

**UNIVERSIDAD ANDRÉS BELLO**

**FACULTAD DE INGENIERÍA**

**INGENIERÍA CIVIL INFORMÁTICA**

**Trabajo 2 Etiquetado – Aprendizaje Supervisado (SVM)**

**ANDRÉS EDUARDO VALENZUELA GONZÁLEZ**

**SANTIAGO - CHILE**

**SEPTIEMBRE, 2017**

Contenido

[**1.** **Introducción** 3](#_Toc492252878)

[**2.** **Descripción del Problema** 3](#_Toc492252879)

[**3.** **Desarrollo** 4](#_Toc492252880)

[3.1. Describe extensamente su experiencia de etiquetado.. 4](#_Toc492252881)

[3.2. ¿Qué características tienen los usuarios relacionados? 5](#_Toc492252882)

[3.2.1. Distribución de Datos 6](#_Toc492252883)

[3.3. ¿Hay relación entre algún lugar de fuera y dentro de USA? 7](#_Toc492252884)

[3.4. Calcule la entropía de sus conjuntos de datos en términos de ubicación geográfica. 7](#_Toc492252885)

[3.5. Calcule la Información mutua para los valores de *World* 8](#_Toc492252886)

[**4.** **Conclusión** 9](#_Toc492252887)

# **Introducción**

Para comprender el desarrollo de este informe, se utilizaron las dos mil etiquetas de la tarea anterior para ser vectorizadas y luego utilizar los algoritmos de *SVM* (y otros creados por el alumnado) para lograr un aprendizaje supervisado.

Junto con las etiquetas proporcionadas (en formato *csv*) por el servidor de la *Universidad Nacional Andrés Bello*, se proporciono además el conjunto de perfiles que fueron etiquetados por el alumnado (en formato *dat*), los cuales fueron unidos para luego generar una bolsa de palabras, de las cuales, además, se descontaron las *stopwords*.

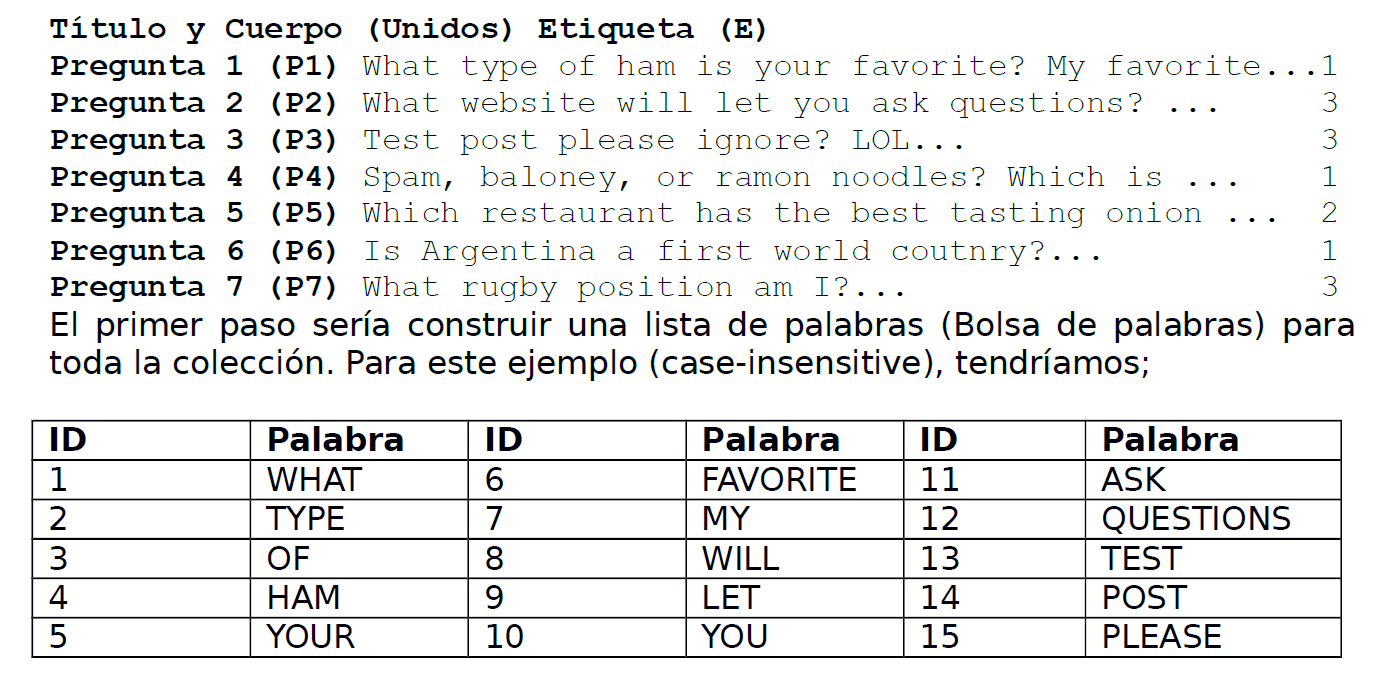
Para realizar el aprendizaje supervisado se debio modelar el problema mediante vectores.   
Si asumimos que cada una de las palabras es un atributo, entonces podríamos crear una representación vectorial que indique la presencia o ausencia de esta palabra en cada una de las descripciones del usuario etiquetadas en la primera tarea. Para la etiqueta se utilizará un valor desde 1 hasta 4:

