**Tarea 1:**

**Anclajes en preguntas**

**Índice**

|  |  |
| --- | --- |
| Introducción | Pág. 2 |
| Descripción del estudio | Pág. 3 |
| Presentación y Análisis de resultados | Pág. 4 |
| Anexos | Pág. 7 |
| Conclusión | Pág. 13 |

**Introducción**

En las comunidades de preguntas y respuestas (cQA), son sitios donde a grandes rasgos, sus miembros responden preguntas realizadas por otros usuarios, donde estos poseen distintas motiviaciones, pero en donde las preguntas poseen una caracteristica esencial que es el periodo, donde la pregunta podria capturar el interes de otros usuarios con similar inquietud.

Para este trabajo, se clasificaron las preguntas conforme a sierte estandares: periodicas, ráfagas, ráfagas múltiples, permanentes reciclables, permanentes no reciclable, moda y otros (cuando no se se estuviera seguro a cual correspondiera).

En el presente trabajo se clasifico 650 preguntas, extraidas de Yahoo! Answer, para posteriormente realizar un estudio sobre las clasificaciones de las preguntas. La idea central es obtener caracteristicas del tipo de pregunta asociada a la clasificaicon mencionada.

**Descripción del Estudio**

En el presente trabajo se clasificaron 650 preguntas desde la base de datos de Yahoo! Answers, analizando el periodo de tiempo en el cual la pregunta podria capturar el intereses de los usuarios. Los tipo de preguntas pueden ser periódicas, ráfagas, permanentes, ráfagas múltiples, modas y otros.

Las preguntas tipo periódicas son aquellas que su interés renace cada cierto tiempo, por ejemplo las preguntas relacionadas a fiestas conocidas, navidad, año nuevo, etc, las preguntas tipo ráfagas poseen una subida y bajada abrupta, siendo temas específicos del periodo que se está viviendo. La pregunta logra captar la atención por un periodo corto. También las preguntas de este estilo pueden ser múltiples, que poseen el mismo comportamiento pero de manera reiterada, aunque, el periodo entre estas es indeterminable. También estas las preguntas permanentes, estas, pueden surgir en cualquier instante y el nivel del tema es constante a través del tiempo. Este tipo de preguntas poseen dos categorías, permanentes reciclables y no recíbales. Las …. Y el último tipo de pregunta trabajado son las modas, cuyo interés va en aumento o disminución en el tiempo. Cuando no existiese seguridad respecto que tipo de pregunta corresponde, se etiqueto como otras.

Las preguntas a menudo pueden adquirir un forma compleja de clasificación, dado que algunas preguntas como: “*¿necesito que me orienten sobre esto... por favor se los pido de corazon?”* no existe interrogante a clasificar dado un periodo de tiempo, o que no es posible utilizarla en un tiempo más.

**Presentación y análisis de resultados**

## Experiencia al clasificar

En primera instancia se observó que el anclaje de las preguntas tendió en mayor frecuencias a las de tipo permanentes reciclables, confirmado por la **Tabla 1**. Tipo que posee más de la mitad de las preguntas clasificas.

Se observó que las preguntas que poseían una estructura bien formada tienden a no ser clasificadas como otros, dado a que les da a la pregunta un plus o un grado de seriedad, por ejemplo respecto a las preguntas en español, donde la estructura básica es, llevar los símbolos: ‘¿’ y ‘?’, para comenzar y terminar una pregunta, después del símbolo viene un pronombre en mayúscula, después viene: el sujeto, pronombres personales y para terminar la interrogante el complemento. Pero hay excepciones, por ejemplo la pregunta *“¿si como caca de perro seré virgen de nuevo?”*, no tiene sentido lógico, no posee un cuerpo, dado esto su anclaje es otros. Por lo tanto cuando las preguntas no tenían la intención de solicitar información dado a un cierto periodo, y/o no fueran de utilidad para otras personas, fueron asignadas con el anclaje otros.

