

Pilot recomanador de llibres

SDG Group



CIDAI-POC-2022-01

Drets reservats. Aquest treball està disponible sota la llicència Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

Segons els termes d'aquesta llicència, podeu copiar, redistribuir i adaptar l'obra amb fins no comercials, sempre que l'obra sigui citada adequadament, tal com s'indica a continuació.

En qualsevol ús d'aquest treball, no s'ha de suggerir que el CIDAI doni suport a cap organització, producte o servei específic. No es permet l'ús del logotip CIDAI.

Si adapteu l'obra, heu de llicenciar-la amb la mateixa llicència Creative Commons o equivalent.

Si creeu una traducció d'aquest treball, heu d'afegir la següent exempció de responsabilitat juntament amb la cita suggerida: "Aquesta traducció no la va crear el Centre of Innovation for Data tech and Artificial Intelligence (CIDAI). CIDAI no es fa responsable del contingut ni de l'exactitud d'aquesta traducció. L'edició original en català serà l'edició autèntica i vinculant".

Qualsevol mediació relacionada amb disputes derivades de la llicència es durà a terme d'acord amb les normes de mediació de la World Intellectual Property Organization.

Cita suggerida. CIDAI-POC-2022-01 // Pilot recomanador de llibres per Joan Salvador Bou, SDG Group, CIDAI, 2022. Llicència: CC BY-NC-SA 4.0



1.	INTRODUCCIÓ	4
2.	RESUM EXECUTIU	5
3.	ENFOCAMENT DE LA SOLUCIÓ	6
4.	IMPLEMENTACIÓ DE LA SOLUCIÓ	7
5.	PROPERES PASSES	14

1. Introducció

1.1. Situació actual

La prescripció o mediació lectora de les biblioteques públiques a Catalunya és segurament un dels serveis que més valor afegeix a les seves múltiples funcions. Trobar el llibre o document en general més adequat és un art que té, també, molta tècnica. Les possibilitats tecnològiques actuals permeten dibuixar un enfocament on es poden aprofitar models algorítmics per dotar a les Biblioteques d'un sistema d'informació que permeti millorar el servei al brindar ofertes personalitzades de llibres.

1.2. Necessitats i objectius

El Departament de Cultura de la Generalitat de Catalunya necessita fer un ús avançat de les dades de les biblioteques públiques que hi ha disponibles per treure'n conclusions i prendre decisions basades en l'evidència estadística. Aquestes necessitats passen per:

- Ser capaços de determinar la importància i potencia de les dades per ajudar a prendre les millors decisions per tal de poder oferir un millor servei a la ciutadania.
- Disposar d'una eina tecnològica destinada a la recomanació de llibres als usuaris habituals de biblioteques. Idealment, en base a les seves pròpies preferències.
- Millorar el format de visualització de les dades mitjançant una eina fàcil d'usuar i de compartir-ne els resultats.

L'enfocament a partir del qual creiem que s'han de satisfer aquestes necessitats passa per aconseguir els següents objectius en aquest projecte:

- Validar que la integritat i la qualitat de les dades és correcta.
- Desenvolupar un algorisme de recomanació de llibres als usuaris de la xarxa de biblioteques de Catalunya
- Construir una eina de visualització de dades per a visualitzar els resultats de l'algorisme.

2. Resum executiu

2.1. Aspectes clau del projecte

Per al desenvolupament del projecte s'han dut a terme diferents anàlisis i processos entre els quals destaquen els següents:

- Realització d'un breu procés d'auditoria de dades on s'han detectat alguns punts de millora en les extraccions, sobretot pel que fa a garantir la qualitat dels registres.
- S'ha acotat el conjunt dels documents potencials als escrits en alguna de les tres llengües que més préstecs concentren: català, castellà i anglès.
- L'aplicació amb èxit de metodologies conegudes en anàlisi textual per obtenir mesures unificades en els títols.
- S'ha limitat l'estudi als 5.000 llibres més prestats, cosa que genera una gran granularitat en les recomanacions.
- La creació d'un model de recomanació per semblança, considerant únicament patrons de comportament similars, que recomana als usuaris 10 llibres d'entre els 5.000 llibres més prestats. Un sistema de recomanació basat en el contingut (en anglès content-based approach) es basa en l'afirmació d'usuaris similars que han llegit llibres similars.
- S'ha pogut millorar significativament la capacitat de l'algorisme mitjançant la cerca d'hiperparàmetres.
- L'algorisme és capaç, en el conjunt de validació, de recomanar el 98% d'aquests llibres.
- Els resultats analitzats demostren que, tot i ser una prova de concepte, les respostes són consistents en termes de complexitat i tipus de llibre concret.

