Universidad de los Andes Facultad de Ingeniería	Ingeniería de Sistemas y Computación **Maestría** MINE 4102 – Análisis de Información sobre Big Data Semestre: 2017-2	Accredited ABET Engineering Accreditation Commission

Taller 1

Objetivo

- Recolectar información sobre la Web a partir de estrategias de *crawling*
- Conocer y gestionar de forma básica una fuente semiestructurada de datos: RSS
- Utilizar expresiones regulares y su uso en ambientes de gestión de información semiestructurada
- Utilizar XQuery sobre documentos semiestructurados XML
- Utilizar el entorno Unix en la construcción de soluciones Web para el descubrimiento y recolección de información.

Prerrequisitos

- Herramientas y lenguajes para desarrollo de aplicaciones Web. Por ejemplo, Java, JSP, Phyton, etc.
- Conocimiento básico de Unix y ambientes de virtualización
- Conocimiento básico de expresiones regulares (regEx)

Metodología

- Se trabaja de acuerdo con los lineamientos generales del curso.
- Se realiza una entrega por grupo
- Utilice para el documento las pautas de elaboración de documentos técnicos que encuentra en Sicua+.

Enunciado

1. **Descubrimiento de información utilizando *crawling*;**
   Construya de forma dinámica el directorio de personas vinculadas con Uniandes a partir de la información disponible en la Web de cada una de las dependencias académicas. Incluya directivos, profesores, personal de laboratorio, personal administrativo y posiblemente estudiantes allí referenciados.

   - Descubra dinámicamente el conjunto de unidades académicas y de apoyo administrativo
   - Descubra dinámicamente las personas que dirigen y que trabajan en cada una de dichas unidades
   - Descubra las 3 noticias más recientes que publica cada una de las unidades, con su detalle.
   - Construya y despliegue una tabla con el nombre la información completa de las personas (nombre, correo, cargo, oficina, extensión telefónica, página Web, página en Google Scholar, redes sociales, unidad académica, etc).
   - Permita que la tabla sea filtrada por cualquiera de los atributos posibles para las personas, según la voluntad del usuario. En particular, se quiere poder desplegar categorías de profesores o empleados, los que son tocayos, los que tienen cuenta en alguna red social, el directorio de Google Scholar, etc)
   - Para los profesores, debe indicar su currículo, títulos académicos, cursos que ofrece, si es de planta o de cátedra.
   - Para los coordinadores de programa debe incluir el nombre de programa, información de SNIES y descripción general del programa.
   - Tenga cuidado en el proceso de forma que en NINGÚN caso debe afectarse la calidad de servicio ni la disponibilidad de cada una de las fuentes.
   - Sólo debe considerarse información encontrada en sitios Web oficiales de Uniandes. Tenga especial cuidado en que algunas personas publican en sus sitios Web enlaces a sitios fuera de Uniandes. NO debe viajar a sitios fuera del dominio de la universidad.
   - NO ESTÁ ADMITIDO HACER USO DE SERVICIOS DE INFORMACIÓN ELECTRÓNICA DE UNIANDES. En particular, NO es válido buscar la información en sistemas como Banner, Academia, LDAP, etc.
   - OPCIONAL: si una persona es responsable de eventos o noticias, incluya el enlace a la información del evento, la fecha en que ocurre y el lugar. Incluya toda la información sobre el responsable.

2. **Utilización de fuentes de sindicación/suscripción, RegEX y XQuery**:
   Desarrolle un componente sencillo que permita consumir las fuentes indicadas al final del enunciado y filtrarlas por cualquiera de los siguientes elementos: palabras en el título, en la descripción o en las categorías del artículo (post), de acuerdo con lo indicado por el usuario. El criterio de consulta es un elemento y su contenido. El usuario debe poder actualizar el criterio de filtrado en cualquier momento.
   Por ejemplo, mostrar las noticias etiquetadas en Juegos Olímpicos, con Juegos Olímpicos en el título o Juegos Olímpicos en la descripción.

   - Tome el conjunto de fuentes RSS que se indica en la tabla asignada a su grupo. Suscriba dicho RSS en su aplicación. Seleccione al menos 3 categorías de información compatibles entre las fuentes indicadas (por ejemplo, noticias de deportes, tecnología y actualidad mundial). Puede incluir fuentes similares, si encuentra alguna dificultad con las propuestas para su grupo. Infórmelo rápidamente a su profesor.
   - Revise el código fuente de los RSS de las fuentes indicadas y reconozca un conjunto de etiquetas XML propias al RSS, frecuentes en la fuente. Reconozca la estructura de la información y las diferencias entre las fuentes
   - Investigue cuáles son los elementos estándar de RSS y ubíquelos en sus datos.

   Debe ser posible ver la información suscrita, en forma simultánea, de tres maneras:
   - Los elementos sin filtrar. Se muestran sólo los títulos de las noticias
   - Los elementos filtrados utilizando expresiones regulares como mecanismo de filtrado. Se muestran los títulos de la noticia
   - Los elementos filtrados utilizando XQuery como mecanismo de filtrado. Se muestran los títulos de la noticia, la fecha de publicación y el link a la misma.