En el momento del etiquetado hubieron preguntas que no se encontraba la interrogante a resolver, o sea, que más que preguntas eran comentarios. Por ejemplo la pregunta *“¿Le dije a un chico que me gusta por facebook?* no existe un periodo en específico para clasificarla como ráfaga, periódica, etc, en donde se volviera a utilizar, pero, como pregunta tampoco es posible clasificarla como utilidad, para decir que su anclaje pudiera ser permanente reciclable, dado que los usuarios del a internet no obtendrán información de utilidad. Ahora bien dado los argumentos mencionados, se puede decir que su anclaje será “otros”, pero, dado que la pregunta posee un cuerpo (donde el usuario puede dar a conocer detalles de la pregunta), “*Que hago! primero le dije por un face falso que no tenia foto de perfil ni nada y luego una amiga me convencio de decirle por mi face real y lo ise pero no me responde nosotros solo somos conocidos ni nos hablamos ni nada pero como creen que el reaccione?”* , dado al cuerpo de la pregunta, especifica cual es la problemática de esta, dado a los detalles entregados. Dado que podrían existir usuarios con una situación similar, fue clasificada como permanente reciclable.

Al momento de etiquetar, fue posible identificar una características de las preguntas que permitió identificarlas como periódicas, las fechas importantes. Por ejemplo, navidad, año nuevo, Halloween. Dado que son preguntas que renacen cada cierto tiempo y a su vez son específicas, ejemplo: *“Are NBA players paid extra to play on Christmas?”*, el usuario desea obtener información sobre algo específico, pero en una fecha específica, que cada cierto tiempo renacen.

En el transcurso del etiquetado aparecieron preguntas tipo “políticas”, en donde los usuarios, deseaban obtener más información al respecto y/o desean expresar su opinión al respecto. En la **tabla 2**, las preguntas con categoría “Politics &amp; Govermment”, poseen el 76% de las preguntas con anclaje ráfaga. A este tipo de preguntas (anclaje) se les suma la categoría “News &am; Events” con el 50% de anclaje ráfaga.

En el etiquetado, se observó una complicación que se lidio con ella en todo el proceso. El problema era que el clasificante no conocía ciertos temas, ya que aparecían siglas de alguna empresa, siglas que incluso en ingles no son conocidas, algunos temas de materia específicas. Complicó el etiquetado ya que no era posible de primera instancia entender el contexto de la pregunta, por lo que dificultada la opción de anclaje, y fue posible apreciar que algunas fueron clasificadas como otros. Por ejemplo:



Imagen 1: Pregunta dificultosa.

Para una persona sin conocimientos de YouTube, no sería posible saber quién es este personaje (youtuber), lo más cercano o global conocido, podría ser un personaje de la sería un persona de la serie Dragon Ball. La mejor respuesta a esta pregunta, es la mostrada en la imagen, que no es posible responder si no se tiene un conocimiento del tema.

# Características

Para el dato obtenido de la **tabla 6**, nos permite entender que los datos obtenidos no poseen un sesgo respecto a la categoría, si no que están dispersos en las distintas categorias. Esto desde un punto de vista a estudiar es bueno, dado a que si la entropía fuera 0, no se podría realizar un estudio al respecto. Algo similar ocurre con los datos de la **tabla 5**, a que cada idioma posee gran diversidad de categorías, que permite realizar el estudio.

De la **tabla 7,** la categoría que posee una menor incertidumbre es “Business &amp; Finance” dada que posee una menor entropía entre todas las categorías. Dado que existe mayor probabilidad que exista una pregunta con anclaje permanente, no solo porque este anclaje posee la mayor frecuencia sino también porque existen cuatro tipos de anclaje que poseen frecuencia 0, por lo tanto la muestra obtenida es más limpia. Pero caso contrario ocurre con la categoría sport, que posee una entropía de 1.58, siendo la más alta, y se puede deber porque no es sesgada en comparación a la otra, dado que solo existe un anclaje con frecuencia cero.

Se detallaran las características encontradas al etiquetar:

* **Periódicas**: las características más apreciadas en el etiquetada ya fueron mencionadas, donde era posible apreciar que las fechas específicas. En la **tabla 4**, se hace una comparación entre el anclaje v/s el número de respuestas, se puede decir que las respuestas no son tantas como en otro casos, pero esto se puede deber a que como este tipo de preguntas nacen cada cierto tiempo, también solo serán respondidas cada cierto tiempo y como no son preguntadas o solicitadas fuera de un periodo indicado, no serán respondidas. También en la **tabla 3**, es posible apreciar lo indicado, poseen pocos seguidores y se puede deber a lo mismo. Lo mismo sucede con la **tabla 1,** que nos muestra la probabilidad que aparezca una pregunta periódica que es 2.15385%, numero bastante bajo, que se puede deber a lo ya mencionado.
* **Ráfagas:** en este tipo de anclaje es apreciable que donde posee más cantidad de preguntas es en la categoría “Politics &amp; Governmment” y “News &amp; events”, dadas a las razones ya fueron dichas anteriormente. Donde la catidad de respuestas entre ambas categorías son (**grafico 2**) 548, número que corresponder a 71.9% del número de respuestas de este tipo de anclaje (**tabla 4**).
* **Ráfagas Múltiples:** dado a la poca cantidad de preguntas con las que se pudo trabajar no fue posible obtener una gran cantidad de caracteristicas, un ejemplo de este tipo fue: *“Why do people say Marilyn Monroe was miserable when she was beautiful and famous?”***,**  donde el internet de la pregunta nacio por alguna razon y muere abruptamente, pero esto podria volver a ocurrir. Pero es intersante que a pesar que de la poca cantidad de preguntas posee una gran cantidad de respuestas, eso se puede deber lo mencionado, que, estas preguntas pueden reutilizarse en el tiempo.
* **Permanentes Reciclable:** para este tipo de anclaje es posible apreciar que es la que posee mayor cantidad de preguntas, respuestas. También en la **tabla 2**, se aprecia que es el anclaje en donde siempre posee el número máximo de preguntas por categoría, por lo tanto podemos decir que este tipo de anclaje puede estar en cualquier tipo de tema. Dado a que los usuarios de internet pueden estar solicitando información de cualquier tipo en cualquier instante, por lo tanto existen muchas preguntas que puede ser reutilizables.
* **Permanentes No Reciclables:** no fue posible obtener características de este anclaje, dado que las preguntas asignadas fueron 2, siendo sus seguidores 7, y el número de respuesta 10.
* **Modas:** este tipo de anclaje tiene una cualidad bastante interesante, que a pesar de la poca cantidad de preguntas (**tabla 2**)y de los pocos seguidores que posee, el número de respuestas es bastante alto. Si obtuviéramos un promedio de respuestas por preguntas, seria aproximadamente 11. Una de este tipo de preguntas fue: “*Xbox one or Playstation 4?”,* estas consolas de videos juegos fueron totalmente en un aumento desde el momento que mencionaron su estreno, pero al pasar el tiempo estas han ido disminuyendo su interés, pero de una forma paulatina, no estrepitosamente como en otros casos. De esta forma es una característica principal de las preguntas tipo moda, siendo que el interés va en aumento o disminución paulatina en el tiempo.
* **Otros** ya se mencionó unas características encontradas para este tipo de preguntas. Preguntas que no tienen sentido lógico, preguntas que son incompletas o que no poseen sentido tanto la pregunta como tampoco el cuerpo de esta: *“¿No quiero estudiar ni trabajar?”*. He incluso se aprecia que este tipo de preguntas poseen gran número de respuestas, se puede deber a que como la pregunta no requiere un conocimiento del tema, no es necesario que cierta gente pueda contestar (dada la exigencia de conocimiento requerida).

**Anexos**

**Probabilidades**

| PERIODICA | 2.15385% |
| --- | --- |
| RAFAGA | 10.1538% |
| PERMANENTER | 56.4615% |
| PERMANENTE-NR | 0.307692% |
| M-RAFAGA | 1.23077% |
| DRIFT | 4.76923% |
| OTROS | 24.9231% |

Tabla 1: Probabilidad de cada tipo de anclaje.