1 = *Undetermined* 2 = *Non-USA* 3 = *World* 4 = *USA only*.

Cada etiqueta debe estar asociada con un único valor numérico. *Extracto del enunciado Tarea 2, Espacio Vectorial, 28 de agosto 2017 [consulta: 13 septiembre 2017, 20:12 hrs]. Disponible en:* <https://gitlab.com/Choapinus/SistemasInteligentes/blob/master/Tarea%202/enunciado_informe/Tarea%202.pdf>

# **Descripción de la Vectorizacion**

Para explicar el proceso de vectorizacion, se considerara el siguiente ejemplo que trabaja con preguntas *cQA* (cada perfil es una pregunta):



Se representaron las preguntas como vectores indicando la frecuencia de cada palabra en la pregunta con su índice correspondiente respecto a la bolsa de palabras (Por bolsa de palabras entenderemos a las palabras ordenadas respecto al orden de llegada desde los perfiles, con las *stopwords* descontadas). Luego, se generaron los vectores con formato *SVM multiclass*, descrito a continuación:

<target> <feature>:<value> <feature>:<value> ... <feature>:<value>

En donde:

* **Target:** Valor de 1 a 4, representa la clase a la que pertenece.
* **Feature:** id de la palabra dentro de la bolsa de palabras.

# **Desarrollo**

## Describe extensamente su experiencia de etiquetado: ¿Qué observó?, ¿Qué dificultades encontró?, ¿Qué observa de las preguntas que fueron en la clase *“Undetermined”*? Discuta esto en extenso y ejemplificando sus observaciones.

Durante el etiquetado se observó un mundo de perfiles similares en estructura de descripción y otra variedad de perfiles distintos al primero, pero similares entre sí (con una pobre descripción y solo la puesta en evidencia de la ubicación geográfica). Además, una gran cantidad de perfiles de la clase *USA only* demostraron creer que *Estados Unidos* es el único país en *América* e incluso, en el mundo porque si bien daban a conocer su ubicación, solo escribían la abreviación del estado (o solo *US* en ciertos casos), la abreviación de la ciudad en donde viven, la abreviación del pueblo en donde viven o el sector en donde viven **sin especificar explícitamente de que país pertenecen**. En el caso de *Georgia*, se puede mal interpretar debido a que este es un estado de *Estados Unidos*, pero a su vez es un país.  
En casi el total de los perfiles, el inglés fue el idioma predominante (otros idiomas fueron el español, portugués, coreano y chino), pero cabe destacar que existieron perfiles que no describían a una persona sino a una empresa, describían provenir de un estado que también es considerado un país, algunos se enorgullecían al comentar que eran veteranos de guerra o que pertenecían a la armada, pero no exponían el nombre de su amado país, algunos eran demasiado ambiguos en especificar su localidad, otros constaban de pasajes de la biblia y/o poseían una redacción paupérrima junto con una pésima ortografía. Esto y otros factores de menor importancia dificultaron el etiquetado y protagonizo la selección de la etiqueta *Undetermined* a un poco más de un cuarto del *dataset* total.

