Desafío

IA x la Identidad

Descripción del problema

Te vamos a dar acceso a un repositorio GIT donde se encuentran las herramientas que vamos a utilizar y que consisten principalmente en un programa que transcribe (parcialmente) imágenes a texto. Te vamos a pedir que clones el repositorio para armarte uno propio en el que vas a realizar tu desarrollo. Para inscribirte, vamos a pedirte que nos indiques tu usuario en GitLab. Si no tenés, por favor create uno.
El proyecto actual está diseñado para el procesamiento de escaneos de documentos mecanografiados o impresos. A partir de un archivo de imagen, aplica un OCR y realiza algunas funciones de post-procesamiento para mejorar el resultado, incluyendo un corrector ortográfico. El OCR utilizado puede encontrarse acá: Tesseract-OCR y un poco de documentación acá: TessDoc. El post-procesamiento consiste en un filtro de caracteres no válidos, unión de palabras por saltos de página, y un corrector ortográfico, JamSpell. Vas a tener que instalar estas herramientas para poder correr el programa. Para eso (para que funcione JamSpell), vas a necesitar una máquina con Linux. En caso de usar otro sistema operativo, podés bajarte de acá una máquina virtual con todo lo necesario instalado[1]. La contraseña del usuario es *identidad1234*.

El repositorio tiene programas escritos en Python y utiliza Tesseract y JamSpell.
Actualmente, el programa principal captura el texto en un único bloque, y el desafío consiste en modificarlo para que detecte los diferentes fragmentos de texto que componen una nota periodística, como por ejemplo título, copete, epígrafe, etc.
Específicamente, tenemos un programa que toma dos parámetros de entrada de tipo string, indicando directorios de entrada y salida. Por default, se procesan las imágenes ubicadas la carpeta de origen input_data/ y se guardan los resultados en la carpeta de destino out_data/, Para ello, ejecutar desde la terminal:

```
python3 run.py
```

o desde el intérprete de python:

```
procesar_imgs("input_data", "out_data")
```

En caso de querer usar otros directorios, usar los flags -e y -s respectivamente. Por ejemplo, si quiero leer imágenes desde el directorio "entrada/" y que la salida sea en "salida/",

[1] Importante: en la importación de la máquina virtual seleccionar 2 o más CPUs para su uso. Si nunca usaste máquinas virtuales, sencillamente instalá VirtualBox y luego importá como en este tutorial.

ejecutar python3 run.py -e entrada/ -s salida/ o bien, llamar a la función procesar_imgs(“entrada”, “salida”). Para más información, leer el readme.

El programa genera un reporte que guarda en el directorio de salida con formato .csv y, por cada imagen del directorio de entrada, un archivo .json conteniendo un diccionario. Este diccionario tiene como claves los tipos de fragmentos que pueden aparecer en la nota. Para cada clave, el valor asociado es una lista (posiblemente vacía) de cadenas de caracteres, con el contenido de la transcripción de los bloques de ese tipo.
Actualmente, como el programa no puede detectar los distintos fragmentos de texto, una sola de las claves del diccionario generado tiene como valor una lista no vacía: la lista con una única cadena de caracteres que contiene todo el texto detectado en la nota, junto. El resto de las claves del diccionario generado tienen como valor la lista vacía.

Ejemplo

Consideremos la siguiente imagen. Con color naranja indicamos los fragmentos que queremos reconocer.

Ratifican las Abuelas una identificación cuestionada

El cadáver de Liliana Pereyra fue identificado en una exhumación realizada en Mar del Plata. Quedó confirmada la aseveración del antropólogo norteamericano Clyde Collin Snow.

Las Abuelas de Plaza de Mayo informaron ayer que el cuerpo exhumado el 9 de marzo último en el cementerio de Mar del Plata pertenece a Liliana Pereyra —desaparecida en octubre de 1977 junto con su esposo Eduardo Cagnola— y que los peritajes demuestran que fue asesinada con un disparo de Itaka y que dio vida a un bebé.

En conferencia de prensa realizada ayer el organismo informó que el cuerpo registrado como "NN" en el cementario Parque de Mar del Plata fue exhumado a instancias de la madre de la joven, Jorgelina Azzari de Pereyra y analizado por el antropólogo forense Clyde Collin Snow quien confirmó la identidad del cadáver.

Como se recordará Snow prestó declaración ante la Cámara Federal que sustancia el juicio contra los comandantes, y demostró —a través de diapositivas— el procedimiento de inhumación y análisis de los restos de la joven indicando rastros de un parto reciente y la destrucción del cráneo por un disparo de Itaka a poca distancia.

Jorgelina Azzari de Pereyra relató que conocía el nacimiento de su nieto por testimonios de ex detenidos desaparecidos que compartieron el cautiverio con Liliana, mencionando a Sara Olars de Osatinsky, destinada por los represores a atender a las embarazadas que estaban recluidas en la Escuela de Mecánica de la Armada.

Allí, "se le hizo adelantar el parto para desocupar el local" y pocos días después "sin su bebé, Liliana fue llevada nuevamente a la Base de Buzos Tácticos de Mar del Plata" mientras el niño fue retirado por un subprefecto Héctor Fabres o Febres.

Denunciando demoras en la tramitación judicial de la inhumación en la que consten las verdaderas razones de su fallecimiento —ya que figuraba como "muerta en un enfrentamiento"— letrados de Abuelas de Plaza de Mayo señalaron que "todos asistimos en estos días al proceso que se sigue a las juntas que es una parte importante, pero sólo una parte de la Justicia argentina".

"La otra, la de todos los días, es la que permanentemente pone trabas para el esclarecimiento de este tipo de casos, ya que hay jueces que ni siquiera nos reciben en su despacho".

