Solemne I – 1era Parte: Inteligencia Artificial

***Profesor:* Alejandro Figueroa**

***Ponderación:* 1**

***Fecha de entrega:* viernes 22 de Agosto a las 23:59 hrs (Chile continental).**

***Ayudante:* Nicolás Olivares**

***Método de entrega:* mail al ayudante (nicolivares@gmail.com)**

Descripción del Problema

Hoy en día nos encontramos todo el tiempo conectado al Internet mediante diferentes medios (e.g. computadores de escritorio, teléfonos móviles, y tablets). Principalmente, vemos a la Web como una fuente de recursos y servicios, cuyo potencial es el de satisfacer nuestras necesidades de información y de interacción social. Por ejemplo, nos dirigimos a nuestro outlet de noticias favorito para leer acerca de los últimos acontecimientos noticiosos del país o del mundo, también nos informamos de las noticias tecnológicas, farándula, salud, etc. En cambio, las redes sociales las utilizamos como recurso para compartir y diseminar información, tips, nuestras opiniones, como también para compartir otros recursos como vídeos y fotos. Una clase de servicio que está a medio camino de ser un recurso de información y de interacción social son los sitios de pregunta-respuesta. En ellos encontramos plataformas que nos permiten satisfacer necesidades mucho más específicas, i.e. un usuario tiene una pregunta, para la cual necesita una respuesta. Normalmente, las respuestas a las preguntas emitidas en estas plataformas no son fácilmente encontrables en Internet, es decir, son preguntas que cuya resolución involucra el procesamiento de diversas fuentes de información, los conocimientos de un experto, o bien simplemente, el emisor no tiene tiempo para encontrar una respuesta fidedigna en otro sitio de la Web.

Ya sea para las redes sociales, noticias o un sistema de pregunta-respuesta, una pieza fundamental es identificar "named entities", si uno pretende obtener valor agregado de la información provista por los usuarios. Por ejemplo, ver si la pregunta fue hecha anteriormente, ergo si hay respuestas que puedan ser entregadas al usuario en el momento de la emisión de la pregunta. Esto disminuiría el tiempo de espera que debe incurrir el usuario hasta que otro usuario de la comunidad le de una respuesta satisfactoria.

¿Qué son "*named entities*"? Son nombres que se utilizan para representar un referente. Por ejemplo, los nombres "*Ford*" y "*Ford Motor Company*" se utilizan para indicar el referente "*la compañía creada por Henry Ford en 1903*". Es decir, para un mismo referente podemos tener nombres distintos, que trabajan como sinónimos. Hay diversos tipos de "*named entities*", pero en el ámbito de esta tarea nos preocuparemos de cuatro clases: **organizaciones**, **personas** y **ubicaciones**. Todo lo que no caiga en esta tres clases es denominado "**token**", por ejemplo puntuación, preposiciones, sustantivos, verbos que no son parte del nombre de una entidad. Merece la pena rescatar el hecho de que las "named entities" son fundamentales porque textos como las noticias, así como también muchas preguntas en una plataforma de pregunta-respuesta, tratan acerca de entidades. Entonces su reconocimiento facilita establecer la relación entre una pregunta/búsqueda y una noticia/respuesta, también la relaciones entre diversos documentos.

En esta tarea vamos a considerar el caso especial de plataforma de pregunta-respuesta Yahoo! Answers. Aquí los miembros forman una comunidad, donde cada uno de ellos puede emitir una pregunta y esperar (generalmente hasta cuatro días) para que los otros miembros de la comunidad le provean de respuestas. Finalmente, el emisor de la pregunta puede escoger la que a su parecer es la mejor respuesta. Hay muchas facetas que comentar acerca de este tipo de plataformas, sin embargo para esta tarea consideraremos que 1) no todas las respuestas provistas a

una pregunta son legítimas, por ejemplo hay propaganda, respuestas engañosas, doble sentido, chistes, etc.; y 2) que cada pregunta está compuesta de tres partes un título que normalmente plantea el objetivo de la pregunta, un contenido que provee de detalles adicionales que deberían ayudar a la resolución de la pregunta, y finalmente una secuencia de respuestas ordenadas cronológicamente.

