

Universidad Andrés Bello

Facultad de Ingeniería

Ingeniería Civil Informática

**Trabajo 1**

**Etiquetado – Ubicación geográfica de Usuarios de PLATAFORMAS cQA**

**Sistemas Inteligentes**

David Andrés Molina Garrido

Profesor

Alejandro Figueroa Amenábar

*Santiago - Chile.*

*Agosto, 2017.*

**Introducción**

Las comunidades de Pregunta-Respuesta (cQA) son sitios donde, en resumen, sus miembros responden a preguntas realizadas por otros usuarios obteniendo cierto puntaje o rango el usuario que lo hace de manera correcta. La cualidad de estos sitios y que lo hace diferente a otro es su finalidad, ésta está destinada a encontrar respuestas especificas a cierta pregunta, ya sean éstas complejas o de conocimiento general en muchos casos. Dentro de estas páginas se encuentran StackExchange, Quora o Yahoo!. En este informe nos enfocaremos básicamente en la descripción del perfil de estos usuarios, y en como proveen, ya sea de manera explícita o implícita, antecedentes de su ubicación geográfica o de los lugares a los cuales ha estado arraigado en su vida.

En este informe trataremos acerca de la experiencia que tomó etiquetar un cierto número de esas descripciones hechas por los usuarios y clasificarlas en 4 clases diferentes:

-**USA only** (Si está relacionado con sólo lugares dentro de USA).

-**Non-USA** (Si está relacionado con sólo lugares fuera de USA).

-**World** (Si está relacionado con lugares dentro y fuera de USA).

-**Undetermined** (Si no provee ninguna pista de algún lugar geográfico relacionado al usuario).

A su vez veremos diferentes relaciones, otros aspectos de cada clase y en conjunto, para poder obtener datos tangibles como probabilidades y otras métricas relacionadas al etiquetado.

**Problema**

La problemática se basa en el etiquetado de 2000 descripciones de usuarios y clasificarlas en 4 clases según su ubicación geográfica. Dentro de estas descripciones encontramos algunas donde se expresa muy claramente y sin doble interpretación el lugar donde ha vivido y/o donde vive el usuario, como otras donde hay que recurrir (levemente) a inferir la información. Para clasificar de manera correcta estas preguntas ocuparemos las clases descritas anteriormente y utilizaremos métodos de cálculo de probabilidades. Algunas de las preguntas y cálculos para analizar el problema serán:

- ¿Qué observó? ¿Qué dificultades encontró? ¿Qué observa de las preguntas que fueron

en la clase “Undetermined”? Discuta esto en extenso y ejemplificando sus

observaciones.

- ¿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?

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

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

- Calcule la información mutua para los valores de fuera y dentro de USA(World).

Experiencias durante el proceso de etiquetado

(dificultades y observaciones)

Al momento de clasificar los perfiles de los sitios CQa se tomó en cuenta todo el texto de éste, y se verificó en detenimiento para poder obtener la información geográfica del usuario. Dentro de los perfiles hubo algunos donde resultó bastante fácil identificar su ubicación, esto porque solo había vivido en un solo lugar y además lo expresaba enfáticamente. Otros perfiles resultaron un tanto más complejas de identificar, por ejemplo, porque hacían mención a lugares específicos, pero de los cuales tenían cierto interés en visitar o solamente les gustaba. Además, se infirió la ubicación de ciertos usuarios que realizaron labores en ciertas organizaciones específicas, como es el caso de soldados retirados y gente graduada en distintas universidades.

Si bien es cierto el etiquetar 2000 perfiles no fue tarea fácil, tampoco fue caótico el realizarlo, esto porque unos conjuntos de los perfiles existentes fueron bastante fáciles de identificar y no requería de mucho tiempo el obtener la ubicación, algunas con muy poco texto y que contenían la información precisa y concisa del usuario (su ubicación).

Dentro de la clase “Undetermined” se encuentran aquellos perfiles donde no había ningún atisbo de la ubicación del usuario, simplemente porque no la colocó o porque no se logró inferir del texto (muy rebuscada).