- S'ha utilitzat l'algorisme per a tots els usuaris que hagin fet com a mínim 10 préstecs i que hagin reservat algun dels 5.000 llibres del conjunt seleccionat prèviament.
- La creació d'una eina de visualització que permet veure les recomanacions dels usuaris usant els resultats de l'algorisme.

2.2. Representació visual dels resultats del Recomanador

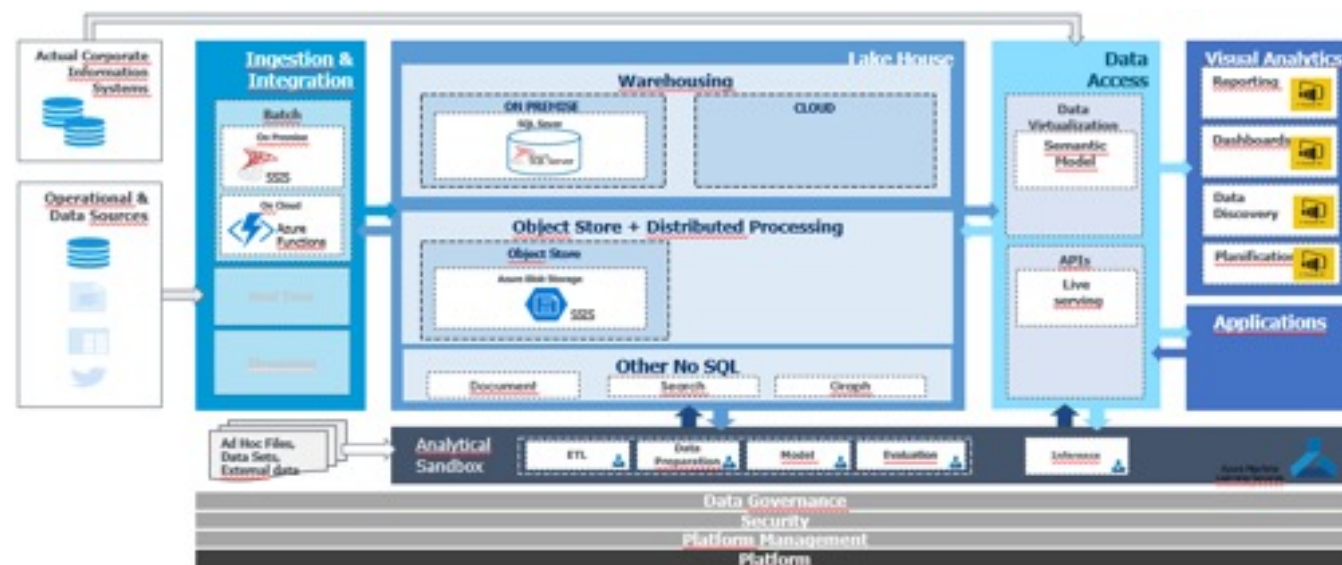
Pel que fa a la representació visual dels resultats del recomanador de llibres, destaquem:

- La visualització s'ha realitzat amb l'eina Power BI.
- Permet validar els resultats de l'algorisme de forma ràpida i senzilla.
- Mostra els 10 primers registres bibliogràfics recomanats per l'algorisme a cada usuari.
- Aspecte visual semblant a la plataforma ARGUS¹.

¹ Argus és el nom que rep el catàleg bibliogràfic de les biblioteques públiques de Girona, Lleida, Tarragona i Terres de l'Ebre i que gestiona el Departament de Cultura de la Generalitat de Catalunya

3. Enfocament de la solució

Per poder convertir el pilot inicial en una realitat, el primer pas ha de ser la implementació d'una ETL que permeti l'extracció, transformació i càrrega de les dades dels fitxers d'entrada a un aparador de dades (datamart). Aquest aparador de dades serà la font d'informació que nodrirà l'algorisme del recomanador (allotjat en Microsoft Azure), el qual un cop processat retornarà un fitxer amb la informació sobre els deu llibres recomanats per a l'usuari en qüestió. Aquest fitxer serà carregat a l'aparador de dades, punt des d'on es visualitzaran els llibres recomanats mitjançant l'eina Power BI.