La madre de Silvia Pereyra

El objetivo es que el programa genere, para esta imagen, un diccionario similar a este:
{
"Diario": [],
"Fecha": [],
"Volanta": [],
"Título": ["Ratifican las Abuelas una identificación cuestionada"],
"Copete": ["El cadáver de Liliana Pereyra fue identificado en una exhumación realizada en Mar del Plata. Quedó confirmada la aseveración del antropólogo norteamericano Clyde Collin Snow."],
"Destacado": [],
"Cuerpo": ["Las Abuelas de Plaza de Mayo informaron ayer que el cuerpo exhumado el 9 de marzo último en el cementerio de Mar del Plata pertenece a Liliana Pereyra -desaparecida en octubre de 1977 junto con su esposo Eduardo Cagnola- y que los peritajes demuestran que fue asesinada con un disparo de Itaka y que dio vida a un bebé. En conferencia de prensa realizada ayer el organismo informó que el cuerpo registrado como

‘NN’ en el cementerio Parque de Mar del Plata fue exhumado a instancias de la madre de la joven, Jorgelina Azzari de Pereyra y analizado por el antropólogo forense Clyde Collin Snow quien confirmó la identidad del cadáver. Como se recordará Snow prestó declaración ante la Cámara Federal que sustancia el juicio contra los comandantes, y demostró -a través de diapositivas- el procedimiento de inhumación y análisis de los restos de la joven indicando rastros de un parto reciente y la destrucción del cráneo por un disparo de itaka a poca distancia. Jorgelina Azzari de Pereyra relató que conocía el nacimiento de su nieto por testimonios de ex detenidos desaparecidos que compartieron el cautiverio con Liliana, mencionando a Sara Olars de Osatinsky, destinada por los represores a atender a las embarazadas que estaban recluidas en la Escuela de Mecánica de la Armada. Allí, ‘se le hizo adelantar el parto para desocupar el local’ y pocos días después ‘sin su bebé’, Liliana fue llevada nuevamente a la Base de Buzos Tácticos de Mar del Plata mientras el niño fue retirado por un subprefecto Héctor Fabres o Febres. Denunciando demoras en la tramitación judicial de la inhumación en la que consten las verdaderas razones de su fallecimiento -ya que figuraba como ‘muerta en un enfrentamiento’- letrados de Abuelas de Plaza de Mayo señalaron que ‘todos asistimos en estos días al proceso que se sigue a las juntas que es una parte importante, pero sólo una parte de la Justicia argentina’. ‘La otra, la de todos los días, es la que permanentemente pone trabas para el esclarecimiento de este tipo de casos, ya que hay jueces que ni siquiera nos reciben en su despacho."],
        "Epígrafe": ["La madre de Silvia Pereyra"],
        "Firma": []
}

Actualmente el programa genera un diccionario con las mismas claves pero donde todo el texto está en la lista correspondiente a “Cuerpo”, y las demás entradas del diccionario tienen listas vacías. Además, según la nota, las columnas no son identificadas por separado, por lo que la transcripción puede tener muchos errores.

Para posibilitar la evaluación, vamos a pedirte que preserves el funcionamiento del programa. Es decir, que se pueda correr de la misma manera y que genere una salida similar, es decir, un reporte y, para cada imagen, un diccionario que contenga las mismas claves. La diferencia debe consistir solamente en los valores del diccionario.

También vamos a evaluar la documentación e interpretabilidad del código, por lo cual es muy importante que el código esté bien organizado, con comentarios donde sean necesarios, indicando qué se computa en cada función o parte del código. En el repositorio original vas a poder ver el nivel de detalle que esperamos.
Además, te pedimos que no modifiques el archivo README. Si hay algo que modificarías, por favor indicalo y explicalo en el informe.

Una de las tareas que vamos a realizar dentro de los primeros días de la competencia es etiquetar manualmente -o semi manualmente- algunos datos que serán de utilidad para el desarrollo de los programas. Para facilitar esta tarea, la vamos a hacer de manera colaborativa entre todxs lxs participantes. Es decir que los datos etiquetados los vamos a

guardar en un espacio accesible a todxs. Aportar a esta base de datos etiquetados también suma puntos.

Al finalizar el tiempo de trabajo de la competencia, vamos a pedirte que entregues un informe. El mismo debe contener una explicación de los desarrollos aportados, funcionamiento del programa, mejoras que se podrían hacer, si hubiera más tiempo, y otros comentarios relevantes que consideres. En el mismo informe debe figurar el enlace al repositorio GIT correspondiente, que debe estar accesible (en modo reporter) para los usuarios de los jurados.
El informe completo no debe superar las dos páginas. La idea es que no dediques un montón de tiempo al informe, sino que sirva para que el jurado pueda entender tu trabajo.

Entrega

Cada participante (o equipo participante) deberá enviar el informe a través de un formulario, que estará disponible en el sitio de la competencia desde el día 14 de abril hasta el 28 de abril a las 23:59. En el informe debe figurar el enlace al repositorio donde está el desarrollo, y que sea accesible como ‘reporter’ para los jurados. Puede realizarse más de un envío, teniendo en cuenta que se evaluará solamente el último envío realizado por cada participante.

Evaluación

La evaluación de las soluciones presentadas estará a cargo de un jurado experto en el área y se tendrán en cuenta los siguientes aspectos.

a) exactitud de la tipificación de bloques en el conjunto de datos de evaluación
b) interpretabilidad y documentación del código presentado
c) presentación y contenido del informe presentado
d) aportes a la base de datos etiquetados manualmente
e) otros recursos aportados por la solución propuesta