Considere la pregunta "*did the betrayal of king edward cause ww11 ?*" que provee como contenido los detalles "*betrayed by parliament and the church , was it treason or the influence of Nazi spyrings ?*". Durante el período la pregunta que se mantuvo abierta, es decir, que se les permitió a los otros miembros responder, se obtuvieron las siguientes respuestas:

Tiempo (epoch)	Snippet de Respuesta
1402778444000	No-one ` betrayed ' King Edward at all - if you mean Edward VIII - later the Duke of Windsor . He quit . He also happened to be a Nazi sympathiser , not very bright , very selfish and a womaniser . WWII was caused by Hitler 's desire to rule the world .
1402779238000	I think you have bought the cover story this is propaganda , absolutely opposite the truth , the man was trying to rebuild the western alliance in the threat of rising german aggression .
1402780098000	no the Zionists got rid of Edward 8th because he liked Hitler Hitler angered the Zionist bankers for creating an alternative economy based on labor and ditching their Central banking scam Hitler did n't want war they - the bankers did and its happening again with Putin and Russia - Putin has ditched central banking
1402788524000	Oh do shut the fcuk up .
1402830862000	No americans cause WW2 by financing Hitler from 1924 in 1932 the Nazis were Broke and could not raise the money to contest the 1933 elections the Duponts used JP Morga n to collect the money from FDR Lindberg Prescott Bush Standard Oil GM ITT Ford IBM Bendix Cocoa Cola Birds eye all helped raise 840 Million US dollars more than enough to run in the 1933 elections and to Buy enough seats to form a coalition that gave the Nazis the 51 % needed to get Hitler elected to Chancellor No Hitler Fund No Hitler as Chancellor No WW2 this link proves i am telling the Truth https:\/\/ www.google.com.au\/#q=americans+who+funded+the+nazis

Objetivo de la Primera Parte de la Solemne 1

Nos enfocaremos en el etiquetado manual de un conjunto de tripletas <título, contenido, respuestas> extraídas de Yahoo! Answers. En el etiquetado, el alumno debe marcar las palabras que conrresponden a cada una de las cuatro clases. Sin embargo, para agilizar este proceso los "tokens" no deben anotarse. En el ejemplo anterior, tendríamos el siguiente título de pregunta: "*did the betrayal of king/PERSON edward/PERSON cause ww11 ?*", y su contenido respectivo está dado por el texto "*betrayed by parliament and the church , was it treason or the influence of Nazi spyrings ?*". En cuanto a la secuencia de respuestas, tenemos:

Tiempo (epoch)	Snippet de Respuesta
1402778444000	No-one ` betrayed ' **King/PERSON Edward/PERSON** at all - if you mean **Edward/PERSON VIII/PERSON** - later the **Duke/PERSON of/PERSON Windsor/PERSON** . He quit . He also happened to be a Nazi sympathiser , not very bright , very selfish and a womaniser . WWII was caused by **Hitler/PERSON** 's desire to rule the world .
1402779238000	I think you have bought the cover story this is propaganda , absolutely opposite the truth , the man was trying to rebuild the western alliance in the threat of rising german aggression .
1402780098000	no the Zionists got rid of **Edward/PERSON 8th/PERSON** because he liked **Hitler/PERSON Hitler/PERSON** angered the Zionist bankers for creating an alternative economy based on labor and ditching their Central banking scam **Hitler/PERSON** did n't want war they - the bankers did and its happening again with **Putin/PERSON** and **Russia/LOCATION** – **Putin/PERSON** has ditched central banking
1402788524000	Oh do shut the fcuk up .
1402830862000	No americans cause WW2 by financing **Hitler/PERSON** from 1924 in 1932 the Nazis were Broke and could not raise the money to contest the 1933 elections the **Duponts/PERSON** used **JP/ORGANIZATION Morga n/ORGANIZATION** to collect the money from **FDR/ORGANIZATION Lindberg/ORGANIZATION Prescott/ORGANIZATION Bush/ORGANIZATION Standard/ORGANIZATION Oil/ORGANIZATION GM/ORGANIZATION ITT/ORGANIZATION Ford/ORGANIZATION IBM/ORGANIZATION Bendix/ORGANIZATION Cocoa/ORGANIZATION Cola/ORGANIZATION Birds/ORGANIZATION eye/ORGANIZATION** all helped raise 840 Million US dollars more than enough to run in the 1933 elections and to Buy enough seats to form a coalition that gave the Nazis the 51 % needed to get **Hitler/PERSON** elected to Chancellor No **Hitler/PERSON** Fund No **Hitler/PERSON** as Chancellor

	No WW2 this link proves i am telling the Truth https:\\ www.google.com.au\/#q=americans+who+funded+the+nazis

Adicionalmente, para facilitar el proceso de etiquetado de entidades se ha provisto de un tag "<suggestions>" que contiene sugerencias. Estas contienen errores, ya sea que faltan palabras que son parte de una entidad, o hay palabras que realmente no pertenecen a una entidad. Simplemente están, con el objetivo de uniformar criterios y proveer una ayuda para el caso de haber ambigüedad. Nótese que las anotaciones dentro de este tag son sugerencias, no la respuesta a la tarea de etiquetado.