ETL-Warehousing

- Disseny i creació de l'aparador de dades Biblioteques
- Implementació ETL
- Emmagatzematge en el Microsoft Azure
- Càrrega dades resultat algorisme Recomanador.

Algorisme Recomanador

- Enteniment de negoci
- Disseny pipeline ML tècnic
- Revisió i tractament de dades
- Modelització
- Selecció de l'algorisme
- Inclusió de criteris de negoci
- Desenvolupament i integració pipeline ML

Visualització

- Visualització dels llibres recomanats mitjançant Power BI



4. Discussió i conclusió

4.1. Primeres iteracions a ARGUS

4.1.1 Descobriments de les dades

Durant la fase de descobriment de les dades s'han trobat alguns punts de millora que estan relacionats amb la qualitat de les dades origen i que afecten al rendiment del algorisme. Aquests punts de millora es poden resumir en els següents conceptes:

- Inconsistències als fitxers
- Duplicació de dades
- Camps buits

Per tal de solucionar aquest obstacle s'acorda amb l'àrea funcional el filtratge i la remediació d'alguns valors que permetin millorar la qualitat de la dada.

4.2. Anàlisis i filtres a les dades

4.2.1 Filtres i descobriments als registres bibliogràfics

Processament dels registres bibliogràfics

En el cas dels registres bibliogràfics s'han eliminat tots els documents que no hagin estat reservats des de 2013. Amb aquest filtre s'ha reduït el conjunt de 900.000 documents a uns 300.000 aproximadament.

Després s'han identificat tots els llibres duplicats reescriuint els autors i títols en lletra minúscula.

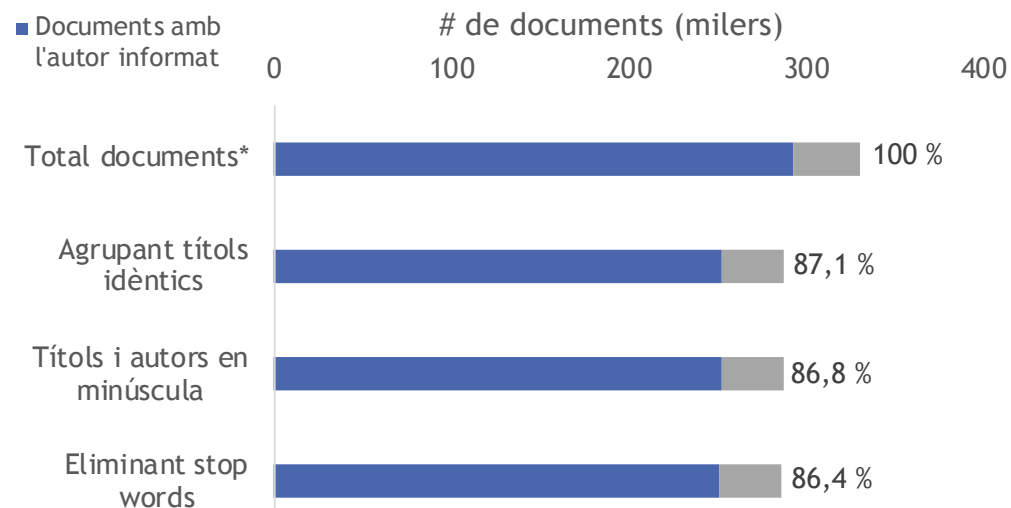
Com que hi ha casos on el títol del llibre no correspon amb l'idioma en què està escrit s'ha avaluat l'efecte d'aplicar paraules buides (stop-words) de diferents idiomes a un mateix títol.

Per avaluar aquest efecte s'han aplicat paraules buides castellanes als títols de llibres en castellà i després s'han aplicat català, de l'anglès i les dues alhora.

En les següents taules resumim els resultats obtinguts al aplicar a títols en castellà les paraules buides d'altres idiomes.

Llistat d'idiomes	Títols afectats
Català	2,9%
Anglès	1,8%
Català i anglès	2,9%

Reducció només en castellà	obra pictòrica completa van eyck
Reducció en castellà i català	obra pictòrica completa eyck



Comentant els resultats obtinguts a l'anàlisi de les dades es va decidir aplicar els següents filtres per construir les dades per al recomanador.

