UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

Aplicació de 'Large Language Models' per a l'Anàlisi i Consulta de Documents Multimodals

Entrega 1:
Contextualització i Abast del projecte

Autor: Albert Reina i Buxó
Director: Marc Alier Forment
Tutor GEP: Fernando Barrabes Naval
Data: 27 de febrer de 2025

Resum

Aquest projecte és lliura com a Treball de Final de Grau en el Grau d'Enginyeria Informàtica de la Facultat d'Informàtica de Barcelona.

L'objectiu principal del projecte és fer ús de "Large Language Models" per a extreure la informació d'imatges, taules i textos en documents amb la finalitat de calcular "embeddings" multimodals per a fer consultes a bases de dades semàntiques.

Resumen

Este proyecto se entrega como Trabajo de Fin de Grado en el Grado de Ingeniería Informática de la Facultad de Informática de Barcelona.

El objetivo principal del proyecto es usar "Large Language Models" para extraer la información de imágenes, tablas y textos en documentos, con la finalidad de calcular "embeddings" multimodales para hacer consultas en bases de datos semánticas.

Abstract

This project is submitted as a Bachelor's Thesis in the Computer Engineering program at the Faculty of Computer Science of Barcelona.

The main objective of the project is to use Large Language Models to extract information from images, tables, and texts in documents in order to compute multimodal embeddings for querying semantic databases.

Índex de continguts

1. Introducció i contextualització..... 4
1.1 Context..... 4
1.2 Identificació del problema..... 4
1.3 Actors implicats..... 5
1.4 Definicions..... 5
2. Justificació..... 7
2.1 Solucions prèvies..... 7
2.2 Justificació..... 8
3. Abast..... 9
3.1 Objectius..... 9
3.2 Impediments i riscos..... 9
4. Metodologia..... 10
4.1 Gestió del projecte..... 10
4.2 Control de versions..... 11
4.3 Validació i proves..... 11
5. Referències..... 12

1. Introducció i contextualització

El projecte *Aplicació de 'Large Language Models' per a l'Anàlisi i Consulta de Documents Multimodals* és un treball de Fi de Grau de modalitat A (centre) i pertany als estudis de Grau en Enginyeria Informàtica de la Facultat d'Informàtica de Barcelona, Universitat Politècnica de Catalunya, dins de l'especialitat de l'Enginyeria del Software. Aquest projecte està dirigit per Marc Alier Forment, del Departament d'Enginyeria de Serveis i Sistemes de la Informació(ESSI).

1.1 Context

En els últims anys, els avenços de la intel·ligència artificial han fomentat el desenvolupament de models Large Language Models que han obtingut una capacitat, cada cop més gran, per a entendre i generar textos amb un grau de coherència i precisió elevadíssims. Però aquests models també tenen altres aplicacions que estan fora de l'àmbit del llenguatge natural. Un d'aquests àmbits és l'extracció d'informació en diversos formats, com per exemple, imatges, taules, textos. També es poden emprar per a fer anàlisis de dades en grans conjunts de dades o conjunts de dades amb formats difícils de processar per a les persones.

Actualment, la gestió i recuperació d'informació afrontant el repte de gestionar grans quantitats de dades en diferents tipus de formats com els que hem comentat anteriorment, text, imatges i taules. Tradicionalment, els sistemes de recuperació d'informació es basen en la cerca textual, això no és suficient per a processar els diferents tipus de dades que tenim en l'actualitat i és en aquest context que sorgeix la necessitat de millorar els models actuals i millorar l'eficiència en aquesta recuperació d'informació multimodal amb l'objectiu de millorar les consultes sobre les bases de dades semàntiques.

1.2 Identificació del problema

Donats els fets exposats en l'apartat anterior, el creixement en la mida de les dades a tractar pels sistemes actuals ens ha portat a un augment del nombre de documents que combinen diferents tipus de formats de dades. Tot i això, els sistemes actuals no són capaços de processar aquests diferents formats, ja que, estan dissenyats per a fer cerques textuals. Això redueix la seva capacitat de comprendre i processar les dades de manera semàntica.

El problema principal recau en el fet que la majoria de les bases de dades semàntiques utilitzen estratègies de cerca de coincidència exacta en paraules clau i, per tant, es dificulta la interpretació quan la informació es presenta en formats de complexitat més elevada.

Un dels altres problemes de l'enfocament tradicional és la integració de diferents tipus de dades en un mateix entorn de consulta. Això representa un problema actualment, ja que no hi ha metodologies estandarditzades per a fer les extraccions de dades de la informació multimodal de manera eficient. Per tant, sense una bona extracció i representació de les dades multimodals, les consultes a les bases de dades semàntiques perden precisió, afectant l'eficiència i la qualitat dels resultats obtinguts per aquestes consultes.

