

Université Cadi Ayyad
Ecole Supérieure de Technologie d'Essaouira

Rapport de projet de fin d'étude
Dans le cadre de la formation Informatique Décisionnelle et Science
Des Données (IDSD)

Sous thème

**La réalisation d'une plateforme contre
la violence aux femmes**

Réalisé par

MAKAN Abderrahmane – DRIOUCHE Aymane

Encadré par

Dr. Hanane Grisssete

Année Universitaire : 2023/2024

Motivation

Dans le cadre générale de notre formation informatique décisionnelle et sciences des données (IDSD) au sein de notre Ecole Supérieur de Technologie, les étudiants doivent choisir un projet proposé par les enseignants un projet pour appliquer notre connaissance de cette deux ans et aussi dans un projet de qualité pour faire connaitre un peux la partie pratique et touché le monde de marché, par ce que la vision de cette formation est de former des techniciens supérieur dans de le domaine d'informatique qui peut réaliser des grandes projet et aussi pour suivis les études.

Dans ce contexte-là, l'idée de la création d'une plateforme dédiée à la lutte contre les abus et violences envers les femmes est un domaine de la technologie et de l'analyse des ressources et fournir des méthodologies pour visualiser la violence envers les femmes.

Il existe plusieurs plateformes contre la violence envers les femmes, mais on a vu qu'il existe encore des ambiguïtés dans ce exceptionnel sujet de femme, du coup dans notre projet fin d'étude nous allons proposer une nouvelle idée pour la création de notre plateforme avec des nouvelles outils et de nouvelles visions, en appliquant des nouvelles méthodes et outils et aussi notre modeste connaissance en domaine de l'informatique en général et en l'intelligence artificielle particulier.

Remerciement

Je tiens à remercier Dieu tout-puissant qui m'a donné la force et la patience pour accomplir ce modeste travail.

Tout d'abord, j'adresse mes remerciements et ma profonde gratitude à toutes les personnes qui ont contribué au succès de mon projet et qui m'ont aidé lors de la rédaction de ce rapport. Il est agréable de payer une dette de gratitude au personnel de l'École Supérieure de technologie d'Essaouira et plus particulièrement à tous nos professeurs et superviseurs qui nous ont préparés théoriquement pour passer ce projet. Je souhaite également exprimer mes remerciements pour leur disponibilité et tous leurs précieux et sages conseils, dans ce projet.

Je tiens également à adresser mes plus sincères remerciements à mon encadrante Dr. Hanane Grissette pour leur accueil chaleureux, leur collaboration, leur explication, leur qualités professionnelles et humaines, leur précieux conseils et leur encouragement durant la période du ce projet. elle m'a donné les outils nécessaires pour accomplir les tâches avec plus de succès et d'intérêt.

Enfin, je tiens à remercier tous ceux qui ont contribué directement ou indirectement à l'accomplissement de ce travail trouvent l'expression de nos remerciements les plus chaleureux.

Résumé

Dans notre rapport de projet de fin d'études, nous présentons une plateforme dédiée à la lutte contre les abus et violences envers les femmes. Nous commençons dans notre page principale de plateforme par présenter des statistiques sur la violence envers les femmes, cette statistique basées sur une étude utilisant des données provenant de Twitter. Dans notre plateforme aussi nous proposons également d'autres plateformes existantes dans ce domaine tel que "*Global Fund for Women*" et "*unwomen*". Ainsi que nous proposons des cours et des livres pour des solutions et des conseils pour lutter contre ce type de violence.

Nous cherchons à collecter les données nécessaires à partir de Twitter, puis à les filtrer afin de créer notre propre modèle de traitement. Le but est de visualiser ces informations sur la page principale de notre plateforme. Enfin, nous développons une modèle de recommandation des livres basé sur l'intérêt de chaque utilisateur, visant à fournir le maximum possible de contenu pour chaque utilisateur.

Ces équipements nécessitent des algorithmes et des modèles pour gérer ces idées, ce qui est à la fois efficace et peu complexe.

ABSTRACT

In our end-of-studies project report, we present a platform dedicated to the fight against abuse and violence against women. Our main page starts with statistics on violence against women, based on a study using data from Twitter.

Our platform also offers other platforms in this area, such as "Global Fund for Women" and "UN Women". We then offer books and courses that deal with violence against women and can help women, as well as solutions and tips to combat this type of violence.