Filtres aplicats a les extraccions d'Argus

- MAT TYPE a
- BIB LVL m
- I TYPE 4
- P TYPE ≤ 6

Filtres aplicats als registres bibliogràfics.

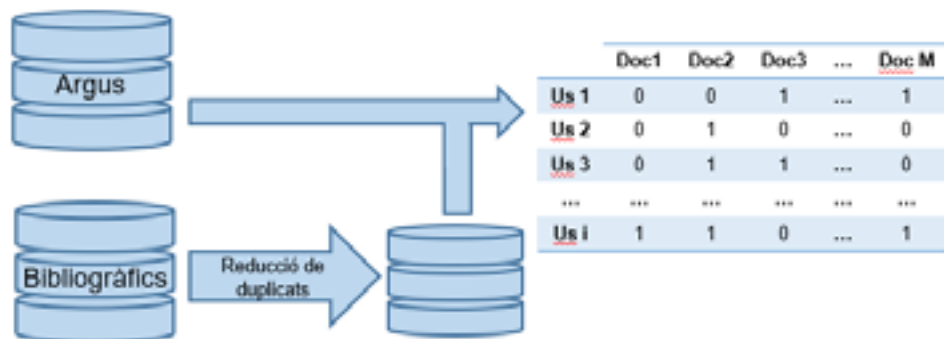
- Eliminació dels duplicats reescriuint els autors i els títols en minúscules.
- Entrenament del recomanador amb els 5.000 llibres que han estat més prestats en el conjunt de dades que tenim.

4.3. Modelització

4.3.1 Breu introducció a un recomanador per similitud

Construcció de les dades: Dades utilitzades per al recomanador

Partim de les dades emmagatzemades a Argus i després del tractament de dades comentat en etapes anteriors, on tenim una base de dades transaccional on es registren tots els préstecs i dels registres bibliogràfics, que és una taula mestra amb la informació dels documents. Amb aquestes dues fonts de dades s'ha creat una matriu on cada fila representa un usuari i cada columna un document diferent.



Aquesta nova representació l'anomenarem taula d'usuari-items (U-I). Per aquesta nova matriu un 0 indicarà que l'usuari no ha fet el préstec

del document i un 1 indicarà que sí l'ha fet. Si un usuari fa el préstec d'un document més d'una vegada també s'indicarà amb un 1. La matriu U-I tindrà 395.050 files (una per usuari) i 276.119 columnes (una per document).

4.3.2 Cerca dels usuaris més similars. Definició de semblança

Concepte de veí: Com identifiquem els veïns?

Per definir una mesura de semblança el primer pas és disposar d'una mesura de distància, així definim la distància entre usuaris com el nombre de préstecs diferents entre els usuaris.

Per exemple:

	Doc1	Doc2	Doc3	Doc4	Doc5
Us1	0	0	1	1	1
Us2	0	1	0	1	0

$d(Us1, Us2)$

Ara bé, la mesura que volem és una de semblança. Per convertir la distància en semblança podem utilitzar diverses famílies de funcions que compleixen que a mesura que la diferència augmenta la semblança disminueix.

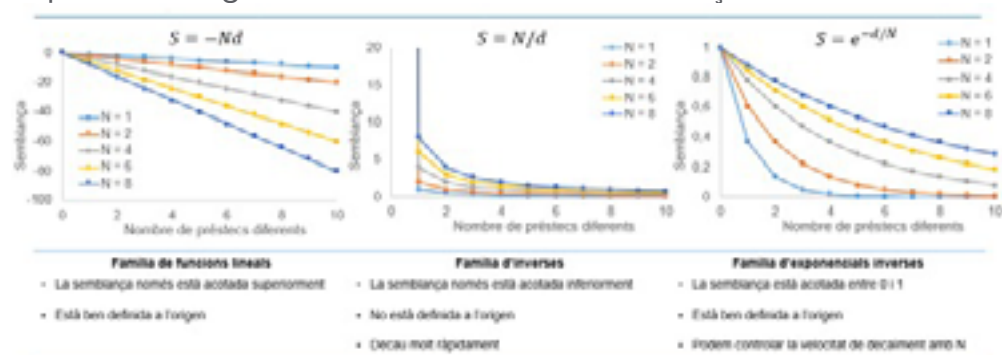
Entre les possibles candidates estan les següents funcions:

- La família de funcions lineals
- La família de funcions inverses
- La família de funcions exponencials inverses

Es descarta la distància cosinus perquè té en compte les coincidències. Aleshores, tots els usuaris seran pràcticament iguals perquè hi

haurà moltes coincidències a causa dels llibres que no han llegit i les diferències en els préstecs es diluiran.