Com a resultat, aquest projecte intentarà generar un pipeline, amb l'ajut d'un model o la conjunció de varis, que permeti reduir al màxim aquests problemes exposats.

1.3 Actors implicats

Desenvolupador: El projecte està realitzat per un desenvolupador, autor d'aquest treball. El desenvolupador és l'encarregat de dissenyar, desenvolupar i documentar el model.

Director del treball de fi de grau: El projecte és dirigit per en Marc Alier Forment, professor del Departament d'Enginyeria de Serveis i Sistemes de la Informació (ESSI). Encarregat de guiar i supervisar el correcte desenvolupament del projecte.

Especialistes en inteligencia artificial: Donat que el projecte es desenvolupa en el marc de les intel·ligències artificials, en concret dins de l'àmbit dels Large Language Models, els especialistes en intel·ligència artificial desenvolupen i document les diferents tècniques i models i, per tant, són part interessada en el projecte.

Empreses que analitzen grans quantitats d'informació: Qualsevol empresa que tingui la necessitat d'analitzar, gestionar grans quantitats d'informació en documents és part interessada en implementar el sistema del projecte.

Desenvolupadors de Codi Obert: Tots els desenvolupadors de les comunitats Codi obert interessants en reutilitzar part del codi implementat en el projecte amb l'objectiu de millorar o crear nous sistemes basats en el del projecte.

Usuaris finals: Aquesta categoria inclou tots aquells professionals en l'àmbit de la tecnologia i que fan ús de la intel·ligència artificial per a realitzar gestió documental, investigació. Anàlisis de dades i altres aplicacions.

1.4 Definicions

IA: Sigles d'intel·ligència artificial.

Model: Estructura entrenada amb dades per a executar tasques específiques.

LLM: Sigles de Large Language Model.

Large Language Model: Model d'intel·ligència artificial basat en xarxes neuronals, entrenat per a entendre i generar llenguatge natural.

Multimodalitat: Capacitat d'un sistema d'IA per a processar i combinar diferents tipus de dades.

Embeddings: Representacions numèriques d'un espai vectorial amb significats semàntics com paraules, imatges, etc.

Pipeline: Seqüència de passos d'un flux de processament de dades que transforma analitza i modela la informació.

Tasques NLP: Conjunt de processos dins del processament de llenguatge natural.

Base de dades semàntica: Sistema d'emmagatzematge optimitzat per a organitzar la informació basant-se en el significat o el context de les dades en lloc de coincidències exactes.

Repositori de codi: Espai on s'emmagatzema i es gestionen les versions del codi font d'un projecte.

Branca: Línia de desenvolupament dins d'un repositori.

2. Justificació

2.1 Solucions prèvies

El camp de la intel·ligència artificial existeix des de fa més de cinquanta anys i ha anat evolucionant a passos agegantats des del seu inici. Per això, prèviament a fer aquest treball he recollit algunes solucions, tant de codi obert com propietàries, que utilitzen els Large Language Models per a extreure informació de documents i generar-ne embeddings multimodals.

La primera solució és el model Luminous de la companyia alemanya Aleph Alpha. Luminous destaca per la seva capacitat de processar tant text com imatge, de manera multimodal. Això els permet integrar diferents tipus de dades en una única representació i facilita la recuperació de la informació i les anàlisis semàntiques de les dades generades pel model.

Una altra solució de codi obert és Spark NLP. És una biblioteca de processament de llenguatge natural que ofereix pipelines i models preentrenats per a resoldre tasques NLP. Spark NLP consta d'una extensió anomenada Spark OCR que permet extreure text d'imatges i documents escanejats i d'aquesta manera, generar els embeddings multimodals a partir de text i imatges.

Per una altra banda, tenim Amazon Titan. Titan forma part de la suite de models que ofereix Amazon i conté models multimodals que poden processar text i imatges per igual. Tots aquests models estan dissenyats per a integrar-se en aplicacions amb requisits de comprensió i generació de continguts i que reben les dades per diferents canals.

Una de les solucions més importants actualment és Gemini. Aquest model Large Language Model està dissenyat per Google i compte amb una gran capacitat per a processar tota mena de dades, per exemple text, imatges, àudio i vídeo entre d'altres. Aquesta multimodalitat converteix a Gemini en una de les solucions més robustes del mercat, pel que a multimodalitat es refereix.