La clase “Undetermined” obtuvo el 3er lugar en la distribución de las clases, quedando no muy alejada del 2do lugar, esto denota que no fue menor el porcentaje de esta clase. Muchas de los usuarios etiquetados con esta clase hacían mención a intereses en particular más que colocar su biografía o alguna reseña de su vida, algunas de los usuarios que quedaron en esta clase eran fans de bandas, anime o en otros casos simplemente usuarios que colocaban poco o casi nada de texto, con lo cual no se lograba inferir absolutamente nada.

El proceso de etiquetado si bien es cierto demandó bastante tiempo en poder ser completado, tuvo diferentes matices en su dificultad. Además, logró poder establecer un criterio para poder estandarizar el etiquetado y que no haya irregularidades en ello.

Características y Distribución

La distribución en el conjunto de datos fue la siguiente:

|  |  |  |
| --- | --- | --- |
| Clases | Total por Clase | Probabilidad |
| USA only | 817 | 0,4085 |
| Non-USA | 579 | 0,2895 |
| World | 77 | 0,0385 |
| undetermined | 527 | 0,2635 |
| Total Conjunto | 2000 | 1 |

Tabla 1: Probabilidad y total de cada clase

Como vemos en la *Tabla 1*, la distribución en el conjunto de los datos estuvo liderada por la clase USA only, con mayor diferencia respecto de la 2da y 3ra clase, las cuales tuvieron una distribución más equiparada.

USA only, Non-USA y World a parte de tener una distribución diferente que cualquier otra clase tuvieron características específicas, propias de su clase. Además, se encontraron diversas relaciones a partir de las ubicaciones más prominentes de cada clase, estas son:

**USA only:** Dentro de las clases, la que obtuvo una mayor distribución fue USA only, y el estado que abarcó el mayor número en esta clase fue California, seguido de Texas. Para estas dos ubicaciones en particular, la particularidad que se encontró en los perfiles de los usuarios fue la siguiente:

California: Por lo general los perfiles de los usuarios de este estado, hacían mención a su ubicación con bastante énfasis, colocándolo primeramente en su descripción su estado. Frases como “I'm a teenager from CA (california girls, we're unforgetable...haha)” se pueden inferir de lo orgullosos que están del lugar donde viven.

Texas: En los perfiles de los usuarios de Texas, se encontró a bastante gente casada, con familia y marinos retirados. Además, compartían con California el énfasis de colocar en la mayoría de las veces su estado en el primer lugar dentro su descripción.

Por lo demás, se debe tomar en cuenta que California y Texas son uno de los estados más grandes y con mayor población en USA.

**Non-USA:** Para esta clase las ubicaciones con mayores usuarios y de las cuales se obtuvo una relación entre éstos fueron:

UK, England, Australia: Los perfiles del Reino Unido, Inglaterra y Australia, de manera particular no afloró una relación muy característica, pero compartían en conjunto la relación con el idioma Ingles. Estos países o conjunto de países (UK) suelen ser bastante parecidos a lo que es USA, y teniendo en cuenta que USA-only fue la clase mayoritaria se puede ver una relación entre todos éstos en el sentido de recurrir mayormente (y apoyado por el idioma) a este tipo de plataformas.

**World:** Para esta clase las ubicaciones con mayores usuarios (que no eran de USA) y de las cuales se obtuvo una relación entre éstos fueron:

Canada: Se infiere de que una de las razones por las cuales Canada es la más frecuente en la clase World es por su cercanía con USA. Al estar en la frontera y compartir además el idioma hace que sea uno de los países con mayor grado de personas que utilizan la plataforma y que se sientan arraigados a estos dos países.