**Relaciones**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **PERIODICA** | **RAFAGA** | **PERMANENTER** | **PERMANENTE-NR** | **M-RAFAGA** | **DRIFT** | **OTROS** |
| Arts &amp; Humanities | 0 | 1 | 15 | 0 | 0 | 1 | 8 |
| Business &amp; Finance | 0 | 3 | 19 | 0 | 0 | 0 | 3 |
| Consumer Electronics | 0 | 4 | 14 | 0 | 1 | 4 | 2 |
| Education &amp; Reference | 1 | 1 | 15 | 0 | 0 | 0 | 8 |
| Entertainment &amp; Music | 0 | 3 | 12 | 0 | 1 | 1 | 8 |
| Health | 0 | 0 | 19 | 0 | 0 | 0 | 6 |
| Games &amp; Recreation | 0 | 3 | 8 | 0 | 1 | 5 | 8 |
| Science &amp; Mathematics | 2 | 0 | 14 | 0 | 0 | 0 | 9 |
| Beauty &amp; Style | 1 | 2 | 12 | 0 | 0 | 1 | 9 |
| Sports | 1 | 5 | 13 | 1 | 0 | 2 | 3 |
| Social Science | 0 | 2 | 12 | 0 | 0 | 0 | 11 |
| Cars &amp; Transportation | 0 | 1 | 21 | 0 | 0 | 0 | 3 |
| Dining Out | 2 | 0 | 12 | 0 | 0 | 2 | 9 |
| Food &amp; Drink | 0 | 1 | 16 | 0 | 0 | 1 | 7 |
| Home &amp; Garden | 2 | 2 | 17 | 0 | 0 | 1 | 3 |
| Local Businesses | 2 | 0 | 17 | 0 | 0 | 2 | 4 |
| Family &amp; Relationships | 0 | 0 | 16 | 0 | 0 | 1 | 8 |
| News &amp; Events | 1 | 12 | 5 | 0 | 0 | 0 | 7 |
| Pets | 0 | 0 | 14 | 0 | 0 | 1 | 10 |
| Politics &amp; Government | 0 | 14 | 5 | 0 | 0 | 2 | 4 |
| Environment | 0 | 4 | 12 | 0 | 0 | 2 | 7 |
| Society &amp; Culture | 0 | 4 | 11 | 0 | 2 | 1 | 7 |
| Travel | 2 | 1 | 14 | 0 | 0 | 3 | 5 |
| Computers &amp; Internet | 0 | 1 | 20 | 0 | 2 | 0 | 2 |
| Pregnancy &amp; Parenting | 0 | 0 | 17 | 1 | 0 | 1 | 6 |
| Yahoo Products | 0 | 2 | 17 | 0 | 1 | 0 | 5 |

Tabla 2: Relación entre categoría y su anclaje.

Grafico 1: Categorías vs Seguidores.

|  |  |
| --- | --- |
| Anclaje | Seguidores |
| PERIODICA | 4 |
| RAFAGA | 25 |
| PERMANENTER | 65 |
| PERMANENTE-NR | 0 |
| M-RAFAGA | 7 |
| DRIFT | 3 |
| OTROS | 83 |

Tabla 3: Anclaje vs Seguidores.

Grafico 2: Categorías vs N Respuestas.

|  |  |
| --- | --- |
| Anclaje | N Respuestas |
| PERIODICA | 139 |
| RAFAGA | 762 |
| PERMANENTER | 3257 |
| PERMANENTE-NR | 10 |
| M-RAFAGA | 229 |
| DRIFT | 354 |
| OTROS | 1915 |

Tabla 4: Anclaje vs N respuestas.

**Entropías**

|  |  |
| --- | --- |
| **Idioma** | **Entropía** |
| ESPN | 1.73486 |  |
| ENG | 1.54655 |  |

Tabla 5: Entropía por idioma.

| ENTROPIA DE MANERA GENERAL | 1.73264 |
| --- | --- |

Tabla 6: Entropía de Manera general.

| **Categorías** | **Entropía** |
| --- | --- |
| Arts &amp; Humanities | 0.65424 |  |
| Business &amp; Finance | 0.530425 |  |
| Consumer Electronics | 1.25933 |  |
| Education &amp; Reference | 0.766381 |  |
| Entertainment &amp; Music | 1.11318 |  |
| Health | 0.397258 |  |
| Games &amp; Recreation | 1.33689 |  |
| Science &amp; Mathematics | 0.812473 |  |
| Beauty &amp; Style | 0.977641 |  |
| Sports | 1.58635 |  |
| Social Science | 0.577697 |  |
| Cars &amp; Transportation | 0.434322 |  |
| Dining Out | 1.04917 |  |
| Food &amp; Drink | 0.644284 |  |
| Home &amp; Garden | 1.0256 |  |
| Local Businesses | 0.993313 |  |
| Family &amp; Relationships | 0.571168 |  |
| News &amp; Events | 0.999409 |  |
| Pets | 0.587593 |  |
| Politics &amp; Government | 0.945918 |  |
| Environment | 0.857425 |  |
| Society &amp; Culture | 1.25244 |  |
| Travel | 1.15332 |  |
| Computers &amp; Internet | 0.898609 |  |
| Pregnancy &amp; Parenting | 1.04122 |  |
| Yahoo Products | 0.88804 |  |

Tabla 7: Entropía para cada categoría.