Anonimització automatitzada de dades en documents administratius

L'objectiu d'aquest document és explicar els punts clau que hauran de recollir-se perquè una Prova de Concepte (PdC) quedi correctament documentada per a la seva possible utilització posterior. Es detallen a continuació cadascun dels apartats clau.

1 Introducció

- ### Objectiu de la PdC

Desenvolupar una solució basada en IA per anonimitzar automàticament dades personals en documents administratius, millorant l'eficiència i la seguretat en la publicació d'informació pública.

- ### Context i justificació

La PdC s'emmarca en el compliment de la normativa de protecció de dades (LOPDGDD, RGPD) i de transparència (Llei 19/2013), i respon a la necessitat d'agilitzar l'anonimització de documents per garantir el dret d'accés a la informació pública.

- ### Abast de la prova de concepte

La PdC s'ha aplicat a una mostra de documents reals de la Diputació de Barcelona, centrant-se en documents PDF verticals i en un conjunt limitat de dades personals.

2 Descripció del problema

- ### Problema a resoldre

L'anonimització manual de documents és lenta i poc escalable. Cal una solució automatitzada que garanteixi la protecció de dades i faciliti la transparència.

- ### Serveis afectats

Tots els serveis que generen documents susceptibles de ser publicats o compartits amb la ciutadania.

- **Beneficis esperats de la solució**
    - Pel ciutadà
      Millora de l'accés a la informació pública.
    - Pel personal de l'administració
      Reducció de tasques manuals de poc valor.
    - Altres possibles beneficis i/o beneficiaris
      Millora de la transparència i del govern obert.

3 Escenari de la prova de concepte

- **Requisits Funcionals / No Funcionals**

  Requisits funcionals:

  - Detecció i anonimització automàtica de dades personals en PDF
  - Tipus de dades: noms, DNI, adreces, matrícules, empreses (en contextos específics)
  - Interfície intuïtiva per carregar i descarregar documents
  - Manteniment del format original
  - Autenticació d'accés i control de permisos

  Requisits no funcionals: consideracions sobre rendiment, seguretat, etc

  - Compliment del RGPD
  - Alta fiabilitat i baixa taxa d'error
  - Escalabilitat i compatibilitat amb navegadors
  - Traçabilitat i registre d'operacions
  - Optimització de recursos i mantenibilitat

- **Recursos necessaris (humans, tècnics, financers)**

  Humans: Personal de la DIBA i de l'empresa proveïdora (Omnios)
  Tècnics: Infraestructura AWS (S3, Lambda, DynamoDB, SQS, API Gateway, Cognito, CloudWatch)
  Financers: 10.674,29 € (IVA inclòs)

- **Ajuntaments i/o organismes col·laboradors a la PdC (funcions i contactes)**

  No aplica

4 Metodologia

- **Enfocament General (e.g., supervisat, no supervisat)**
  - Finetuning d'un LLM: Entrenament amb documents etiquetats per detectar dades sensibles
  - Prompting avançat: Instruccions específiques per a cada tipus de dada
  - Adaptabilitat: Ajust del model segons el tipus de document

- **Eines i tecnologies utilitzades**
  - Frontend: React + TypeScript
  - Backend: Python + Poetry
  - Infraestructura: AWS (serverless)
  - Seguretat: Cognito + CORS

- **Procés de desenvolupament i proves**
  - Desenvolupament iteratiu amb validació contínua
  - Proves amb documents pseudoanonimitzats reals

5 Implementació

- **Descripció de la solució proposta**
  - Plataforma web per carregar documents PDF
  - Anonimització automàtica amb LLM (Claude 3.5)
  - Aplicació de màscares sobre el document
  - Descàrrega del document anonimitzat

- **Arquitectura del sistema**
  - Frontend allotjat a S3
  - Backend amb API Gateway i Lambdas
  - Processament asíncron amb SQS
  - Emmagatzematge a S3 i metadades a DynamoDB

- Monitorització amb CloudWatch

6 Flux de dades i algorismes utilitzats

- **Recol·lecció de dades**
  - Documents reals de la DIBA amb dades pseudoanonimitzades
  - Tipologies: informes, resolucions, comunicacions, decrets

- **Pre-processament de dades**
  - Conversió de PDF a text
  - Normalització i segmentació
  - Integració de criteris d'anonimització

- **Algorismes Utilitzats**
  - LLM Claude 3.5 per detectar dades sensibles
  - Càlcul de coordenades per aplicar màscares
  - Aplicació de màscares amb Apryse

- **Flux de processament**
  - Pujada del document
  - Extracció de text i detecció de dades
  - Generació de màscares i coordenades
  - Aplicació de màscares en la descàrrega
  - Monitorització contínua

7 Resultats

- **Resultats obtinguts**
  - Alta precisió en PDF verticals
  - Integració tècnica exitosa
  - Limitacions en altres formats i etiquetes

- **Mètriques d'avaluació**
  - Temps de processament
  - Precisió en la detecció

- Taxa d'error
- Escalabilitat

Comparació amb Objectius Inicials

- Objectius assolits parcialment
- Cal ampliar formats i funcionalitats

8 Seguretat i privacitat de les dades. Avaluació de riscos

- Entorn segur (AWS)
- Compliment del RGPD
- Risc de reidentificació si no s'afina l'algorisme

9 Anàlisi i Discussió

Lliçons Apreses

- La IA és viable per a l'anonimització, però cal supervisió
- Cal explorar models open source com Salamandra (projecte AINA)
- Necessitat d'estandarditzar formats i etiquetes

Limitacions de la PdC

- Només s'ha provat amb un tipus de PDF
- No s'han testat llistes blanques ni etiquetes avançades

10 Conclusions

Resum de troballes

- La PdC valida la viabilitat de l'anonimització automatitzada
- Cal afinar el sistema per cobrir tots els requisits normatius

Propers passos suggerits

- Reentrenar el model amb aprenentatge supervisat
- Provar-lo en un entorn acotat (departament pilot)
- Ampliar funcionalitats i formats

11 Recomanacions per a futures Implementacions (**)

- **Requisits funcionals**
  - Compatibilitat amb més formats (PDF escanejat, DOCX, CSV, etc.)
  - Llistes blanques i edició manual de màscares
  - Integració amb gestors d'expedients i portals de transparència

- **Cos ètic i de serveis**
  - Evitar riscos de re-identificació
  - Prevenir biaixos algorítmics
  - Garantir la interoperabilitat i la seguretat

12 Referències (Si aplica)

- **Bibliografia i recursos consultats**

  No aplica

- **Documentació addicional**

  No aplica

13 Annexos (Si aplica)

- **Codis fonts**

  No aplica

- **Dades de prova**

  No aplica

- **Gràfics i visualitzacions**

  No aplica