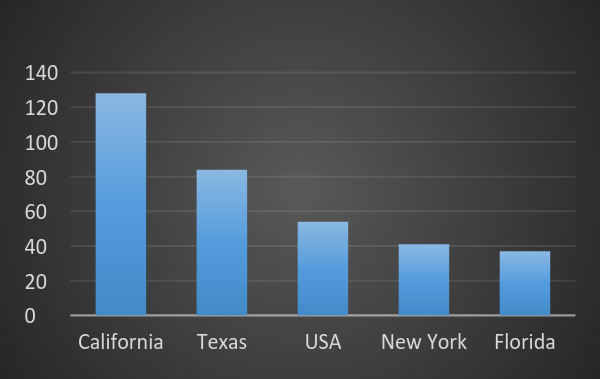
Germany: Esta ubicación fue la segunda con mayor número de usuarios, y si bien es cierto no comparten el idioma, es uno de los destinos turísticos preferidos para los estadounidenses y por ello la relación entre éstos.

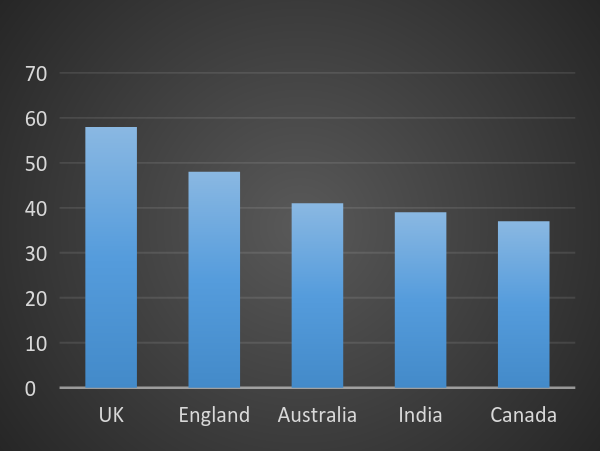
Las siguientes tablas y gráficos muestran los lugares más prominentes de cada clase, con la frecuencia y probabilidad respectiva (en su clase).

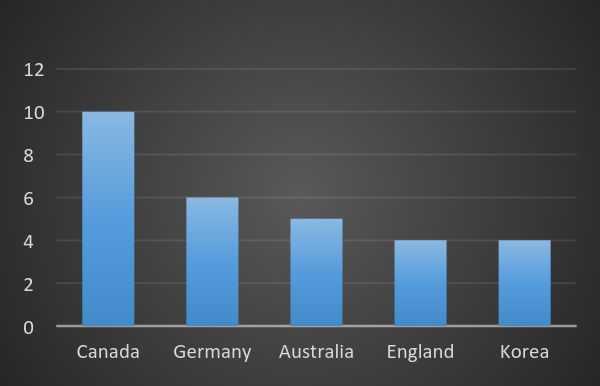
|  |  |  |
| --- | --- | --- |
| USA only |  |  |
| Ubicación | Frecuencia | Probabilidad |
| California | 134 | 0,164014688 |
| Texas | 84 | 0,102815177 |
| USA | 54 | 0,066095471 |
| New York | 41 | 0,050183599 |
| Florida | 37 | 0,045287638 |
|  |  |  |
|  |  |  |
| Non-USA |  |  |
| Ubicación | Frecuencia | Probabilidad |
| UK | 58 | 0,100172712 |
| England | 48 | 0,082901554 |
| Australia | 41 | 0,070811744 |
| India | 39 | 0,067357513 |
| Canada | 37 | 0,063903282 |
|  |  |  |
|  |  |  |
| World |  |  |
| Ubicación | Frecuencia | Probabilidad |
| Canada | 10 | 0,12987013 |
| Germany | 6 | 0,077922078 |
| Australia | 5 | 0,064935065 |
| England | 4 | 0,051948052 |
| Korea | 4 | 0,051948052 |

Tablas 2: Frecuencias y probabilidades de ubicaciones más prominentes

de cada clase.







Como se puede observar en las tablas y en los gráficos anteriores las ubicaciones más prominentes en base a su frecuencia se encuentran en USA-only, además, acompañadas de un porcentaje más alto que en las demás clases. Esto se debe a la alta frecuencia en ubicaciones como California y Texas, en contraste por ejemplo con la clase World, donde si bien es cierto hay algunas ubicaciones que son más frecuentes, éstas no generan altos porcentajes al ser muy variadas las ubicaciones de esta clase.