Finalment, OpenAI ofereix diferents serveis que permeten la generació d'embeddings multimodals. Tot i que el seu producte estrella, ChatGPT 4 compta amb la capacitat de processar tant text com imatges no genera embeddings multimodals de manera directa i, per tant, no seria una solució a tenir en compte. Però sí que hi ha altres serveis que generen aquests embeddings i poden ser interessants en el context del nostre projecte.

Per exemple, l'API de embeddings d'OpenAI, tot i que actualment només compte amb la generació d'embeddings de text es preveu que pugui generar embeddings d'imatges i taules en un futur.

OpenAI també ofereix un model anomenat CLIP (Contrastive Language-Image Pretraining) que té per objectiu generar representacions multimodals imatge-text que permetin fer cerques semàntiques de manera eficient.

2.2 Justificació

Un cop exposades les solucions que hi ha al mercat actualment i tenint en compte el que aporten i el que no, podem afirmar que, no existeix una solució que, de manera eficient, processa la informació recollida en documents en diferents modalitats i en concret text, imatges i taules i posteriorment genera embeddings multimodals que permeten fer consultes sobre les bases de dades semàntiques on s'emmagatzemen aquests embeddings.

Això obre la porta a la creació d'un nou model o conjunció de diversos models en un nou sistema que ens permet fer aquest processament multimodal dels documents i la generació d'embeddings multimodals de text, imatge i taules. Aquest sistema permetria unificar el processament de dades i la cerca d'aquestes, millorant l'eficiència del procés.

A més, aquesta solució permet simplificar el procés d'integració amb altres eines o sistemes que requereixin diferents models per a fer el processament i generació d'embeddings de les dades que tracten. Unificant tot el processament en un mateix sistema i eliminant possibles incompatibilitats entre altres sistemes.

Per tant, la realització d'aquest projecte implica una millora en les actuals solucions que s'ofereixen en el mercat, tant de codi obert com propietàries.

3. Abast

Aquesta secció està dedicada a l'especificació dels objectius i subobjectius que es pretenen assolir amb aquest projecte, els diferents impediments i riscos que poden sorgir durant tot el procés de desenvolupament del projecte.

3.1 Objectius

El principal objectiu del projecte és desenvolupar un sistema que implementi un o més models Large Language Model que permeti extreure la informació de tipus text, imatge o taules de documents i la processi per a obtenir embeddings multimodals (únicament en les modalitats definides abans text, imatge i taules). Aquests embeddings han de ser compatibles amb la base de dades semàntica escollida.

Per tal d'aprofundir una mica més en aquest objectiu ens cal definir alguns subobjectius que ens permetin especificar una mica més.

- Crear un sistema que implementi un o més models LLM (Large Language Model).
- Dotar aquest sistema de la capacitat de rebre les dades en el format correcte (documents).
- Implementar el codi que permet generar els embeddings multimodals.
- Assegurar que els embeddings generats són compatibles amb les bases de dades utilitzades.

3.2 Impediments i riscos

En tots els projectes apareixen obstacles i impediments que dificulten les tasques pendents de realitzar per a assolir els objectius marcats a l'inici. Per tant, és important estar preparats per a afrontar aquests obstacles de la millor manera i reduir l'impacte en el projecte.

Seguidament, s'exposen alguns dels possibles impediments i riscos que es poden trobar durant el desenvolupament del projecte.

Limitacions de recursos i temps: Els recursos limitats i la manca de temps degut als terminis ajustats del projecte poden resultar en una limitació de capacitat per a assolir tots els objectius marcats en el projecte de manera satisfactòria.

Desafiaments d'implementació: Els problemes de seguretat, eficiencia o comptabilitat en les tecnologies emprades per al desenvolupament poden representar un obstacle en el correcte desenvolupament del projecte.

Inexperiencia en les tecnologies: La falta d'experiència en talguna de les tecnologies emprades per l'equip de desenvolupament poden resultar en retards significatius en les tasques derivades del projecte.

Requeriments computacionals: Alguns dels models emprats per a l'extracció d'informació i generació d'embeddings són molt demandants de la maquinaria, això podria alentir el procés de proves de les diferents versions.

4. Metodologia

Amb l'objectiu de completar un projecte d'aquestes dimensions és necessari definir una metodologia de treball per tal de tenir un control de les tasques i els recursos i, d'aquesta manera, guiar el procés de desenvolupament mantenint el control i la qualitat esperada.

Tot seguit es defineix aquesta metodologia de gestió.