We're looking to collect the necessary data from Twitter and then filter it to get a clean data set. The purpose is to view this information on the main page of our platform. Finally, we develop a book recommendation model for each user, where they enter the title of a book and their favorite genre, and then the model recommends books in the same of their interests. This equipment requires algorithms and models to handle these ideas, which is both efficient and minimally complex.

Table des matières

Chapitre 1 : Introduction générale	5
I. Introduction	6
II. Présentation de projet	6
III. Objectif de projet.....	6
i. Objectif personnel	6
ii. Objectif professionnel	6
IV. Fonctionnement	7
V. Planification.....	7
VI. Conclusion.....	7
Chapitre 2 : Cadre théorique et technique d'étude	8
I. Introduction	9
II. Service des statistiques	9
i. Real-time data collection	9
ii. Data visualisation	11
III. Système de Recommandation.....	13
IV. Les Framework et les outils.....	18
V. Conclusion.....	20
Chapitre 3 : Méthodologie Proposé	21
I. Introduction :	22
II. Méthodologies proposes :.....	22
i. Le modèle de recommandation des livres :	22
ii. Statistiques sur la violence contre les femmes :	32
III. Conclusion.....	35
Chapitre 4 : Expérimentation Analyse de résultats	36
I. Introduction	37
II. Le système de recommandation	37
i. Introduction.....	37
ii. Dataset et la collection des données.....	37
iii. Partie d'analyse des données	39
iv. Partie de prétraitement des données :	40
v. Partie de predction des genres	41
vi. La creation de la fonction de recommandation :	43

Introduction générale

III. Experementation.....	44
IV. Implémentation.....	46
V. Conclusion.....	50
Chapitre 5 : Conclusion.....	51
I. Introduction	52
II. Les perspectives en future	52
III. Conclusion.....	53
Référence	54

Listes des figures

Figure 1 : Architecture de système de recommandation.....	13
Figure 2 : Architecture Content-Based.....	14
Figure 3 : Architecture Collaborative Filtering	16
Figure 4 : Architecture de CBOW	31
Figure 5 : Architecture de Skip-Gram	31
Figure 6 : Affichage des données bassons sur les notations	39
Figure 7 : Affichage des livres bassons sur le genre.....	42
Figure 8 : Metrics pour chaque genre	45
Figure 9 : Page de login	46
Figure 10 : Page inscription	46
Figure 11 : Page de connexion	47
Figure 12 : Page d'accueil	47
Figure 13 : Page des plateformes	49
Figure 14 : Page des livres	49

Liste des abréviations

Abréviation	Désignation
CSS	▪ Cascading Style Sheets
HTML	▪ Hyper Text Markup Language
MVT	▪ Modèle-Vue-Template
MVC	▪ Modèle-Vue-Contrôleur
MIT	▪ Structured Query Language Server
ORM	▪ Object-Relational Mapping
DOM	▪ Document Object Model
API	▪ Application Programming Interface
CMS	▪ Content Management System

Chapitre 1 : Introduction générale

I. Introduction

Dans ce chapitre, nous allons présenter notre contexte d'étude de ce projet, aussi nous présentons notre objectif personnel et professionnel de notre sujet de travail proposé tout en expliquant les fonctionnalités nécessaires pour réaliser ce dernier.

II. Présentation de projet

Dans le cadre de la formation informatique décisionnelle et science des données au sein de l'école supérieur de technologie, les étudiants doivent choisir un projet proposé par les enseignants. Ce projet s'inscrit dans le processus d'apprentissage consistant à valider nos connaissances et concrétiser ce qu'on a appris tout au long de ces deux années de DUT par le biais du projet de fin d'études (PFE).

Notre projet a pour le but de réaliser une plateforme pour les femmes. Autrement dit l'objectif est de mettre en vision la violence contre les femmes, est de proposer quelques outils de solution.

Pour y arriver, nous avons dans un premier temps, fait une analyse des plateformes de femme existantes. Nous avons ensuite défini un modèle de notre plateforme le design et le contexte de plateforme, puis nous avons choisi les différents outils à utiliser et enfin nous avons entamé la réalisation du projet.

III. Objectif de projet

i. Objectif personnel

- La maîtrise des nouveaux langages de programmation utilisée au cours de la réalisation de ce projet.
- Savoir gérer un projet de manière professionnel et structuré.
- Développer nos connaissances techniques et enrichir notre jargon informatique.
- Être créatives et avoir des nouvelles idées.
- Se familiariser avec le travail en équipe et réaliser un projet comme le suivant qui nous aide à faire un échange d'idées et d'information et décortiquer le sujet et de répartir les tâches.

ii. Objectif professionnel

- Le projet vise à aider les individus qui sont toujours en cas de violence.
- Faciliter à l'utilisateur de naviguer dans ce sujet international de femme.