La probabilidad total (en todas las clases) de ciertas ubicaciones fue la siguiente:

-Tomando el caso de las más prominentes, las cuales vemos en las tablas anteriores (*Tablas 2*), su nueva probabilidad, ahora tomando en cuenta su ocurrencia en las demás clases:

|  |  |  |
| --- | --- | --- |
| Ubicación | Frecuencia | Probabilidad (Frecuencia/2000) |
| California | 143 | 0,0715 |
| Texas | 91 | 0,0455 |
| USA | 86 | 0,043 |
| New York | 52 | 0,026 |
| Florida | 38 | 0,019 |
| UK | 62 | 0,031 |
| England | 52 | 0,026 |
| Australia | 46 | 0,023 |
| India | 42 | 0,021 |
| Canada | 47 | 0,0235 |
| Germany | 12 | 0,006 |
| Korea | 4 | 0,002 |

Tabla 3: Probabilidad de ubicaciones más prominentes con respecto a todas las clases.

Como vemos, las ubicaciones que tienen una distribución mayor siguen siendo prácticamente las mismas, estas ubicaciones no solo gozan de ser frecuentes en la clase USA-only, sino que además tienen una alta frecuencia en la clase World. Además, podemos notar que las ubicaciones que no pertenecen a USA-only son ubicaciones en las cuales se parecen mucho y comparten bastantes relaciones tanto en idioma, costumbres y linaje.

Relación entre algún lugar de USA y fuera de USA

La relación que se encontró y que se infiere en los puntos anteriores, fue que efectivamente se nota una clara relación en los países de habla inglesa con respecto a USA en general. Esta relación se denota en la distribución de la ubicación de los usuarios.

Para el caso de las ubicaciones de habla inglesa más prominentes de la clase Non-USA como:

-UK

-England

-Australia

-Canada

-London

-Ireland

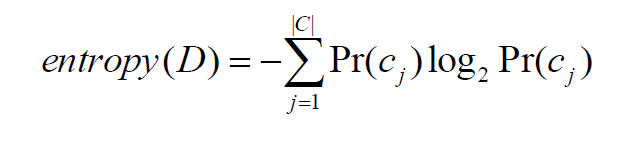
Éstas, tuvieron un porcentaje del 40.75% dentro de su clase, por lo cual vemos que de los países fuera de USA que acceden a estas plataformas en general son los países de idioma Ingles.

Además, como se explica anteriormente son países que tienen una conexión turística importante y que comparten una relación, en el caso de USA, ancestralmente importante.

Entropía en los Conjuntos de Datos

La entropía es un cálculo relacionado a la probabilidad de variables aleatorias para poder obtener su grado de incertidumbre, entre más incertidumbre necesitaremos más información, entre menos incertidumbre más puro y con mayor certeza será nuestro dato.

Para este caso ocuparemos la siguiente fórmula de cálculo de entropía para encontrar esta pureza en los datos:



Con esta fórmula calcularemos la incertidumbre y pureza de nuestros datos, para cada clase, en general y en términos de ubicación geográfica.

La Entropía en general se calcula para datos de 2 clases (o variables) diferentes, así el valor de la entropía variaría entre 0 y 1 (0% a 100%), en nuestro caso seguiremos ocupando esta fórmula teniendo siempre presente que, por definición de entropía, al tener mayor entropía conlleva más incertidumbre en los datos.

Ocupando las probabilidades de cada clase (*Tabla 1*) Los datos de entropía obtenidos son:

|  |  |
| --- | --- |
| Clases | Entropía |
| USA only | 0,527615339 |
| Non-USA | 0,517731594 |
| World | 0,180911413 |
| undetermined | 0,507006973 |

Tabla 4: Clases su Entropía.

Como observamos, la incertidumbre se mantiene regular entre USA-only, Non-USA y Undetermined, esto ya que poseen distribuciones parecidas, no así como es el caso de World, donde es menos probable, por lo tanto, más seguro que ningún perfil caerá en su clase.