Entregable

- Muestre los resultados solicitados en una aplicación Web sencilla, que ofrezca las ***funcionalidades*** solicitadas. No olvide relacionar el número de grupo y sus integrantes en la página Web de resultados.
- Elabore un documento de **máximo 4 páginas** en el cual relacione;
  - ✓ Forma de acceso a la aplicación Web que muestra los resultados.
  - ✓ **Métodos y tecnología** concretos utilizados en cada uno de los retos propuestos
  - ✓ **Expresiones en XQuery y las RegEx** que le utiliza en la solución
  - ✓ El **algoritmo básico** para resolver cada uno de los retos, de manera que puedan percibirse los elementos interesantes para poner en valor en la solución
  - ✓ **Análisis de resultados obtenidos**, dificultades, logros y posibilidades de generalización de la solución. Analice la calidad de los resultados obtenidos desde el punto de vista de la información entregada al usuario. Analice problemas encontrados, mejoras posibles, retos por resolver para hacer un mejor trabajo de entrega (delivery) de información al usuario. Proponga posibles extensiones de valor agregado.
  - ✓ Analice **en qué cambia su solución** si se quiere construir el directorio completo de la Universidad, incluyendo las publicaciones de los profesores, de la librería o unidades de investigación y responsables de los órganos de gobierno y de administración de la Universidad**.**

- **Su solución DEBE permanecer en línea hasta el final del semestre.**
- **Su solución en línea y la mostrada en el momento de sustentación DEBE CORRESPONDER DE MANERA EXACTA CON LO ENTREGADO EN SICUA+, si hay cualquier diferencia, por mínima que sea, puede ser considerado FRAUDE**
- **Los resultados de su solución deben ser comprobables en el momento de ejecución con los datos que se encuentran en las fuentes correspondientes.**

Aspectos que el grupo decide

1. El filtraje se realiza para evitar el despliegue de los artículos etiquetados según la selección del usuario, o para desplegar únicamente los correspondientes a las etiquetas seleccionadas (consultas tipo NOT IN)
2. ¿El usuario decide qué tipo de filtraje hacer? ¿O se hace el escogido por el grupo en el punto anterior?

Requerimientos técnicos

1. Realice el proceso de filtrado utilizando el mínimo número posible de expresiones regulares.
2. La interacción con el usuario debe ser en una aplicación Web gráfica, sencilla pero intuitiva y bien presentada.

3. Desarrolle y despliegue la aplicación solicitada en el ambiente UNIX provisto en el curso.

Recomendaciones

Diseñe este componente de forma que pueda ser extensible, configurable e integrable con otros entornos Web. Estos entornos pueden ser talleres o tareas posteriores u otros sitios web que encuentren este servicio interesante y la solución apropiada.

Asignación de fuentes de datos

Grupo	Fuentes
1, 6	El Tiempo, elmundo.com, WRadio
2, 7	El Colombiano, CNN en Español, BBC Mundo
3, 8	GoogleNews, LinkedIn, FlipBoard
4, 9	Mashable, Reuters News, DW
5,10	Wired, Lifehacker, BBC Technology

Evaluación

La evaluación se hace siguiendo los lineamientos establecidos para los trabajos prácticos y talleres del curso. La entrega se hace así:

Entregable	Fecha y hora límite de entrega	Porcentaje en la evaluación
Proyecto de software desarrollado, aplicación funcional y demostración	Lunes 11 de septiembre, 14:00	70%
Informe	Lunes 11 de septiembre, 14:00	30%

El cumplimiento de las restricciones técnicas es parte integral de los dos entregables. No satisfacerlos invalida LOS DOS entregables.

Se espera que cada miembro del grupo haga una contribución igualmente significante al desarrollo de esta actividad y a las tareas definidas al interior del grupo. El trabajo por debajo de este rango tiene una penalización proporcional sobre la evaluación global de la tarea

Los resultados serán sustentados en sesión de 20 minutos por grupo en horario definido en Sicua+, el día de la entrega.

Entregables

.

Archivo de la entrega: <Taller1_NN_login1_login2_login3>.zip.

Donde NN es el número del grupo y login1 y login2 son los correspondientes a los miembros del grupo en Uniandes.

Contenido: Archivo zip con el proyecto de software y archivo .pdf con el informe de análisis. Nombre del archivo de análisis: Taller1_NN_login1_login2_login3.pdf

El no seguimiento del formato de entrega del taller tiene una penalización de **0.5/5.0** en la nota final. La no presentación a la sustentación de los resultados produce una nota final en la tarea de 0.0/5.0. El grupo COMPLETO tiene UNA oportunidad de sustentación.

La hora de entrega DEBE ser la indicada en el enunciado, independientemente de la disponibilidad eventual posterior del enlace de entrega en Sicua+. El taller se rige por las normas definidas en las reglas de juego de trabajos prácticos. En particular, entregas tardías tienen como evaluación 0/5.