4.1 Gestió del projecte

La metodologia escollida per a fer la gestió del projecte és la metodologia ***Kanban***, que és una gestió visual basada en columnes i targetes. Aquest mètode permet fer un seguiment molt clar del progrés del projecte i veure quines tasques queden pendents per a completar els objectius. Per tal de fer ús d'aquesta metodologia usarem l'eina ***Trello***, on les diferents columnes representen els possibles estats de les tasques i les targetes les mateixes tasques. D'aquesta manera facilitem la priorització de les tasques més importants o urgents i aquelles que poden bloquejar altres tasques. A més, ***Trello*** ens permet definir en cada targeta (tasca) tot un seguit d'informació addicional i camps que són molt útils per a fer una gestió eficient d'aquestes.

Els possibles estats de les tasques són els següents:

Pendent: Tasques no començades.
En Procés: Tasques ja començades.
Bloquejat: Tasques que no es poden realitzar per un motiu.
Proves: Tasques pendents de ser provades o en proves.
Validació: Tasques pendents de validar abans de donar-les per acabades.
Finalitzada: Tasques acabades.

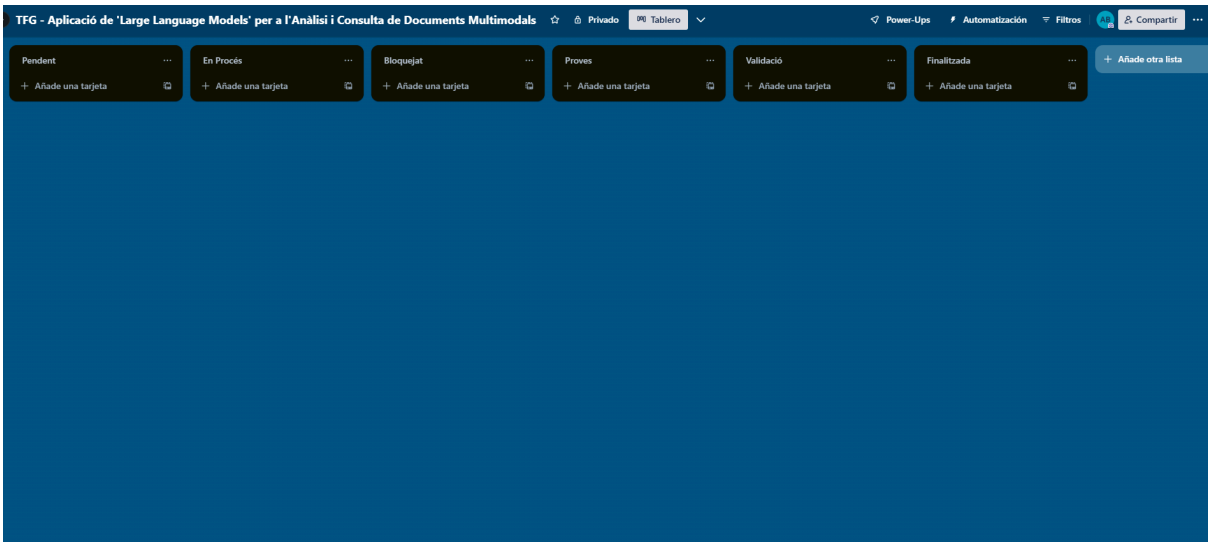

Figura 4.1: Tauler del Trello del projecte. Font: elaboració pròpia

4.2 Control de versions

Per a dur un control de versions del codi font del sistema s'utilitzarà *Git*. *Git* és una eina de control de versions de codi obert que permet fer un seguiment dels canvis en el codi font durant tot el projecte. Permet, addicionalment, treballar de forma col·laborativa amb altres desenvolupadors mitjançant l'ús de repositoris i branques.
Hi ha diferents plataformes que permeten usar *Git*, la més important és *Github*, que serà la que farem servir en aquest projecte. *Github* és un gestor de repositoris *Git*.

Dins del sistema de *Git* existeixen diferents maneres de treballar. En aquest projecte usarem *Gitflow*, que és un model de ramificacions que permet treballar de manera clara i estructurada. La utilització d'aquesta metodologia implica la definició d'una seria de branques i regles per a la correcta gestió. Les branques definides per *Gitflow* que s'implementaran en el projecte són les següents:

Main: És la branca principal del projecte i la que conté el codi estable del projecte. Els diferents commits que rebi aquesta branca representen les versions de codi que han estat provades i estan llestes per a implementar en producció.

Develop: És la branca principal del desenvolupament del projecte. Els diferents commits que rebi la branca seran les funcionalitats acabades i pendents de provar en el conjunt del sistema.

Tasca: Per a cada tasca de desenvolupament es crearà una nova branca amb el nom de la tasca a realitzar en ella. Rebrà els commits relacionats amb la funcionalitat de la tasca i és on es faran les primeres proves. Aquesta branca es crearà des de la branca develop sempre.

