Universidad Andrés Bello
Facultad de Ingeniería
Ingeniería Civil Informática

Sistemas Inteligentes

Trabajo 1 Etiquetado de Anclaje Temporal de Preguntas

Profesor: Alejandro Figueroa
Ayudante: Alexander Espina

Horario: miércoles 14:00-15:30
Jueves 12:10-13:40

Fecha de Publicación: lunes 7 de marzo de 2016
Fecha de Entrega: miércoles 30 de marzo de 2016
Lugar: Horario de clases

Aspectos Generales

- El trabajo es individual.
- La entrega del informe impreso debe ser realizada de manera presencial, en horario de clases.
- Lea atentamente las indicaciones esbozadas en el syllabus del curso.

Comunidades de Pregunta-Respuesta

Las comunidades de PreguntaRespuesta (cQA) son sitios en donde, a grandes rasgos, sus miembros responden a preguntas realizadas por otros usuarios. Algunos ejemplos de estas pueden ser StackExchange, Quora o Yahoo! Answers. Los elementos centrales de una cQA, podrían ser identificados como; preguntas, respuestas, categorías, calificaciones hechas a la calidad de las respuestas, etc. Una de las razones que hace importante el uso de estas comunidades es que los usuarios pueden obtener respuestas específicas a sus preguntas. Este tipo de inquietudes particulares es, muchas veces, difícil resolverlas en la web tradicional. En realidad, es posible realizar preguntas simples como "*¿Cuál es la capital de Chile*?" hasta aquellas que persiguen realizar una encuesta sobre algún tema específico (e.g., "*¿Cuál es el mejor restaurante de pizzas en Santiago*?").

En esta tarea nos enfocaremos en el **anclaje temporal** que pueden tener algunas preguntas. Por anclaje temporal, entendemos el periodo de tiempo durante el cual una pregunta podría capturar la atención de los usuarios de la comunidad. Por ejemplo, la pregunta *"¿Por qué el cielo es azul?*" tiene el potencial de capturar la atención de los usuarios en cualquier época del año, indistintamente. Otro tipo de preguntas, tienen la facultad de capturar mucha atención en un período corto de tiempo y posteriormente, pasan a ser ignoradas casi totalmente por la comunidad. Ejemplos de este último tipo pueden ser las consultas *"¿Cuánto cuestan los tickets para el concierto de Rolling Stones de esta noche?*" La utilidad de determinar el anclaje temporal de las preguntas se basa principalmente en la idea de evitar el fenómeno de duplicate-asking (preguntar varias veces los mismo), y para establecer vínculos entre preguntas relacionadas. Otro aspecto importante en el anclaje temporal de las preguntas es la validez o el interés en el tiempo que puedan tener sus respuestas. La misma pregunta puede tener distintas respuestas en el tiempo "*¿Quién va a ganar el Barcelona o el Real Madrid?*".

Tipo	Pregunta	Respuestas	Ejemplos
Periódicas	El interés de la pregunta renace cada cierto tiempo bien definido.	Las respuestas pueden ser reciclables. Es decir, las mismas respuestas pueden ser usadas en cada uno de estos "renacimientos" de interés.	"*How do you cook a Christmas Turkey*?", "*What are good ideas for Valentine's Day*?", "*When is Yom Kippur*?"
Ráfagas	El interés de la pregunta nace y muere abruptamente. Es decir, la pregunta captura la atención de mucha gente por un periodo corto, sin embargo, ese interés muere rápidamente.	Las respuestas de esas preguntas capturan el interés de las personas durante ese breve periodo de tiempo. Aún cuando, sus respuestas conservan validez, es muy poco probable que sean consultadas posteriormente.	"*Who killed Osama Bin Laden?*", "*Will Trump win tonight's SC primary?*", "*Why Mark Zuckerberg gave away 99% of Facebook?*", "*Did Obama killed Scalia?*", "*When will Hurricane Sandy hit NYC?*"
Permanentes	El interés de la pregunta puede surgir en cualquier instante indistintamente. El nivel de interés por el tema es constante a través del tiempo.	Las respuestas a estas preguntas podrían o no reciclarse en el tiempo.	"*How to make green beer?*", "*How do you remove acne?*", "*What is the capital city of the United States?*", "*What is the time difference between Santiago and London?*"
Ráfagas Múltiples	Se comportan igual que las de ráfagas, pero reiteradamente. Sin embargo, el periodo entre cada uno de estos eventos es indeterminable.	Las respuestas no son reciclables en el tiempo. Es decir, las respuestas dadas en una instancia anterior del evento, no van a ser útiles en la nueva ocurrencia.	*"Who will win tonight Real Madrid or Barcelona?", "Will the GOP win this election?", "Are you for or against of gun control?", "Are you pro-life or pro-abortion?", "Did Angelina Jolie and Brat Pitt get divorced?", "How much did the stock market crashed yesterday?", "How many red cards has Luis Suárez received this year?"*
Modas/Drift	El interés va en un aumento o disminución paulatina en el tiempo.	Las respuestas son reciclables, pero llega un momento en el cual alcanzaran un peak de atención, después descenderán y su consulta será muy rara.	"*How do I install Windows 8?*", "*How do I make furry nails?*", "*Do you keep your finger warm with furry nails?*"