Mediante estas probabilidades y ocupando la fórmula de la Entropía para la sumatoria de sus probabilidades, nuestra entropía general es:

|  |  |
| --- | --- |
| Entropía General | 1,733265319 |

Entropía por ubicación geográfica

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Clase | Ubicación | Frecuencia | Probabilidad | Entropía |
| USA only | California | 134 | 0,16401469 | 0,42776721 |
| Texas | 84 | 0,102815177 | 0,337426545 |
| USA | 54 | 0,066095471 | 0,259048295 |
| New York | 41 | 0,050183599 | 0,216624542 |
| Florida | 37 | 0,045287638 | 0,202197478 |
| Non- USA | UK | 58 | 0,100172712 | 0,33251716 |
| England | 48 | 0,082901554 | 0,297820273 |
| Australia | 41 | 0,070811744 | 0,270491483 |
| India | 39 | 0,067357513 | 0,262156607 |
| Canada | 37 | 0,063903282 | 0,253566059 |
| World | Canada | 10 | 0,12987013 | 0,382449149 |
| Germany | 6 | 0,077922078 | 0,28689538 |
| Australia | 5 | 0,064935065 | 0,256159639 |
| England | 4 | 0,051948052 | 0,221651249 |
| Korea | 4 | 0,051948052 | 0,221651249 |

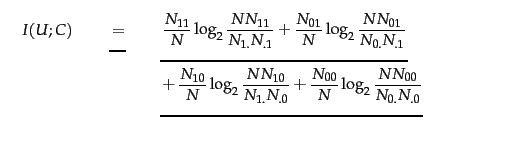
Tabla 5: Entropía por Clase y Ubicaciones

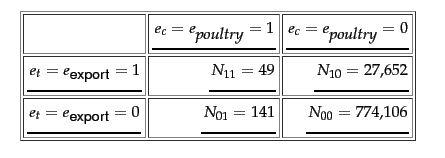
En esta tabla se encuentran las Entropías separadas por ubicación geográfica y clase. Como podemos observar de la clase USA-only la ubicación con mayor incertidumbre es California, y esto porque justamente hay más probabilidad de que algún perfil caiga en esta ubicación a futuro, no así las que tienen poca probabilidad y prácticamente “se sabe” que no caerá un nuevo perfil de usuario en ellas.

La Clase que más certeza nos da (esto por su poca ocurrencia en los datos), es la clase World, junto con ubicaciones como Korea e Inglaterra (no siendo éstas las de menor distribución en la clase). Lo que hace particular el cálculo de la entropía en esta clase es la alta variación en sus datos, y por ello la alta certeza de que éstas no se repetirán.

Información Mutua

La Información Mutua mide la reducción de la incertidumbre de una variable aleatoria X, debido al conocimiento del valor de otra variable aleatoria Y. La información mutua que calcularemos en este caso será para cada ubicación perteneciente a la clase World. Cada ubicación se calculará de manera separada dejando como la variable Y al resto de ubicaciones pertenecientes a las clases. Para el cálculo de esta información ocuparemos la siguiente fórmula:





Como muestran las imágenes anteriores, se genera en cada cuadrante una sumatoria específica de casos para poder calcular en la fórmula, Ejemplo:

|  |  |  |
| --- | --- | --- |
|  | World | Resto de Clases |
| USA | 39 | 54 |
| Resto de Ubicaciones | 183 | 1907 |

Tabla 6: Ejemplo Información Mutua

Las tablas a continuación muestran las ubicaciones (variable aleatoria X), y a su lado la información mutua de éstas con respecto a las demás ubicaciones (variable aleatoria Y).