Fix: Les branques de fix s'utilitzaran per a solucionar els problemes crítics que puguin aparèixer a la branca Main i, per tant, es crearan a partir d'aquesta i sempre retornaran a main. És important que aquestes branques es tornin a ajuntar amb Main i no amb altres branques que no continguin les versions estables del projecte.

4.3 Validació i proves

També és important, per tal de mantenir la qualitat durant tot el procés de desenvolupament, fer proves del codi i validar que les tasques fetes compleixen amb els criteris que s'esperen. Per tant, es definiran un conjunt de proves dins l'entorn de *Github*, utilitzant els sistemes de *Github Actions* per a assegurar que el codi del repositori funciona correctament.

Addicionalment i per a fer la validació de les tasques, durant el desenvolupament del projecte es mantindrà una comunicació regular entre el desenvolupador i autor del projecte i el director d'aquest, el Marc Alier Forment, amb l'objectiu de garantir el compliment dels objectius definits anteriorment amb la qualitat esperada. Per a dur a terme aquestes validaciones es convocaran reunions periòdiques per a discutir els avenços del projecte i les següents passes d'aquest.

5. Referències

Li, X., Zhang, Y., & Wang, J. (2023). Multimodal Large Language Models: A Survey. arXiv. https://arxiv.org/abs/2311.13165

Chen, L., Huang, T., & Zhao, M. (2024). Leveraging Large Language Models for Multimodal Search. arXiv. https://arxiv.org/abs/2404.15790

Ghosh, A., & Patel, S. (2021). Efficient Multi-Modal Embeddings from Structured Data. arXiv. https://arxiv.org/abs/2110.02577

Rodríguez, P., & Martínez, L. (2024). Semantic-Aware Representation of Multi-Modal Data for Data Ingress: A Literature Review. arXiv. https://arxiv.org/abs/2407.12438

Management Solutions. (2023). El auge de los large language models. Management Solutions. https://www.managementsolutions.com/sites/default/files/minisite/static/72b0015f-39c9-4a52-ba63-872c115bfbd0/llm/pdf/auge-de-los-llm-03.pdf

Google Cloud. (2024). Obtén incorporaciones multimodales. Google Cloud. https://cloud.google.com/vertex-ai/generative-ai/docs/embeddings/get-multimodal-embeddings?hl=es-419

García, J., & López, M. (2023). Bases de datos semánticas: Evolución y aplicaciones. Revista de Software y Bases de Datos, 15(2), 45-67. https://revistas.unla.edu.ar/software/article/view/1278/1113

Pérez, A., & Ramírez, C. (2021). Sistemas de extracción de información: Procesos y técnicas aplicadas. Universidad de Córdoba. https://www.uco.es/investiga/grupos/labindoc/images/files/neoinstrumenta/articulos/neoins2013n3.pdf

Fernández, L., & Torres, R. (2020). Extracción y representación de conocimiento a partir de corpus textuales. ResearchGate. Recuperat de https://www.researchgate.net/publication/265293565_Extraccion_y_representacion_de_conocimiento_a_partir_de_corpus

Zhang, Y., & Liu, H. (2024). Semantic-Aware Representation of Multi-Modal Data for Data Ingress: A Literature Review. arXiv preprint. https://arxiv.org/abs/2407.12438

Kim, D., & Lee, J. (2023). Multimodal Neural Databases: An Approach to Multi-Modal Data Querying at Scale. arXiv preprint. https://arxiv.org/abs/2305.01447

Wikipedia. (2024). Aleph Alpha. Wikipedia, la enciclopedia libre. https://es.wikipedia.org/wiki/Aleph_Alpha

Wikipedia. (2024). Spark NLP. Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Spark_NLP

Wikipedia. (2024). Gemini (modelo de lenguaje). Wikipedia, la enciclopedia libre. https://es.wikipedia.org/wiki/Gemini_%28modelo_de_lenguaje%29

Amazon Web Services (AWS). (2024). Amazon Titan Models. AWS Bedrock. https://aws.amazon.com/es/bedrock/amazon-models/titan/

OpenAI. (2021). CLIP: Connecting text and images. OpenAI. https://openai.com/index/clip/

OpenAI. (2021). CLIP (Contrastive Language-Image Pre-Training). GitHub. https://github.com/openai/CLIP

Atlassian. (s.f.). Flujo de trabajo de Gitflow. Atlassian. https://www.atlassian.com/es/git/tutorials/comparing-workflows/gitflow-workflow