Enunciat del projecte de PROP
Quadrimestre de tardor, curs 22/23

Gestor de Documents

Donades les següents definicions:

- Document: títol, autor, contingut
- Contingut: seqüència de frases
- Frase: seqüència de paraules
- Paraula: seqüència de caràcters
- Un títol o un autor es consideren una frase

Es tracta de construir un entorn per a la manipulació de documents. El programa ha d'oferir un entorn tant còmode com sigui possible, i obligatòriament ha de:

- Permetre carregar documents (individualment o en conjunt), a partir de diferents formats (com a mínim text pla i xml)[1] i guardar documents a diferents formats (com a mínim text pla i xml). A més a més, s'haurà de poder guardar/recuperar documents en un format propietari del sistema (a decidir per l'equip).
- Permetre la gestió de altes, baixes i modificacions de documents
- Permetre diversos tipus de consultes, com a mínim les següents:
  1. Llista de títols d'un autor
  2. Llista d'autors que comencen per un prefix (que pot ser buit)
  3. Contingut d'un document donat el seu títol i autor
  4. Llista de documents per dos mètodes:
     1. Donat un document D i un natural k, obtenir els k documents (tan sols títol i autor) més semblants a D
     2. Donades una expressió booleana formada pels operadors & | i ! (amb les normes de precedència habituals i la possibilitat de parentitzar per canviar aquesta precedència) i conjunts de paraules (delimitats per {}), seqüències de paraules (delimitades per "'"), o paraules soltes com a operands, obtenir tots els documents que contenen una frase que satisfà aquesta expressió. Un exemple d'expressió seria:
        {p1 p2 p3} & ("hola adéu" | pep) & !joan
        Una frase satisfà aquesta expressió si conté les 3 paraules p1, p2 i p3
        i conté la seqüència "hola adéu" o la paraula pep
        i no conté la paraula joan
     3. Opcionalment, es considerarà també el següent mètode: donades p paraules (denotades col·lectivament com a query), i un enter k, obtenir els k documents més rellevants (en quant a contingut) per aquesta query.
- Permetre la gestió d'expressions booleanes del punt 4.2 anterior, per tal de poder reutilitzar-les (com a mínim alta, baixa i modificació)

Els resultats de les diferents consultes s'han de poder ordenar per diferents criteris, i opcionalment es podria operar també amb aquests resultats.

Per tractar els documents es farà servir el model d'espai vectorial, amb més d'una estratègia per l'assignació de pesos a les paraules dels documents. En el cas de la cerca opcional de l'apartat 4.3, la estratègia d'assignació de pesos a les paraules de la query pot ser diferent a les usades amb els documents.

[1] En el text pla, com a mínim la primera línia contindrà el autor, la segona el títol i a partir de la tercera hi serà el contingut. En xml, hi haurà com a mínim etiquetes títol, autor i contingut.

A més dels altres factors de qualitat de qualsevol programa (disseny, codificació, reusabilitat, modificabilitat, documentació,...), es valorarà l'eficiència, flexibilitat i usabilitat d'aquest. En particular, es pressuposa que caldrà l'ús d'algun tipus d'índex intern per assolir aquesta eficiència. Aquest índex s'haurà de poder conservar d'una sessió a la següent.

Funcionalitats principals a entregar al primer lliurament:

Implementació de la part del controlador (o controladors) del domini que permeti efectuar totes les operacions obligatòries mencionades, exceptuant el primer punt.

Reducció de funcionalitat als equips de 3 persones:

- La cerca d'autors per prefix (apartat 2) és opcional
- La cerca per expressió booleana (apartat 4.2) pot no incloure la parentització per canvi de precedència d'operadors

Dates dels lliuraments:

- Primer: divendres 11 de novembre de 2022
- Segon: divendres 16 de desembre de 2022
- Tercer: divendres 23 de desembre de 2022 (lliuraments interactius: a partir del 9 de gener de 2023)