UNIVERSIDADE DA CORUÑA

facultade de informática da coruña

Anteproxecto TFG
GRAO EN CIENCIA E ENXEÑARÍA DE DATOS (GCED)

Datos da/o estudante*:

Nome e apelidos	**Ana Cives Trillo**
DNI	**206304267W**
Enderezo electrónico	**ana.ctrillo@udc.es**
Teléfono	**697938388**

Título (galego)*: Desenvolvemento dun sistema de procesamento automático de facturas comerciais con OCR e NLP

Título (castelán)*: Desarrollo de un sistema de procesamiento automático de facturas comerciales con OCR y NLP

Título (inglés)*: Development of an Automated Invoice Processing System with OCR and NLP

Tipo de proxecto*:

Clásico

Dirección:

Máximo dúas persoas da FIC. Permítese unha terceira persoa só no caso de pertencer a outra organización (que hai que especificar):

- Titor(a) da FIC*: Pablo Alejandro Calviño Padín

- Titor(a) da FIC:

- Director(a) externo/a: Adrián Villegas Duque
  Organización: *Vento Abogados&Asesores*

Breve descrición*:

En determinados ámbitos, traballar con facturas de maneira manual pode volverse unha tarefa complexa e propensa a erros, especialmente cando o volume destas é considerable. Por iso, co obxectivo de optimizar o tratamento destes documentos, reducir a carga de traballo e mellorar a eficiencia na xestión administrativa, o presente Traballo Fin de Grao propón o

desenvolvemento dunha aplicación intelixente orientada á automatización do procesamento de facturas comerciais en formato PDF.
A aplicación basearase en dúas tecnoloxías fundamentais:
OCR (Recoñecemento Óptico de Caracteres): Permite converter texto contido en imaxes ou documentos PDF en texto editable e analizable. Esta fase inicial ten como obxectivo estructurar o contido do documento para que poida ser tratado por algoritmos informáticos.

NLP (Procesamento de Linguaxe Natural): Unha vez dispoñemos do texto, este módulo analiza linguisticamente o contido e identifica información clave, como datas, importes, nomes de provedores, números de factura ou NIFs.

A información extraída será almacenada nunha base de datos estruturada, facilitando así a súa consulta posterior, visualización e integración con outros sistemas. Para o acceso á información, prevese a creación dunha interface gráfica amigable ou dunha API que facilite o uso do sistema.

Obxectivos concretos*:

1. Desenvolver un sistema funcional de extracción de datos desde facturas utilizando OCR, concretamente, empregando Tesseract OCR para procesar as imaxes ou PDFs.
2. Aplicar técnicas de Procesamento de Linguaxe Natural (NLP) para a análise e clasificación automática dos campos relevantes, integrando ferramentas como spaCy, NLTK ou Transformers, co fin de identificar de maneira eficiente as entidades clave presentes nas facturas como datas, importes ou códigos fiscais.
3. Almacenar os datos extraídos nunha base de datos estruturada, mediante o deseño dun esquema eficiente e adaptado ás características dos datos identificados durante o proceso de extracción.
4. Implementar unha interface de usuario ou unha API para facilitar o uso do sistema, mediante a creación dunha interface gráfica sinxela que permita aos usuarios cargar documentos, visualizar os datos extraídos e realizar correccións cando sexa necesario.

Método de traballo*:

Para garantir a execución eficiente e estruturada do Traballo Fin de Grao, o desenvolvemento da aplicación seguirá unha metodoloxía áxil, centrada na entrega progresiva de funcionalidades completas en ciclos curtos (iteracións). Esta aproximación permitirá validar os resultados de forma continua, tanto a nivel técnico como funcional, garantindo que cada módulo implementado cumpra cos requisitos definidos.

Ademais, esta metodoloxía facilitará a detección temperá de posibles problemas ou desviacións, permitindo introducir modificacións no deseño ou na planificación sen comprometer o avance global do proxecto. A interacción frecuente cos usuarios ou supervisores, característica clave das metodoloxías áxiles, tamén contribuirá a aliñar o desenvolvemento coas necesidades reais, favorecendo así a creación dunha solución útil, robusta e adaptada ao contexto específico de uso.

Fases principais do traballo*:

O desenvolvemento deste Traballo Fin de Grao estrutúrase en cinco fases principais, cada unha delas centrada nun aspecto fundamental do sistema:

1. Obtención do texto a partir das facturas (OCR): Nesta fase lévase a cabo a extracción do contido textual presente nas facturas. Utilizarase a ferramenta Tesseract OCR, que permite converter a información contida en imaxes ou PDFs en texto editable.

