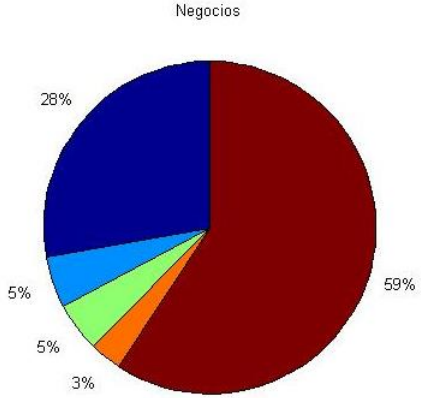
- Validación

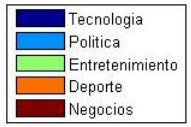












Resultados Esperados

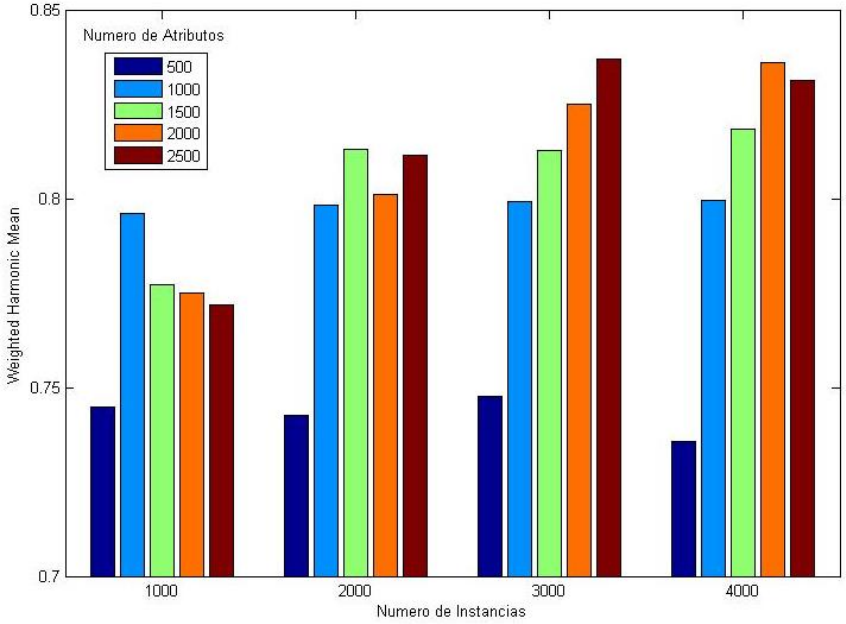
- Tuning

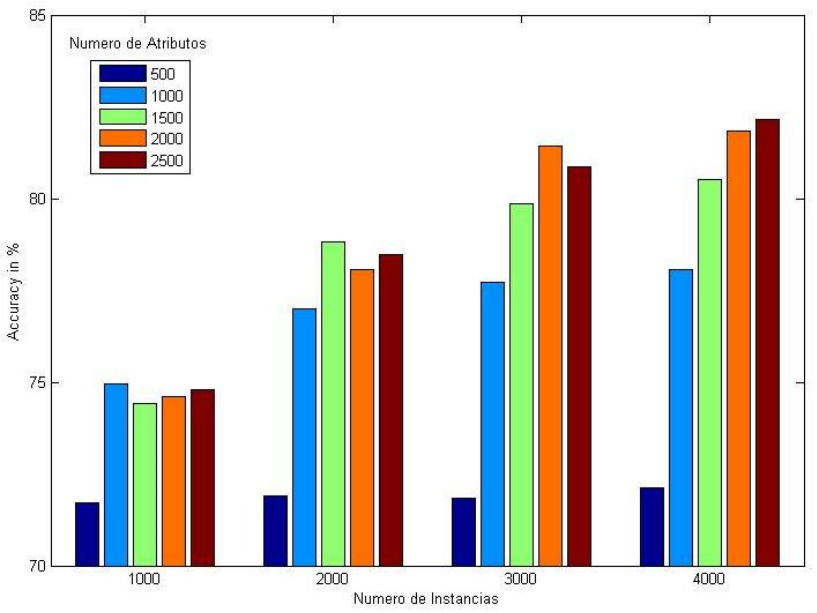
Intel Core i5, 2.67 GHz,

4 GB RAM, S.O Ubuntu 14.04

Tiempo de construcción:

$\approx 20\ min$





Tiempo de validación:

2.500 tweets

$\approx 2.5\ seg$

Conclusiones

- Se logró construir un clasificador relativamente rápido
- El clasificador tiene acierto en más del 70 % de los casos
  - **Excepción: Business**
- Se utilizaron librerías famosas en el área de *Machine Learning*
  - Weka
  - LibSVM
- Nos enfrentamos a un problema “real”

Referencias Bibliográficas

- ***Apache Lucene***. Disponible en: http://lucene.apache.org/
- Chih-Chung Chang and Chih-Jen Lin. ***LIBSVM -- A Library for Support Vector Machines***. Disponible en: http://www.csie.ntu.edu.tw/~cjlin/libsvm/
- Inoshika Dilrukshi, Kasun De Zoysa, Amitha Caldera. ***Twitter News Classification Using SVM***. The 8th International Conference on Computer Science & Education (ICCSE), 2013, pp 287 – 291. Colombo. IEEE, DOI: 10.1109/ICCSE.2013.6553926
- Ukrit Wattanavaekin, Wipawee Amornwat. ***Classification of Twitter News***. School of Information, Computer and Communication Technology, Sirindhorn International Institute of Technology, Thammasat University. 2013.
- ***Waikato Environment for Knowledge Analysis – WEKA***: http://www.cs.waikato.ac.nz/ml/weka/
- ***Twitter API***: https://dev.twitter.com/overview/documentation
- ***Twitter4j library***. Disponible en: http://twitter4j.org/en/index.html