Proyecto – Aplicación de extracción de información

Dr. José Lázaro Martínez Rodríguez

De acuerdo con las actividades referentes a la lectura de archivos PDF y la ejecución de expresiones regulares con C# (o Python), desarrolle una aplicación para la extracción de información. Dicha aplicación consiste en obtener, por medio de **expresiones regulares**, los siguientes elementos a partir de archivos PDF de Curriculum vitae (ver archivos adjuntos):

- Nombre
- Fecha de nacimiento
- Lugar de nacimiento
- Dirección
- Correo electrónico
- Teléfono
- Formación académica
- Último registro de experiencia profesional
- Título (Ing. Lic, etc)

El proceso consiste en leer un archivo (ya sea PDF o Word), aplicar las expresiones regulares, obtener coincidencias y mostrar la información solicitada en pantalla. Se recomienda realizar una interfaz gráfica con la que se desplieguen los datos en un formulario.

Recomendaciones:

- Debe crear una expresión regular por cada elemento de información requerido
- Investigue las bibliotecas requeridas para correr las regex en c# o python.
- Debe preparar su escenario para presentarlo durante la sesión en grupo el día de la entrega.
- Se aceptan equipos de dos personas. Se evaluará su contribución equitativa durante la presentación.
- De no seguir las rúbricas de la presentación y el reporte, se descontarán puntos de su calificación para esta actividad.
- OPCIONAL: Obtener el texto de una imagen OCR.

Restricciones:

- La misma expresión para cada elemento debe ser funcional para los distintos formatos de CV
- El sistema debe detectar automáticamente el tipo de archivo a leer (pdf o Word). No se le debe indicar manualmente.
- Las expresiones no deben ser a medida para cada formato de curriculum. Es decir, **NO** se deben tener regex para el curriculum “A” y otras regex diferentes para el curriculum “B” sino que la misma expresión debe funcionar independientemente de la estructura del curriculum.

Reporte

Prepare el reporte de la actividad y expórtelo en formato PDF. El reporte debe incluir

- Portada. Institución, tema tratado, integrantes.
- Objetivo de la actividad
- Desarrollo. Indique la manera en que implementó cada etapa en el proceso de la aplicación.
    - La lectura de archivos,
    - la implementación de las expresiones regulares requeridas,
    - la forma de buscar coincidencias de dichas expresiones con respecto al texto leído
- Muestre el código, de preferencia por bloques principales para ir explicando el propósito. Ver recomendaciones al final del documento para formatear código.
- Pruebas. Realizar pruebas con los documentos de ejemplo y además **incluir un cv con el que usted cuente para obtener los elementos antes mencionados**.
- Limitaciones. ¿Qué no hace el sistema o en qué condiciones puede fallar? Esto no es penalizable para el proyecto pero siempre es buena practica incluir las limitaciones.
- Conclusiones. No incluir opiniones (subjetivo) sino los hallazgos encontrados durante el desarrollo de la actividad (objetivo).

- Código. Debe estar indentado, usando formato:
    - Letra consolas/Menlo, una de las dos.
    - Tamaño de letra 9.
    - Espaciado simple.
    - Colocar código dentro de cuadro de texto (use título descriptivo).
    - Puede usar varios cuadros de texto para un programa, según necesite.
    - Puede usar coloreado e indentación proporcionado por su entorno de desarrollo (IDE). Por ejemplo, usando *Visual Studio Code* (no es obligatorio usar este, puede usar otro) se muestra el siguiente código:

```
struct Stack
{
  int top;
  unsigned capacity;
  int *array;
};
```

Código 1 Estructura para una pila

Fecha de entrega 11 de mayo. De no entregar el proyecto (o cometer plagio) no aprobará la materia.
https://cmicdelegacionchihuahua.org.mx/wp-content/uploads/arforms/userfiles/?SD