# Solemne I – 1era Parte: Inteligencia Artificial

***Profesor:* Alejandro Figueroa**

***Ponderación:* 1**

***Fecha de entrega:* viernes 22 de Agosto a las 23:59hrs (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. Examine si en una página de respuestas, es decir, una tripleta <título, contenido, respuestas> hay discrepancias para la clase dada manualmente a una palabra determinada. ¿Qué observa? ¿Qué aprende de ésto? ¿Hay una clase más suceptible a las discrepancias?
2. Los nombre propios en inglés parten, en su gran mayoría, con mayúsculas. Es esperable que las entidades estudiadas en esta tarea también partan con mayúsculas. ¿Qué observa sobre el uso de las mayúsculas en Yahoo! Answers?. ¿Tienen el mismo uso en el título, contenido y respuestas? ¿Qué sucede con su uso en las "named entities" seleccionadas? ¿Qué puede corroborar acerca del uso de la mayúscula y los "tokens"? De las palabras que comienzan con mayúsculas ¿cuántas fueron reconocidas como "named entities" y cuántas como "tokens"?
3. ¿Cuántas de las palabras que fueron sugeridas en una de las tres clases de "named entities" y no fueron elegidas por ud. como tal fueron confusiones producto de errores ortográficos? ¿Qué puede concluir de aquello?
4. ¿Qué sucede con las estructuras como tablas y listas? ¿Facilitan o dificultan la etiquetación? ¿Qué sucede cuando la respuesta es la entidad? ¿En este último caso se obtiene una buena o mala identificación? Ejemplifique. En estos casos, ¿Las sugerencias son o no un aporte?
5. Utilice MontyLingua[1] para hacer un análisis morfológico y sintáctico de las palabras en las tripletas. ¿Qué observa? ¿Las palabras que ud. etiqueto como entidades son suceptibles a ser mapeadas a una raíz o éstas no cambian? ¿Hay una clase más afectada? ¿Qué sucede con los tokens? En cuanto a las

1 web.media.mit.edu/~hugo/**montylingua**

categorías sintáctica ¿Cuál es la distribución en el conjunto de datos? ¿Cuáles son las cinco más prominentes en cada una de las 26 categorías? ¿Cuales son las categorías sintácticas más afectas a ser mapeadas a su raíz? ¿Cuál (es) es la categoría sintáctica más frecuente entre las palabras que etiqueto manualmente como "named entity"?

6. Antes de una "named entity": ¿Cuál es la clase sintáctica más común?, y después de una "named entity" ¿Qué sucede? ¿Es para las tres clases de entidades lo mismo?.
7. Considerando un ejercicio binario 0/1, si una palabra es o no entidad. Calcule la accuracy, la precisión, el recall y el F-Score de la clase positiva (entidad). Para ésto, considere las sugerencias como respuesta automática y sus etiquetas manuales como la "verdad absoluta"? También calcule la entropía del conjunto de datos considerando 0/1.
8. Para modelar el contexto de cada clase, construiremos los vectores de contexto izquierdo y derecho. Para cada clase (incluyendo los tokens). El vector de contexto izquierdo es un histograma de la primera palabra a la izquierda de una clase de entidad. El derecho es su homologo pero al otro lado. Por ejemplo, la siguiente respuesta contenida en nuestro ejemplo de trabajo: "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" tiene siete palabras etiquetadas como persona. El vector de contexto izuiqerdo sería: of=1, liked=1, Edward=1, Hitler = 1, with=1, and=1, -=1, en cambio el de contexto derecho sería: 8th=1, because=1, Hitler=1, angered=1, did=1, and=1, has=1. Haciendo ésto para todas la colección etiquetada ¿Qué observa para los ocho vectores? ¿Qué sucedería si hacemos el procedimiento para el 2do vecino más cercano y para cada lado? Por ejemplo, al lado izquierdo sería rid=1, of=1, he =1, liked=1,

banking=1, again=1, Russia=1. ¿Si consideranos el tercero? Nota que para facilitar el proceso, conviene agregar tres place holders al comienzo y al final de la respuesta. Uno típico es "@#@#".

Finalmente, esboce sus conclusiones finales, para guiar sus conclusiones intente responder la siguiente pregunta ¿Qué indicadores hay en el texto que podría indicarme si una palabra es una "named entity" o no, si su clase respectiva?. Además, piense en: ¿Qué conclusiones cree que también se darían para el caso del idioma español? ¿Qué características piensa ud. harían más fácil/difícil el reconocimiento de entidades en español?. Fundamente y ejemplifique sus respuestas. También, en las conclusiones, explayese en cómo podría la identificación automática de entidades ayudar a las empresas. Por ejemplo, ¿Estaría un banco interesado en reconocer entidades? ¿En qué tipo de casos y para qué? ¿Al gobierno le interesaría identificar entidades en alguna de sus base de datos?

## Además, tenga en cuenta que:

**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.**

**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.**