- Créer une interface pour faciliter la rencontre les autres plateformes de même idée.
- Créer une interface pour faciliter la lecture des cours et les conseils de ce sujet.

IV. Fonctionnement

Notre plateforme est conçue comme un espace sûr et utile pour les femmes qui ont fait face à des difficultés ou à des cas de violence. Elle offre aux utilisateurs la possibilité de consulter une variété de ressources telles que des conseils, des cours et des sources de motivation. De plus, les utilisateurs peuvent accéder à d'autres plateformes similaires sur le sujet pour trouver un soutien supplémentaire. En résumé, notre plateforme vise à fournir les services suivants :

- Créer son compte.
- Consulter des livres.
- Visualiser des statistiques.
- Consulter les autres plateforme.

V. Planification

- **1ère étape :** La conception du projet.
- **2ème étape :** la création d'un modèle basé sur les données de Twitter pour générer les statistiques sur la violence contre les femmes dans le monde entier et ces causes.
- **3ème étape :** la suggestion des courses et les plateformes similaires dans cette vision.
- **4ème étape :** la réalisation d'un système de recommandation des livres basé sur les intérêts de l'utilisateur.
- **5ème étape :** la réalisation de la plateforme

VI. Conclusion

A travers ce chapitre, nous avons précisé le contexte général de notre projet, ainsi que notre plan de travail sur lequel nous nous sommes basés au cours de la réalisation de ce projet.

Chapitre 2 : Cadre théorique et technique d'étude

I. Introduction

Avant de commencer le travail, nous présentons quelques outils théoriques utilisées dans notre idée en général et dans ce projet en particulier. Tout d'abord, nous identifions les algorithmes plus utilisés dans cette étude en général, et après notre algorithme et méthode, après tout ça on va voir les Framework et les outils que on a utilisé dans ce projet. Cela nous aidé à maîtriser le travail requis et les mots techniques au cadre de ce projet.

II. Service des statistiques

i. Real-time data collection

a. Twitter API

API Twitter est une interface de programmation d'application fournie par Twitter qui permet aux développeurs d'accéder aux fonctionnalités de Twitter, telles que la lecture et l'écriture de tweets. L'extraction des données à partir de Twitter est couramment utilisée pour la recherche, l'analyse des données ou encore l'analyse des sentiments, la veille concurrentielle, etc. Cette méthode d'extraction n'est pas valide à l'heure actuelle, car l'API officielle de Twitter présente des limitations, notamment en ce qui concerne la rétroaction historique et le nombre de requêtes. Par conséquent, la solution pour extraire les données à partir de Twitter consiste à utiliser une autre méthode, étant donné que l'API Twitter présente des limitations[1].

Après avoir effectué des recherches, nous avons décidé d'utiliser ScraperAPI, un service tiers qui permet d'accéder aux données de Twitter sans les limitations de l'API officielle. Nous avons utilisé ScraperAPI pour intégrer les fonctionnalités de Twitter dans notre propre plateforme. Nous avons essayé d'extraire toutes les données relatives à la violence contre les femmes. On a accédé aux fonctionnalités de Twitter à travers cette méthode, lire les tweets, récupérer les données des utilisateurs, voici une explication détaillée pour cette procédure :

1. **Création d'un compte dans Scrapperapi** : nous devons créer un compte dans ce site web
2. **Création d'une API playground** : Une fois connecté à notre compte de développeur, créez une API.
3. **Obtention de la clé d'API** : Une fois notre application créée, nous obtenons une clé d'API, on va utiliser cette clé pour authentifier notre accès à l'API Twitter.

4. **Utilisation de l'API Twitter :** Utilise une bibliothèque d'accès à l'API Twitter dans le langage de programmation python pour extraire des tweets contenant des mots-clés pertinents concernant la violence contre les femmes. On peut filtrer les données par des hashtags spécifiques, des mots-clés ou des phrases associées à ce sujet.

```
twitter_data=[]
import requests
payload = { 'api_key': 'd8eb5bdac5e6495f2d2217606edcddaf',
            'query': '("violence against women" OR #VAW OR #domesticviolence OR #genderviolence OR #sexualassault
            'num': '100'}
response = requests.get('https://api.scraperapi.com/structured/twitter/search',params=payload)
data=response.json()
data
data.keys()
```

5. **Stockage des données :** Pour stocker les données collectées dans une base de données ou un fichier, en veillant à respecter les politiques de confidentialité et les conditions d'utilisation de Twitter.

```
df=pd.DataFrame(twitter_data)
df.to_excel("violence.xlsx", engine="xlsxwriter")
print("file exported")
```

6. **Analyse des données :** Une fois que nous avons collecté une quantité suffisante de données, on va les analyser pour en avoir des informations significatives. Cela peut inclure l'identification de tendances, de motifs ou d'autres insights sur la violence contre les femmes sur Twitter.