Cada estudiante debe solicitar al ayudante, un archivo "tar" que contiene las tripletas a ser etiqeutadas. Este archivo "tar" consiste en un conjunto de 26 archivos más pequeños, cada uno correspondiente a una de las 26 categorías diferentes de preguntas en Yahoo! Answers. Cada archivo contiene preguntas de la categoría respectiva.

Id	Nombre	Id	Nombre	Id	Nombre
396545012	Arts & Humanities	396545451	Environment	396545444	Politics & Government
396545144	Beauty & Style	396545433	Family & Relationships	396546046	Pregnancy & Parenting
396545013	Business & Finance	396545367	Food & Drink	396545122	Science & Mathematics
396545311	Cars & Transportation	396545019	Games & Recreation	396545301	Social Science
396545660	Computers & Internet	396545018	Health	396545454	Society & Culture
396545014	Consumer Electronics	396545394	Home & Garden	396545213	Sports
396545327	Dining Out	396545401	Local Businesses	396545469	Travel
396545015	Education & Reference	396545439	News & Events	396546089	Yahoo! Products
396545016	Entertainment & Music	396545443	Pets		

Requerimientos para el Informe

Una vez etiquetadas todas las tripletas contenidas en las 26 categorías, tanto el título como su contedido, y las respuestas, el alumno debe entregar un informe que responda las siguientes preguntas:

1. Combinando las 26 categorías, haga un histograma de las clases de entidades asignadas manualmente a las palabras, es decir la frecuencia de los tres tipos de entidades. Para ésto, también considere los "tokens". La suma de los cuatro valores obtenidos debe darle el número de palabras en contenidas en los 26 archivos.
2. Haga lo mismo que 1) pero para cada una de las 26 categorías por separado. ¿Qué observa? ¿Cómo podría explicar lo observado?
3. Compare las etiquetas asignadas manualmente y las sugerencias: ¿Cuántas palabras que no fueron sugeridas como entidad, pertenecian realmente a entidades? ¿En este ultimo punto, se ve alguna clase más afectada que las otras?. Y al revés, ¿Cuántas palabras fueron sugeridad como entidad, pero realmente no lo eran? ¿Hay alguna clase particularmente más afectada?
4. El mismo análisis que en 3), ahora desarrollelo por categoría. ¿Hay alguna de las 26 categorías más propensas a uno de los dos tipos de errores que se mencionan en los puntos 3 y 4?
5. Suponiendo que las etiquetas manuales son "la verdad absoluta", calcule la accuracy, precision, recall y F1-score de las sugerencias combinando las 26 categorías.
6. Realice lo mismo del punto 5 pero para cada una de las 26 categorías por separado.
7. Calcule la entropía del conjunto de datos, y de cada categoría en particular, utilizando las etiquetas manuales.

Finalmente, entregue sus conclusiones generales sobre los resultados obtenidos. ¿Qué cree que sucedería si realizamos un análisis similar sólo en los títulos de las preguntas? ¿Para qué categoría visualiza que son más útiles las sugerencias? ¿Y las entidades en general? ¿Qué categoría se ve más afectada cuando una pregunta y una respuesta coinciden en ambos aspectos de una entidad, tanto en tipo como en las palabras?.

Además, tenga en cuenta que:

1. **El alumno deberá realizar una presentación de a lo más diez minutos para mostrar los resultados obtenidos más destacables. La presentación está orientada a corregir errores tempranos y no repercutan en las tareas posteriores. Las presentaciones son en horario de ayudantía.**

2. **Cada estudiante debe trabajar sobre su propio conjunto de datos. El utilizar los datos de otro compañero automáticamente le hace acreedor de la nota 1 en la tarea, y deberá tener sus propios datos etiquetados para las partes posteriores. Cada estudiante debe solicitar su conjunto personal de datos al ayudante.**

3. **Para que su tarea sea válida, el alumnos debe entregar sus datos etiquetados junto con el informe de la tarea. De faltar los datos, el alumno obtendrá nota 1.**