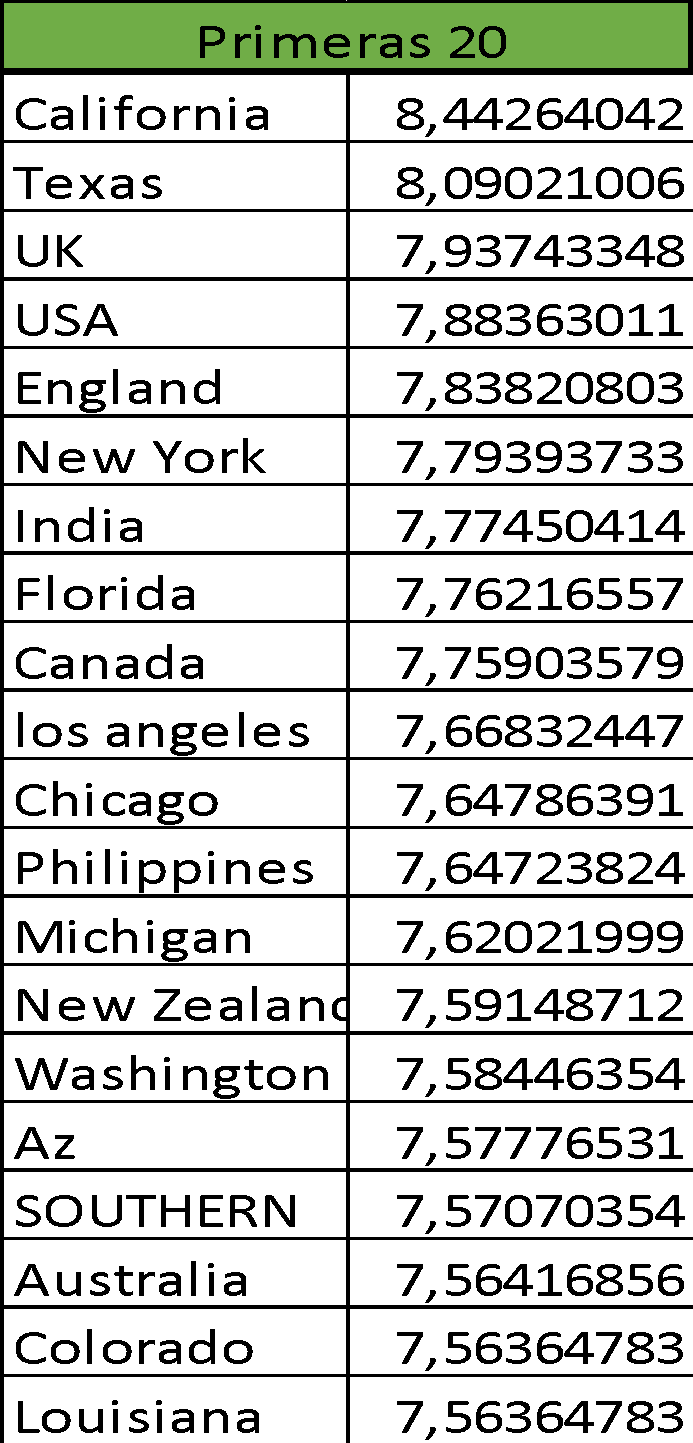
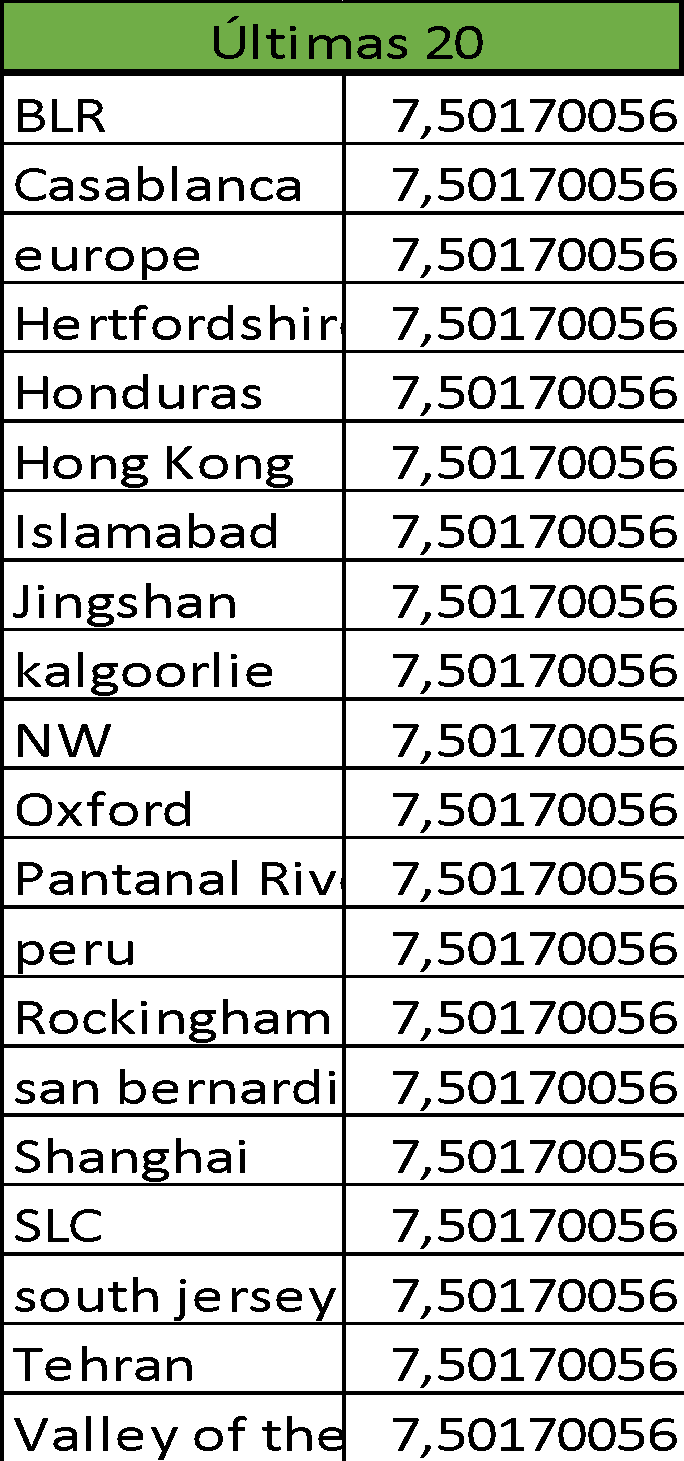