Representació gràfica de les funcions de semblança:

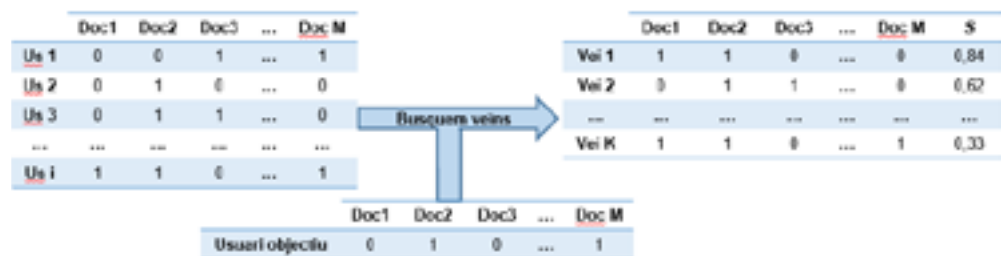


S'escull la família d'exponencials inverses perquè té les millors propietats: acotada, definida i es té control del decaïment.

4.3.3 Construcció de la recomanació

Lògica del recomanador: Selecció del nombre de veïns

Una vegada avaluada la semblança d'un usuari respecte a la resta d'usuaris s'agafen els més semblants i construïm una matriu U-I més petita.



El nombre de veïns que s'ha escollit es determinarà més endavant quan expliquem el procés d'entrenament.

Obtenció de la recomanació

Un cop s'obté el conjunt de veïns calculem un "valor d'usuari mitjà". Per fer aquest càlcul s'utilitzarà la mitjana ponderada on els pesos seran la similitud.

□ Mitjana ponderada:

$$\bar{x} = \frac{\sum_i \omega_i x_i}{\sum_i \omega_i}$$

	Doc1	Doc2	Doc3	...	DocM	S
Vei 1	1	1	0	...	0	0,84
Vei 2	0	1	1	...	0	0,62
...	...	...	...	...	...	...
Vei K	1	1	0	...	1	0,33
Valoració	0,84	0,96	0,32	...	0,78	N/A

□ En el nostre cas:

$$\overrightarrow{Valoració} = \frac{\sum_{i=1}^K S_i \overrightarrow{Vei_i}}{\sum_{i=1}^K S_i}$$

Amb la valoració es pot ordenar els documents de major a menor i recomanar als usuaris els 5 primers (per exemple)

Valoració	0,84	0,96	0,32	...	0,78	N/A	→	Recomanació	Doc2	Doc1	DocM	...	Doc3
-----------	------	------	------	-----	------	-----	---	-------------	------	------	------	-----	------

4.3.4 Modelització

Preparació de les dades: Consideracions a l'hora de preparar l'entrenament

Per entrenar el recomanador s'han fet les següents consideracions amb l'objectiu de reduir-ne els temps i poder obtenir unes mètriques robustes per avaluar el model en diferents aspectes.

L'algorisme podrà treballar amb els 5.000 llibres més recomanats oferint un escenari molt granular de recomanacions.

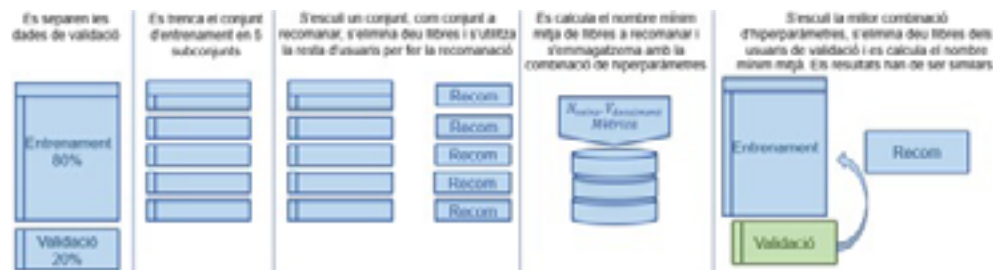
Dins d'aquest subconjunt de llibres només s'utilitzarà per entrenar el recomanador els usuaris que hagin fet més de 50 préstecs.