La etiqueta más simple de obtener fue *World* dado a que fue sencilla de identificar dado a que basto que poseyera una sola referencia a *Estados Unidos* y al menos una referencia a cualquier otro país para obtener mencionada etiqueta.  
La más complicada de obtener fue *USA only* y *Non-USA* debido a que sin internet, el alumnado no tiene conocimiento absoluto de todos los estados de *Estados Unidos* ni de todos los países del mundo.

## ¿Qué características tienen los usuarios relacionados con cada ubicación geográfica (las más prominentes)? ¿Cuál es la distribución en su conjunto de datos? ¿Cuál es la probabilidad de cada uno de los valores de variables?

Para la etiqueta *USA only* se puede apreciar que los tres sectores más prominentes en cantidad son *California, Texas* y *Florida* (98, 66 y 62 veces contados respectivamente de un total de alrededor de 220 sectores distintos)*.* Estos tres estados se ubican al sur de los *Estados Unidos*, cada uno separado por una distancia de no más de 3 estados, por lo cual se infiere que dentro de la clase *USA only* existe una gran probabilidad de que la nueva etiqueta pertenezca a uno de estas tres localidades antes mencionadas.

Por ser de *Estados Unidos*, varios perfiles descrito bajo estos 3 estados **no** **mencionaron al país que pertenecen.** Y en cuanto a relación, los perfiles clasificados bajo la clase *USA only* poseen un rango de edad de entre los 16 a 35 años. Entre mayor es el sujeto descrito (25-30 años), más veces fue provista la ubicación geográfica y la manera de descripción de sujeto fue considerada más por parte de logros y metas. Entre menor sea el sujeto descrito (24 hacia abajo), el perfil comienza a tomar forma de descripción de gustos adolecentes y se provee información respecto al lugar de estudios (lo cual no fue de interés en esta ocasión).

### Distribución de Datos

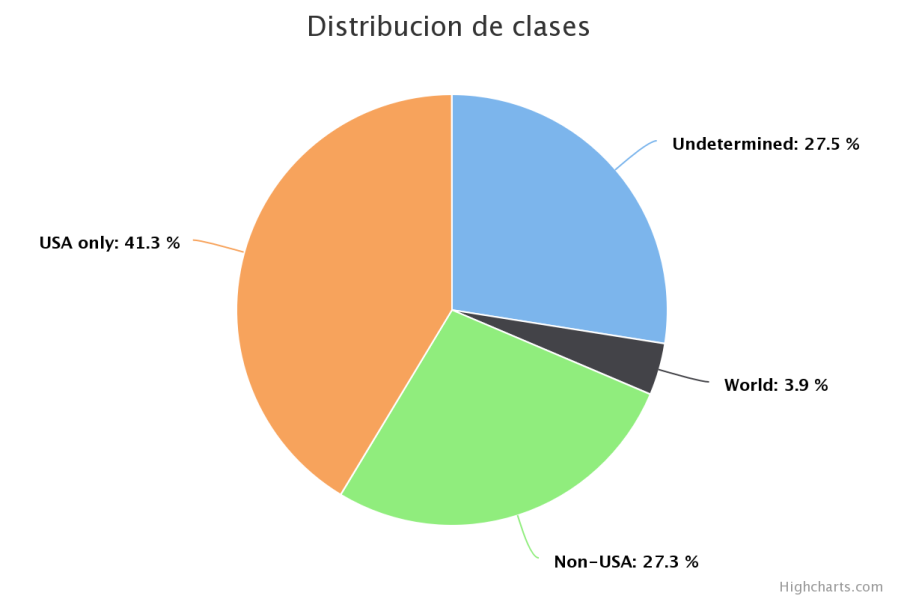


Figura 2. Distribución de datos

Como se puede apreciar en la figura 2, se obtuvo un 41.3% para la clase *USA only* (827 de 2000), 27.3% para la clase *Non-USA* (545 de 2000), un 3.9% para la clase *World* (78 de 2000) y un 27.5% para la etiqueta *Undetermined* (550 de 2000).