Tabla 7: Ubicaciones con mayor Información Mutua Tabla 8: Ubicaciones con menor Información Mutua

Las primeras 20: Como se denota en la *tabla 7*, las primeras ubicaciones con mayor información mutua, ósea, donde menos se ha reducido la incertidumbre (hay más incertidumbre), son justamente las ubicaciones con mayor frecuencia y probabilidad que tenemos, esto lo vemos además en la *tabla 3.*

Las últimas 20: En la *tabla 8*, las ubicaciones que se muestran son las que menor frecuencia en su totalidad tenemos del etiquetado. Es por esto, que se ve reducida mayormente su información mutua.

**Conclusión**

En consecuencia, con el resultado del etiquetado y el análisis posterior nos podemos dar cuenta que existe una tendencia hacia las ubicaciones de USA, esto ya que en el set de datos la clase más frecuente fue USA-only, además de sumarles las ubicaciones de World. Es por esto, que además vemos una alta relación con países cercanos al idioma, situación económica, costumbres y linaje como son algunos sectores del Norte de Europa (Non-USA).

Además, se puede ver que con el cálculo de la Entropía podemos observar los datos con mayor o peor certeza en nuestro análisis, teniendo la menor certeza (de su ocurrencia) en los lugares con mayor probabilidad y más frecuentados, como son los ya mencionados anteriormente.

Por otro lado, se recalca la gran importancia del buen etiquetado, esto ya que es fundamental para poder tener “pureza” en los datos que posteriormente analizaremos ya que demarcan su tendencia. Además, si hubiesen sido pocas preguntas nuestro grado de análisis sería menor ya que no serían más confiables nuestros datos por el hecho de no tomar una muestra representativa de las ubicaciones de la plataforma. Entonces, el tomar una muestra que sea representativa y la vez que marquen la tendencia sería lo ideal para nuestro análisis ya que tendríamos casos más certeros.

La recolección de datos para nosotros debe ser muy importante ya que de aquí obtenemos ciertos patrones (ubicaciones y otros), los cuales depende de cómo los desarrollemos y analicemos para poder sacarle el máximo provecho a la información implícita que éstos contienen.