2. Procesamento da linguaxe natural (NLP): Unha vez obtido o texto mediante OCR, aplicarase Procesamento de Linguaxe Natural (NLP) co obxectivo de analizar e comprender o contido das facturas. Esta fase inclúe tarefas como a tokenización, o etiquetado gramatical (POS tagging) e o recoñecemento de entidades nomeadas, que permitirán identificar automaticamente campos clave como datas, importes, nomes de provedores, números de factura ou NIFs. Para a extracción de información, utilizaranse ferramentas e librarías especializadas como spaCy, NLTK ou modelos baseados en Transformers.

3. Almacenamento e explotación dos datos: A información estruturada extraída será almacenada nunha base de datos deseñada específicamente para este fin.

4. Desenvolvemento dunha interface ou API para a interacción co sistema: implementarase

unha capa de presentación que permitirá aos usuarios interactuar co sistema. Esta poderá consistir nunha interface gráfica sinxela, onde se poidan cargar documentos, visualizar os resultados da extracción e realizar correccións manuais, ou nunha API REST que facilite a integración con outras aplicacións ou fluxos de traballo xa existentes.

5. Documentación: esta fase consistirá na redacción detallada da memoria do TFG, na que se explicará de maneira clara e estruturada todo o proceso realizado ao longo do proxecto. A memoria incluirá a introdución do traballo, os obxectivos perseguidos, a metodoloxía empregada, a descrición das fases de desenvolvemento (incluíndo a análise, o deseño, a implementación), así como os resultados obtidos e apartado no que se expoña a análise de posibles melloras ou futuras liñas de traballo.

Material, medios e recursos necesarios*:

1. Hardware: Ordenador persoal: Un equipo con capacidade suficiente para realizar as tarefas descritas.

2. Software e librerías principais: Tesseract OCR, librería de código aberto que será empregada para a extracción de texto a partir das facturas escaneadas en formato imaxe ou PDF; spaCy ou NLTK, librerías para o procesamento de linguaxe natural (NLP) que serán empregadas para analizar o texto extraído e identificar as entidades clave dentro das facturas (datas, importes, NIF, razón social, etc.).
3. Base de Datos: será necesario empregar unha base de datos para o almacenamento dos datos obtidos durante o procesamento das facturas.

4. Conxunto de datos: O conxunto de datos a utilizar corresponde a un conxunto de facturas escaneadas en formato PDF de diferentes tipos, que inclúen facturas de diversos provedores, diferentes formas de organizar a información.

A ampliar durante o desenvolvemento do proxecto.

UNIVERSIDADE DA CORUÑA

facultade de informática da coruña

Propiedade intelectual do traballo*:

O Regulamento dos Traballos de Fin de Grao da Facultade de Informática da Coruña (aprobado pola Xunta de Centro o 18 de xullo de 2022) establece na sección 4, en relación aos dereitos derivados da propiedade intelectual dos traballos, o seguinte:

4.1. No caso dos traballos desenvolvidos en colaboración cunha entidade externa, a titularidade dos dereitos de propiedade e explotación dos resultados, se for o caso, rexerase polo establecido na relación contractual entre a/o estudante e a entidade externa. Neste caso, quen exerza a dirección académica non será titular dos dereitos de propiedade intelectual, salvo que se establecer doutra maneira nun documento asinado pola/o estudante, o profesorado encargado da dirección e un/ha representante da entidade externa.

4.2. No caso dos traballos desenvolvidos no ámbito do centro, a titularidade dos dereitos de propiedade intelectual, se for o caso, corresponderá á/ao estudante segundo queda recollido no apartado h) do artigo 8 do Real Decreto 1791/2010 do 30 de decembro, salvo que se establecer doutra maneira no anteproxecto asinado pola/o estudante e o profesorado encargado da dirección do TFG.

Os resultados xerados durante a realización do TFG poderán ser utilizados para a docencia na UDC, sempre con autorización e mención explícita de quen ostente a súa autoría.

Indique a continuación se o traballo se realiza en colaboración cunha entidade externa ou no ámbito do centro, e neste último caso, o acordo sobre os dereitos derivados da propiedade intelectual do traballo.

O traballo realízase en colaboración cunha entidade externa: **Si**

Nome da empresa: Vento Abogados&Asesores

Se o traballo non se realiza en colaboración cunha entidade externa, indique se os dereitos derivados da propiedade intelectual son compartidos entre a/o estudante e as/os directores:

En A Coruña, a 19 de abril de 2025

Asinado:

Estudante

CALVIÑO PADIN PABLO ALEJANDRO - DNI 46908191J

Digitally signed by CALVIÑO PADIN PABLO ALEJANDRO - DNI 46908191J
Date: 2025.05.19 12:30:05 +02'00'

Titoras/es e Directoras/es

VILLEGAS DUQUE ADRIAN - 39487149J

Firmado digitalmente por VILLEGAS DUQUE ADRIAN - 39487149J
Fecha: 2025.05.18 18:30:11 +02'00'