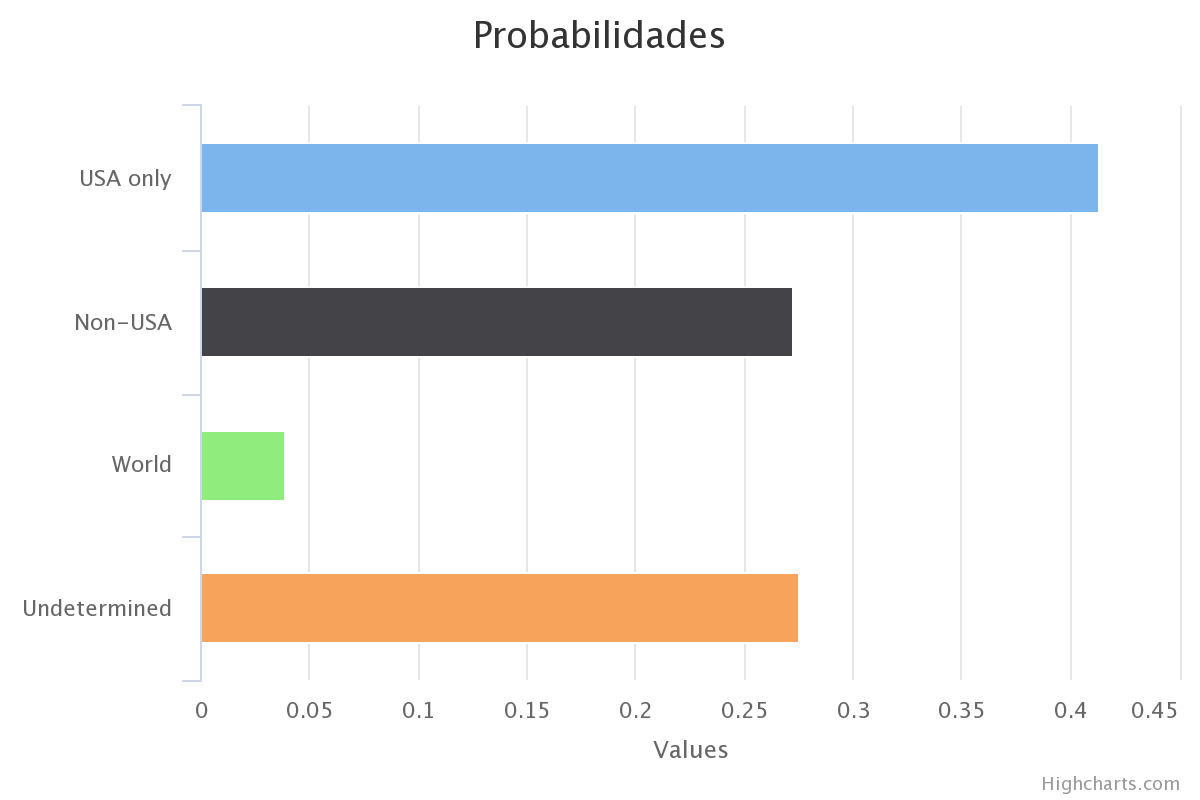


Figura 3. Probabilidad de cada clase

En la figura 3, para las probabilidades de cada clase se tiene que:

* ***USA only*** tiene un 0.4135 de probabilidad de ocurrir.
* ***Non-USA*** posee un 0.2725 de probabilidad de ocurrir.
* ***World*** posee un 0.039 de probabilidad de ocurrir.
* ***Undetermined*** tiene un 0.275 de probabilidad de ocurrir.

## ¿Hay relación entre algún lugar de fuera y dentro de USA?

La clase *World* demuestra que existen al menos 78 distintas relaciones entre *USA* y algún lugar de fuera como por ejemplo la creencia de que *Estados Unidos* es el único país de *América* y por ende el más importante del planeta, otra relación es el por qué posee esta etiqueta. Algunos perfiles describen viajes por todo *Estados Unidos* y algunos perfiles de la clase *World* también describen viajes turísticos. Una relación más apunta a los comentarios que se hacen dentro de la etiqueta *USA only, tales como* “me gustaría irme a vivir a X parte del mundo y estoy juntando dinero para ello”, mientras que algunos perfiles de la etiqueta *World* ya lo están haciendo, es decir que los perfiles antes mencionados de *USA only* pueden potencialmente convertirse en *World* en un tiempo más, aun así es una mera suposición.