Exemple des données :

	RecordID	Country	Gender	Demographics Question	Demographics Response	Question	Survey Year	Value
0	1	Afghanistan	F	Marital status	Never married	... if she burns the food	01/01/2015	NaN
1	1	Afghanistan	F	Education	Higher	... if she burns the food	01/01/2015	10.1
2	1	Afghanistan	F	Education	Secondary	... if she burns the food	01/01/2015	13.7
3	1	Afghanistan	F	Education	Primary	... if she burns the food	01/01/2015	13.8
4	1	Afghanistan	F	Marital status	Widowed, divorced, separated	... if she burns the food	01/01/2015	13.8

Donc, à travers ces étapes pour collecter les données à partir de l'API Twitter et pour filtrer et analyser les données selon nos besoins, on a procédé à la création d'un fichier de données au format CSV qui contient 12600 rows and 8 columns.

ii. Data visualisation

1. Pourquoi et comment

La visualisation des données est un élément essentiel de l'analyse des données, permettant de présenter de manière claire et concise des informations complexes. Dans le cadre de notre projet sur la violence contre les femmes, nous avons utilisé Plotly plots pour visualiser les données démographiques et les tendances de la violence dans différents pays. Cette partie explique pourquoi nous avons choisi de visualiser ces données, comment nous avons utilisé les Plotly plots pour le faire, et comment nous avons extrait ces visualisations sous forme de fichiers HTML pour les visualiser sur le web.

La visualisation des données à l'aide des Plotly plots nous a permis de mettre en évidence les tendances et les disparités régionales de la violence contre les femmes en fonction de différents facteurs démographiques.

Voici quelques raisons principales pour lesquelles nous avons choisi de visualiser les données :

- **Compréhension claire des données** : Les Plotly plots offrent une représentation visuelle claire des données, ce qui facilite la compréhension des tendances et des modèles.
- **Identification des disparités régionales** : Les cartes choroplèthes nous ont permis d'identifier les pays où la violence est plus prévalente pour certains groupes démographiques, aidant ainsi à orienter les politiques et les interventions de manière plus ciblée.
- **Exploration interactive** : Les fonctionnalités interactives des Plotly plots ont permis aux utilisateurs d'explorer les données de manière personnalisée, en sélectionnant des catégories démographiques spécifiques et en choisissant entre les valeurs maximales et médianes.

- **Engagement du public** : Les visualisations interactives sont plus engageantes pour le public, ce qui peut aider à sensibiliser davantage aux questions de violence contre les femmes.

2. Les étapes de visualisation

Pour visualiser les données générées par notre code python, nous avons suivi les étapes suivantes :

- **1- Extraction des données** : Les données démographiques et les tendances de la violence contre les femmes ont été extraites à partir de rapports nationaux et internationaux. Ces informations incluaient l'âge, le niveau d'éducation, la situation professionnelle, l'état civil, le lieu de résidence, et autres.
- **2- Transformation des données** : Les données extraites ont été converties dans un format approprié pour l'analyse et la visualisation. Cela comprend la manipulation des données afin de les préparer pour être utilisées dans les visualisations.
- **3- Choix des visualisations** : Les visualisations les plus appropriées ont été choisies par nous pour représenter les données. Cela incluait utiliser des graphiques radar pour observer les tendances et les disparités régionales, ainsi que des cartes choroplèthes pour représenter les données géographiques.
- **4- Utilisation de Plotly pour la visualisation** : Plotly a été utilisé comme outil principal pour créer les visualisations interactives. Plotly offre une grande variété de graphiques et de fonctionnalités interactives qui nous ont permis de représenter efficacement les données.
- **5- Création des fichiers HTML** : Après avoir généré les Plotly plots, nous les avons exportés sous forme de fichiers HTML pour les visualiser sur le web. Cette méthode nous a permis de partager facilement nos visualisations avec d'autres utilisateurs et de les intégrer dans des sites web ou des applications web.
- **6- Intégration dans l'application web Django** : Les fichiers HTML des visualisations ont été intégrés dans l'application web *Django* à l'aide de balises `<iframe>`. Cela a permis d'afficher les visualisations de manière interactive dans la page web.