Per avaluar la "precisió" dels recomanadors s'utilitzarà el nombre mínim mitjà de llibres que cal recomanar a cada usuari per assegurar que al menys un llibre dels recomanats serà un bon candidat. Aquesta mètrica es impossible de calcular exactament així que s'obtindrà una cota superior amb el següent procediment amb les dades de validació:

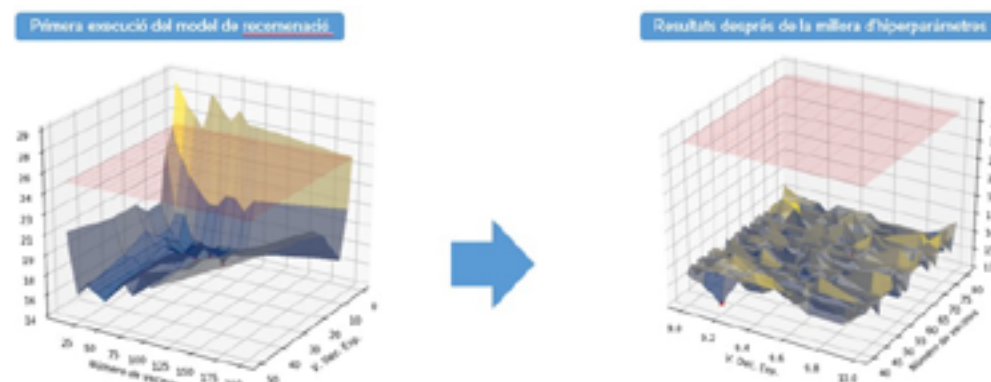
- S'eliminen 10 dels préstecs de tots els usuaris que s'utilitzen per fer la validació
- S'aplica el recomanador a aquests usuaris
- Es compta quants llibres són necessaris recomanar a cada usuari per encertar com a mínim un dels llibres eliminats
- Es calcula la mitjana de tots els usuaris

Entrenament del model: Procés d'entrenament

Per dur a terme el procés per entrenar els models s'ha decidit deixar separat un 20% dels usuaris per fer l'última validació i la resta s'ha utilitzat per obtenir els hiperparàmetres. Per l'elecció dels millors hiperparàmetres s'ha adaptat la lògica de la validació encreuada (cross-validation) a la nostra situació.



Durant el procés de modelització es troben els hiperparàmetres que proporcionen millors resultats en el conjunt de validació i, per tant, proporcionaran les millors recomanacions a futur.



Fàcilment s'observen millores significatives en l'algorisme.

Conclusions i resultats:

Després del procés d'entrenament s'arriba a la conclusió que la millor combinació d'hiperparàmetres és (resultats de la cerca):

- Nombre de veïns: $V = 40$
- Velocitat de decaïment: $N = 9.2$

Resultats del model

Avaluant el nombre mínim mitjan (NMM) de llibres a recomanar obtenim un $NMM = 14$, aquesta mètrica s'ha d'entendre com una aproximació superior de la precisió del model.

Per avaluar la completeness, com que no s'han limitat els préstecs als usuaris, s'avaluarà fent una recomanació per tots els usuaris que tinguin més de 10 préstecs.

El que s'observa és que es recomanen 4.899 llibres dels 5.000 que hi ha disponibles en l'algoritme (98%).

4.4. Visualització

4.4.1 Definició de la visualització

El model generat necessita d'una visualització per validar els resultats de manera gràfica i fàcil d'interpretar. Com l'usuari al final és el ciutadà, s'ha intentat apropar el màxim possible del que seria una visualització per aquest.

Els resultats que es mostren són els 10 bibliogràfics més ben valorats de cara a l'usuari que ha retornat el model. L'usuari pot buscar un ciutadà en concret per veure les recomanacions a aquest, tot i que idealment el ciutadà només hauria de veure les seves recomanacions.

L'objectiu és permetre a l'usuari en qüestió navegar per les seves recomanacions i obtenir informació d'aquest bibliogràfics. D'aquesta manera podrà saber si la recomanació li és interessant a primera vista i procedir a agafar-lo. S'ha pensat en una visualització senzilla i intuïtiva.