## Calcule la entropía de sus conjuntos de datos en términos de ubicación geográfica.

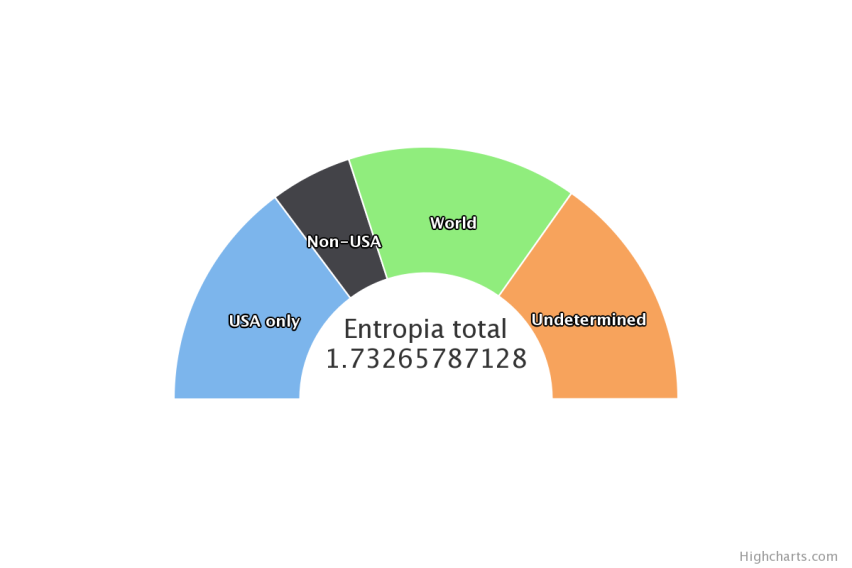


Figura 4. Entropías

Para la figura 4 se tiene que:

* La clase *USA only* posee una entropía de 0.52681585653
* La clase *Non-USA* posee una entropía de 0.511120583212
* La clase *World* posee una entropía de 0.182534900566
* La clase *Undetermined* posee una entropía de 0.512186530969

Como se puede apreciar en la figura 4, la entropía total fue de 1.7326.  
Cabe destacar que *“a menor entropía, mayor certidumbre”,* por lo cual se puede afirmar que la clase que posee la mayor certidumbre es *World* debido a la facilidad de identificar que una persona haya viajado por el mundo y guarde relación con *Estados Unidos* **siempre y cuando nos comente sus viajes**. De otra manera, pasaría a ser una etiqueta más de *Undetermined* o *Non-USA*, dependiendo de si solo comenta que ha viajado por todo el mundo o si no menciona a *Estados Unidos*, respectivamente*.*

## Calcule la Información mutua para los valores de *World*

Antes de exponer la información mutua obtenida de un conjunto de datos selecto de la etiqueta *World*, primero debemos entender que la información mutua (o transinformación) de dos variables aleatorias es una cantidad que mide la dependencia mutua de las dos variables, es decir, mide la reducción de la incertidumbre (entropía) de una variable aleatoria X, debido al conocimiento del valor de otra variable aleatoria Y. *extracto de Wikipedia, Información Mutua, 15 de abril 2017, 00:53 [consulta: 04 septiembre 2017, 1:10 am]. Disponible en:* [*https://es.wikipedia.org/wiki/Informaci%C3%B3n\_mutua*](https://es.wikipedia.org/wiki/Informaci%C3%B3n_mutua)

Información mutua:

|  |  |  |  |
| --- | --- | --- | --- |
| Canadá: 0.00146 | Italy: 0.00917 | Chicago: 0.00249 | Indiana: 0.00302 |
| Texas: 0.000189 | Australia: 2.18e-06 | Delhi: 0.00117 | Virginia: 0.00229 |
| Connecticut: 0.00213 | California: 8.24e-05 | Los Ángeles: 0.00055 | Germany: 0.00302 |
| Tampa: 0.00213 | Miami: 0.00131 | Florida: 5.24e-08 | Alaska: 0.00407 |
| Hawaii: 0.00227 | México: 0.00833 | India: 0.000161 | US: 0.00802 |
| New York: 0.00518 | London: 3.07e-05 | UK: 0.000334 | Maine: 0.00336 |

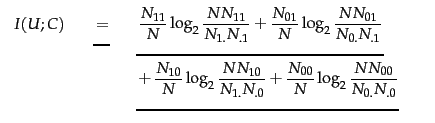


Figura 5. Formula Información Mutua

**Consideraciones:** Para la transinformación se tomó en cuenta los sectores más prominentes ya que una gran cantidad de países representaban una constante cercana a uno, por lo cual se encontraba poco pertinente calcular lo mismo para 30 países distintos. Se adjunta prueba.