Objetivos

La tarea apunta a confrontar al alumno con la etiquetación de un conjunto de datos, y los problemas que conlleva esta tarea. Además, a revisar conceptos aprendidos en asignaturas anteriores como el de probabilidad. Sumado a ésto, se pondrán en práctica conceptos que serán discutidos en clases como el de entropía.

Desarrollo

A cada estudiante se le asignará un número 650 de preguntas donde deberá identificar el anclaje temporal de cada una de ellas. Este conjunto estará dividido en 390 preguntas en inglés, y 260 preguntas en español. Cada una de las preguntas debe ser asociada con alguna de las categorías anteriormente esbozadas. En caso de no poder relacionarlo con ninguna de las categorías, el alumno puede asignarlo a una clase "OTROS". Nótese que a cada pregunta debe asignársele sólo un tipo de anclaje. El alumno deberá entregar un informe donde:

- Describe extensamente su experiencia de etiquetado: ¿Qué observó? ¿Qué dificultades encontró? ¿Qué observa de las preguntas que fueron en la clase "OTROS"? Discuta esto en extenso y ejemplificando sus observaciones.
- ¿Qué características tienen las preguntas relacionadas con cada tipo de anclaje? ¿Cuál es la distribución en su conjunto de datos? ¿Cuál es la probabilidad de cada tipo de anclaje?
- ¿Hay relación entre la categoría de una pregunta y su anclaje? ¿Entre el número de seguidores? ¿Qué sucede con el número de respuestas?
- Calcule la entropía de sus conjuntos de datos, para cada idioma, de manera general y para cada categoría.

Nótese que debe entregar un informe conforme al esquema descrito en el Syllabus del curso, es decir con introducción, descripción del problema, conclusiones, etc.

Formato de Archivo

El alumno recibirá dos archivos, uno por cada idioma, que tienen el siguiente formato:

20150925133057AAs4Ij6 396545451#Environment 2115500306#Global Warming 0 following 23 answers <title>Is it necessary for the current warming to be unusual?</title> <body>Is that one of the evidences of support for AGW? < br/>

 Along the same lines, does it need to be warmer now that in the past 1000, 2000, 5000 or 10000 years as well?</body>

20160108021923AAlelZU 396545144#Beauty & Style 396546060#Skin & Bod y 396547156#Tattoos 0 following 12 answers <title>Do you think it's a stupid tattoo idea?....?</title> <body>Okay so I'm a 16 year old girl (17 in two months) and my favorite band is Metallica...I love all of their stuff but my favorite line from one of their songs is "Live is ours, we live it our way..." I want to get that tattooed on my arm when I'm 18, but my parents hate the idea! They both love the band and song but they don't like the idea of putting that tattoo on my body. The song (Nothing Else Matters) is very meaningful to me and I absolutely love the quote (I have an idea of what it stands for In my head) so should I just say f*** it and not care what they think...after all It is MY body and they don't have much control over that. I was just wondering if you guys think it's a stupid idea as well...</body>

Cada línea del archivo representa una pregunta. Después vienen de cinco a siete columnas separadas por un tabulador, en donde la primera es el identificador de la pregunta que está compuesto por la fecha que se hizo la pregunta y un identificador del usuario. La segunda columna corresponde a la categoría de primer nivel de la pregunta. Todas las preguntas pertenecen a una categoría de primer nivel. Si la línea tiene seis columnas, la tercera corresponde a una sub-categoría, y si la línea tiene siete columnas, la cuarta corresponde a una sub-sub-categoría. Las últimas dos columnas representan el título de la pregunta y su cuerpo. Las otras dos columnas indican el número de respuestas que tiene la pregunta, y los seguidores. Muchas veces, los usuarios no ingresan el cuerpo de la pregunta, esto se indica con un “<body>
</body>”.

Entregable

El alumno debe entregar dos archivos, uno para cada idioma, con el siguiente formato:

Qid etiqueta

Es decir, dos columnas donde la primera es el identificador único de la pregunta (e.g. 20150925133057AAs4Ij6) y la segunda es la etiqueta (i.e., “PERIODICA”, “RAFAGA”,”PERMANENTE-R”,”PERMANENTE-NR”,”M-RAFAGA”,”DRIFT”,”OTROS”). Nótese que la segunda columna debe ajustarse a los valores descritos de manera estricta. Ambas columnas deben estar separadas por un tabulador (i.e., “\t”). Junto a los dos archivos, debe ser entregado un informe impreso.