- **7- Personnalisation et interactivité** : Nous avons ajouté des fonctionnalités interactives aux visualisations, permettant aux utilisateurs d'explorer les données en sélectionnant des catégories démographiques spécifiques et en choisissant entre les valeurs maximales et médianes.

III. Système de Recommandation

Un système de recommandation n'est pas rare dans le monde d'aujourd'hui, où les données sont immenses. Presque toutes les grandes entreprises du commerce électronique et de l'industrie du divertissement ont intégré des systèmes de recommandation à leurs sites Web et applications, permettant aux clients de découvrir facilement leurs produits préférés, films, musique et cours. Dans les systèmes de recommandation, il y a 3 types dans le système de recommandation :

- Collaborative Filtering
- Content-Based Filtering
- Hybrid Approach

Les deux approches les plus populaires utiliser sont la recommandation basée sur le contenu et la recommandation basée sur la collaboration.

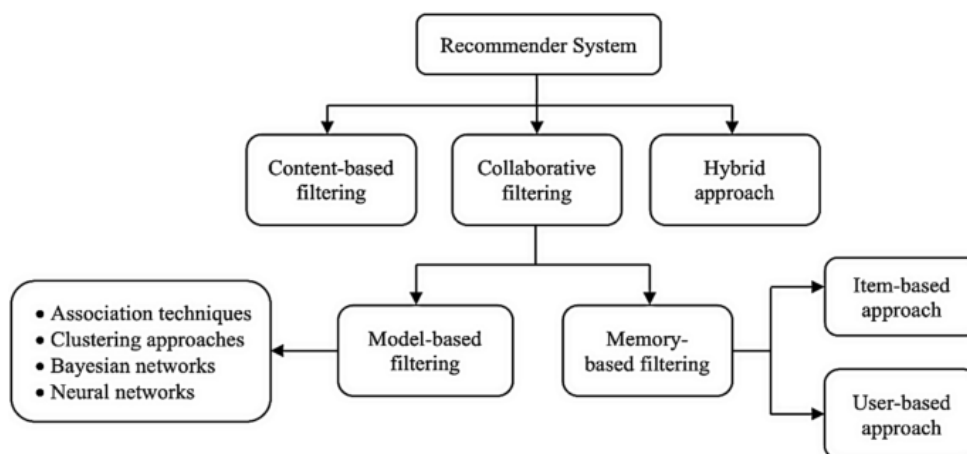


Figure 1 :Architecture de système de recommandation

Dans notre projet en utilisant le modèle word2vec on a créé une système de recommandation basé sur le contenu en utilisant l'ensemble de donnée depuis la grande source de base de données, (Kaggle)

i. Système de recommandation basé sur le contenu :

Dans les systèmes de recommandation basés sur le contenu, tous les éléments de données sont collectés dans différents profils d'éléments en fonction de leur description ou de leurs caractéristiques. Par exemple, dans le cas d'un livre, les fonctionnalités seront l'auteur, l'éditeur, etc. Dans le cas d'un film, les caractéristiques seront le réalisateur, l'acteur, etc. Lorsqu'un utilisateur attribue une note positive à un élément, les autres éléments présents dans ce profil d'élément sont agrégés pour créer un profil utilisateur. Ce profil d'utilisateur combine tous les profils d'articles, dont les articles sont évalués positivement par l'utilisateur. Les éléments présents dans ce profil d'utilisateur sont ensuite recommandés à l'utilisateur, comme le montre la figure.

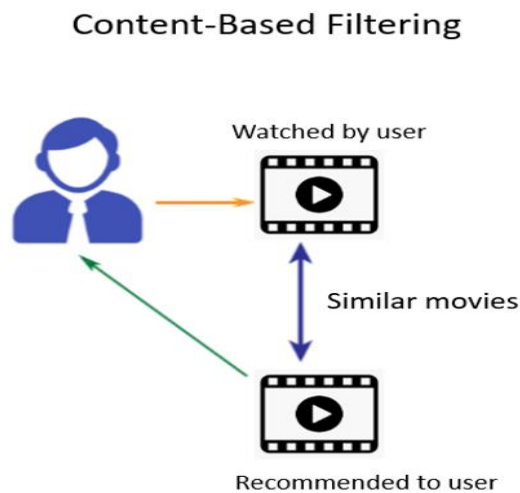


Figure 2 : Architecture Content-Based

ii. Système de recommandation basé sur le filtrage collaboratif :

