UNIVERSIDAD NACIONAL
ESCUELA DE INFORMÁTICA
PARADIGMAS DE PROGRAMACIÓN

Proyecto
Parseador de Twitter y Facebook

Eddy Ramírez Jiménez

7 de septiembre de 2015

1. MOTIVACIÓN

La teoría de lenguajes es amplia y apenas ha tenido éxito en el desarrollo de lenguajes de programación, sin embargo ha tenido poco éxito en el parseo[1] exitoso sin errores de lenguaje natural, que será parte de este proyecto.
Por otra parte, la creación de lenguajes de programación está dada para facilitar el desarrollo de una aplicación particular. Para ello es que se realizan lenguajes de programación en donde es fácil trabajar con ciertas características por encima de otras.
En este proyecto se utilizarán 3 lenguajes de programación a saber: Java, Erlang y Prolog (en la versión Swi-Prolog) donde en cada lenguaje se podrán aprovechar las características propias de los tres, así como las características de sus respectivos paradigmas.
El proyecto consiste en el etiquetado de diversos temas de los tweets y posts de Twitter y Facebook de costarricenses. Para ello el proyecto se divide en diversas etapas: La captura de los textos originales, el análisis de esos textos y finalmente el mostrar resultados gráficamente del trabajo realizado.

2. MÓDULOS DEL SISTEMA

El proyecto constará con diversos módulos los cuales se detallan a continuación.

2.1. CONSULTAR DATOS DE LOS SERVIDORES

Utilizando Erlang, se deben crear hilos que van a estar sensando y consultando los sitios https://dev.twitter.com/ y https://developers.facebook.com/ para obtener los comunicados ticos de esas aplicaciones, y deben de eliminar artículos y preposiciones de dichos textos. En caso de ser necesario, sería muy agradable para futuras etapas del proyecto que busque además hacer una corrección ortográfica (en la medida de lo posible) al texto obtenido.

2.2. ALMACENANDO LOS DATOS

Los datos se van a almacenar en mnesia, la base de datos de Erlang, provisionalmente, mientras otros hilos van a enviar por socket de mnesia a Prolog los datos, quien se va a encargar de realizar la siguiente etapa de etiquetado.
Luego, cuando ya estén cargados con los datos de prolog, se enviarán a un programa en Java para su final almacenamiento en una base de datos relacional libre (se sugiere que sea Postgres o MySql).
En la medida de lo posible, se utilizarán los siguientes datos:

1. Texto del mensaje
2. Hora en que fue emitido
3. Lugar en que fue emitido

[1] Palabra no existente en español, debería leerse *análisis textual*

4. Usuario autor del mensaje
5. Hastags utilizados (puede ser más de uno)
6. Medio, si es Facebook o Twitter

2.3. PARSEANDO LAS HILERAS

Utilizando en Swi-Prolog el DCG[2] con expresiones regulares, se parseará el texto de los mensajes previamente guardados y se etiquetará con respecto a su contenido por temas. Primero se etiquetará si el emisor emite un mensaje negativo, positivo o neutro. Luego indicará si se trata de un mensaje relativo a la política, futbol, educación, chiste, estado de ánimo personal o desconocido (si no calza en ninguna otra opción).
Esta clasificación se hará automáticamente con respecto a algún banco de palabras que constará con al menos 100 palabras de cada tema (pueden ser hashtags o nombres propios). Es importante considerar que las imágenes no deben ser consideradas para ser analizadas.

2.4. MOSTRANDO RESULTADOS

En Java se debe de generar una serie de gráficos, se sugiere utilizar JFreeChart de modo que se muestre la relación entre:

- Cada tema y la hora del día en que se realizaron. Es decir, indicar por fecha si hubo alguna hora en que hubo *"picos"* en que se habló más de un tema
- Un gráfico de pastel con las proporciones de cada tema
- Un gráfico de barras con la frecuencia con que se han emitido los mensajes por hora (independientemente del tema)
- Una relación entre los mensajes con hashtags y los que no tiene.
- El gráfico de los hashtags más utilizados en ciertas horas
- Los usuarios con más mensajes (posts o tweets)
- Relación de densidad de mensajes por hora por cada medio (Facebook o Twitter)
- Tres otros gráficos que cada grupo considere que son interesantes para mostrar.

Todos estos gráficos deben ser realizados por medio de consultas a una base de datos relacional.

[2] http://www.pathwayslms.com/swipltuts/dcg/

3. Flujo del programa

El programa debe de comenzar a funcionar desde Erlang, quien se va a encargar de emitir hilos que sensen el estado de Facebook y Twitter para agregarlos a su base de datos no relacional, con sus respectivas etiquetas. También habrán otros hilos que estarán enviando a Prolog mensajes para que prolog los analice y retorne las etiquetas pertinentes, las cuales serán añadidas a la base de datos no relacional en tuplas.
Una vez que un elemento (tupla) esté completo, entonces un tercer grupo de hilos, enviará esta información a un programa en Java que se encargará de ponerlo en una base de datos.
Finalmente cuando un usuario lo desee ejecutará el módulo de gráficos para poder ver las diferentes gráficas solicitadas en ese módulo.

3.0.1. Inicio del programa

El programa se puede iniciar por aparte cada módulo o se puede utilizar un programa en C que inicie todos los demás.

3.1. Comunicación entre los módulos

Se espera que la comunicación entre cada módulo sea a través de sockets en todo momento.

4. Documentación

El documento debe ser escrito en LaTeX y debe ser entregado en digital y debe de poseer muestras del funcionamiento de cada una de las partes del sistema. Así como resultados preliminares de las corridas.
En caso de haber uno o más módulos incompletos, se debe de explicar la solución planteada para solventar los posibles fallos que no se resolvieron.

5. Aspectos Técnicos

1. El proyecto debe de funcionar apropiadamente en Linux y ser elaborado en los lenguajes indicados en esta especificación.
2. Se permite el uso de todas las bibliotecas que cada lenguaje posea para facilitar el desarrollo del proyecto.
3. Cualquier duda con respecto a uso de bibliotecas externas permitidas deben de ser consultados al profesor

6. Aspectos Administrativos

- El trabajo puede ser realizado en parejas.

- La fecha de entrega es en semana 17 de forma inamovible.
- La evaluación queda sujeta al criterio del profesor.