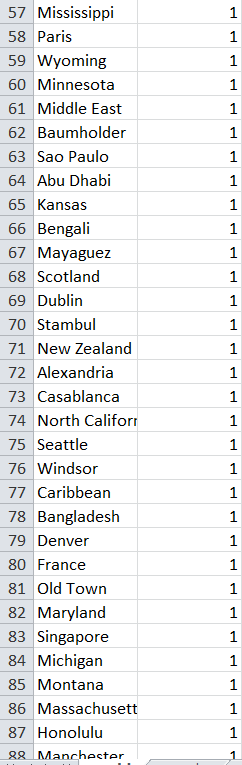


Figura 6. Prueba discriminatoria

# **Conclusión**

Al finalizar el etiquetado se pudo concluir lo siguiente:

* Dejando la clase *USA only* de lado, la segunda mayor clase es *Undetermined*. Esto dificulta en gran medida un proceso de aprendizaje y rescate de resultados ante los perfiles que no especifican nada de su ubicación geográfica, en cambio si el estudio fuese otro y con otras etiquetas, el trabajo sería mucho más arduo, pero se podría rescatar aún más información respecto a los perfiles. Para este punto, en síntesis, entre más datos útiles se puedan rescatar de los perfiles, más análisis se pueden hacer al respecto, por ejemplo gustos/metas/necesidades en común, edad, tipo de descripción de perfil según la edad, cantidad de hombres y mujeres, tendencias políticas, estado civil, clase social, etc.
* La extensión del dataset influye bastante en la tendencia de los datos. Experimentalmente se redujo el espectro de datos al 10% del original poseído (200 perfiles) y los datos distan mucho de los originales. Como se esperaba, la magnitud de cada clase no cambio al 10% de su valor original con 2000 perfiles. Esto tiene sentido solo al pensar que los perfiles provistos por el servidor son aleatorios y no siempre se darán los mismos resultados, es decir que otra persona que haga este mismo experimento posiblemente obtendrá resultados distintos a los que se adjudican a continuación.
* Las cantidades experimentales fueron:
  + Cantidad de perfiles en *Undetermined*: 56
  + Probabilidad de *Undetermined*: 0.028
  + Cantidad de perfiles en *World*: 10
  + Probabilidad de *World*: 0.005
  + Cantidad de perfiles en *Non-USA*: 57
  + Probabilidad de *Non-USA*: 0.0285
  + Cantidad de perfiles en *USA only*: 77
  + Probabilidad de *USA only*: 0.0385
* Entropías:
  + *Undetermined*: 0.144436022153
  + *World*: 0.0382192809489
  + *Non-USA*: 0.146287486709
  + *USA only*: 0.180911413143
  + Entropía total: 0.50971

De todas maneras, lo hecho en el experimento tampoco es incorrecto ya que la entropía, es decir, la certidumbre de los datos se redujo en gran medida, lo cual puede ser beneficioso para algún tipo de entrenamiento supervisado como los *Arboles de Decisión* y los *K-NN* *(K-Nearest Neighbors | K-vecinos más cercanos).*

* Entre todas las etiquetas se pueden apreciar varias tendencias destacables, por ejemplo una gran cantidad de perfiles no especifican su religión, su tendencia política, su lugar de trabajo, su fecha de nacimiento, correo electrónico (se asume que la plataforma tiene una sección en donde detallar este dato) entre otros.
* Por último, se entiende que la información rescatada (ubicación geográfica) se puede aplicar para realizar publicidad enfocada a las regiones más prominentes o realizar un nuevo estudio enfocado en otro tema con base en los países mas destacables ya que se daría por supuesto que habrá una mejor participación comparado con los países con menor cantidad de perfiles. Estudios tales como entender y declarar los años de escolaridad por cada nacionalidad, tipo de estado laboral actual, verificar estado civil, motivos para viajar o no viajar a otros países, ganancia neta/bruta mensual, edad promedio, idioma más hablado (mundialmente), cultura más frecuente, religión más frecuente, etc.