Inicialment es mostrarà informació bàsica com la portada, l'autor i el títol de l'obra, però és ampliable a proporcionar més detall.

Els registres bibliogràfics recomanats es troben ordenats en una fila horitzontal estant a l'esquerra el més ben valorat. Cada registre bibliogràfic es troba representat per la seva portada o, en el cas quan no s'ha trobat la portada, pel títol. Les portades s'han obtingut a partir de les URL de les imatges obtingudes via web scraping.

4.4.2 Funcionament de la visualització

S'ha implementat un buscador d'usuaris per trobar les recomanacions fetes a l'usuari entre tots els resultats. L'objectiu final és que aquest filtratge vingui automàticament i l'usuari no hagi de buscar-se a si mateix.

Per tal de trobar les recomanacions d'un usuari s'ha d'escriure al buscador el número de carnet de l'usuari (codi de barres) i aquest filtrarà les recomanacions.

La navegació entre bibliogràfics es fa seleccionant l'element que volem veure i aquest actualitzarà la selecció del requadre inferior. Segons s'ha anat seleccionant diferents bibliogràfics, la portada i la informació mostrada al requadre inferior anirà adaptant-se a la selecció.

S'ha incorporat un botó a la dreta del cercador d'usuari que permet esborrar el codi d'usuari introduït.

4.4.3 Exemple del funcionament

A continuació, mitjançant la visualització, mostrem un exemple que prova la consistència del model.

L'usuari 2500158737 ha llegit els següents llibres:

- Harry Potter i la Cambra secreta
- La Muntanya que parla
- La Volta al món en 80 dies
- El Noi del vestit
- El Misteri del tresor desaparegut
- La Vila del llac dels barrufets
- Els Barrufets jugadors
- Tor: tretze cases i tres morts
- El Somriure de Mona Ratisa
- El Corredor del laberint

Els llibres recomanats inclouen obres de Harry Potter, més llibres dels barrufets com Doctor Barrufet o Els Jocs de la fam llibre similar a El corredor del laberint que l'usuari havia llegit.

Aquest exemple és una petita mostra que es recomanen llibres similars als que es llegeixen.

4.4.4 Punts a tenir en compte

Per a la implementació de la visualització dins del sistema de biblioteques es fa un llistat de punts a tenir en compte de cara a un futur projecte que s'ha identificat en aquesta POC:

- Integrar-la dins de l'espai personal de l'usuari dins de l'ARGUS sense necessitat d'una doble identificació.
- Definir el mètode d'implementació de la visualització, ja sigui un Power BI incrustat a la web o mitjançant un script de programació web.
- Tenir la imatge de la portada de tots els registres bibliogràfics o la màxima quantitat possible. D'aquesta manera farem fàcil la identifi-

cació visual del document per part de l'usuari.

- Proporcionar a l'usuari la disponibilitat de llibre, per exemple amb una URL indicant a quines biblioteques el pot trobar.
- Proporcionar informació extra del bibliogràfic per ajudar a l'usuari a triar el llibre com la següent:
 - Resum o sinopsi
 - Número de pàgines
 - Editorial
 - Idiomes disponibles
 - Codi ISBN
 - Premis i/o guardons obtinguts

5. Properes passes

Les possibles millores per al recomanador de llibres es poden diferenciar dos conjunts principals:

Referents a les dades

- Fer un anàlisi completa i minuciosa del camp matèria (subjecte) dels registres bibliogràfics per crear una mena de descriptors dels documents per poder incloure'ls a l'algoritme de recomanació.
- Incloure informació dels llibres com nombre de pàgines, data de publicació, etc.
- Tenir identificades les diferents traduccions d'un mateix llibre com un sol document.
- Avaluar la possibilitat d'utilitzar la informació dels usuaris per a que la cerca dels usuaris més similars no es basi només en els préstecs
- Estudiar la possibilitat que els usuaris valorin els llibres i utilitzar aquestes valoracions a l'algoritme.

Referents al model (algoritme)

- Entrenar altres algorismes de recomanació com regles d'associació, xarxes bayesianes, etc. estudiant el seu resultat.
- Valorar la lògica de les recomanacions per afavorir els llibres que no solen ser prestats.
- Crear una sèrie de regles pel nombre de recomanacions que es fan als usuaris.
- Utilitzar més d'un model per fer una recomanació, de forma combinada i/o incloure regles de negoci.