Les approches collaboratives utilisent la mesure de la similitude entre les utilisateurs. La technique commence par la création d'un groupe ou d'une collection d'utilisateurs X dont les préférences, les goûts, et les aversions sont similaires à celles de l'utilisateur A. X est appelé le voisinage de A. Les éléments qui sont appréciés par la plupart des utilisateurs de X sont ensuite recommandés à l'utilisateur A. L'efficacité d'un algorithme collaboratif dépend de la précision avec laquelle l'algorithme peut le voisinage de l'utilisateur cible. Les systèmes traditionnellement collaboratifs basés sur le « *filtering* » résistent au problème du démarrage à froid et aux problèmes de confidentialité, car il est nécessaire de partager données de l'utilisateur. Cependant, les approches collaboratives ne nécessitent aucune connaissance des

connaissances en matière de fonctionnalités de l'élément pour générer une recommandation. De plus, cette approche peut aider à développer sur les centres d'intérêt existants de l'utilisateur en découvrant de nouveaux éléments[2].

Les approches collaboratives encore une fois divisé en deux types :

- les approches basées sur la mémoire
- les approches basées sur les modèles.

Les approches collaboratives basées sur la mémoire : recommandent de nouveaux éléments en tenant compte des préférences de son voisinage. Ils utilisent directement la matrice d'utilité pour la prédiction. Dans cette approche, la première étape consiste à construire un modèle. Le modèle est égal à une fonction qui prend la matrice d'utilité en entrée :

$$\textbf{Model} = f(\textbf{utility matrix})$$

Ensuite, des recommandations sont faites sur la base d'une fonction qui prend le modèle et le profil utilisateur en entrée. Ici, nous ne pouvons faire de recommandations qu'aux utilisateurs dont le profil d'utilisateur appartient à la matrice d'utilitaires. Par conséquent, pour faire des recommandations pour un nouvel utilisateur, le profil de l'utilisateur doit être ajouté à la matrice d'utilitaires, et la matrice de similarité doit être recalculée, ce qui rend cette technique lourde en calcul.

$$\textbf{Recommendation} = f(\textbf{defined model}, \textbf{user profile})$$

$$\textbf{Where user profile} \in \textbf{utility matrix}$$

Les approches collaboratives basées sur la mémoire sont à nouveau subdivisées en deux type:

- Le filtrage collaboratif basé sur l'utilisateur
- le filtrage collaboratif basé sur les éléments.

Dans l'approche basée sur l'utilisateur, l'évaluation d'un nouvel élément est calculée en trouvant d'autres utilisateurs du quartier de l'utilisateur qui ont déjà évalué ce même élément. Si un nouvel élément reçoit des évaluations positives de la part du voisinage de l'utilisateur, le nouvel élément est recommandé à l'utilisateur. La figure 3 illustre l'approche de filtrage basée sur l'utilisateur.

Collaborative Filtering

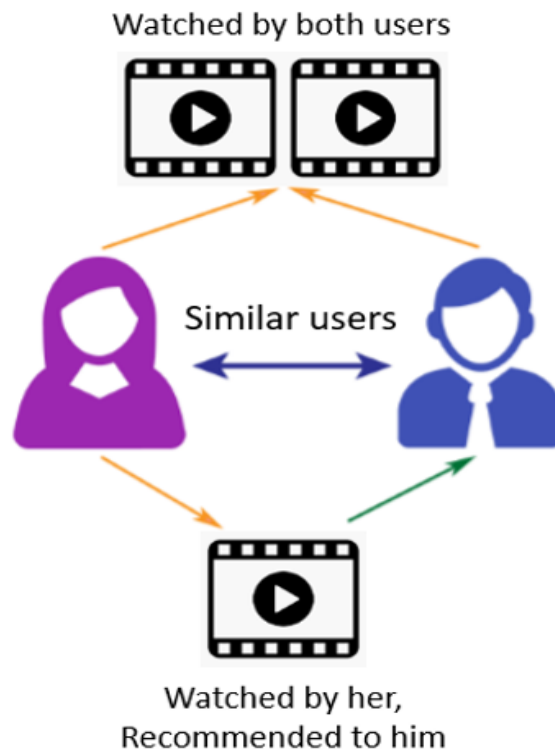
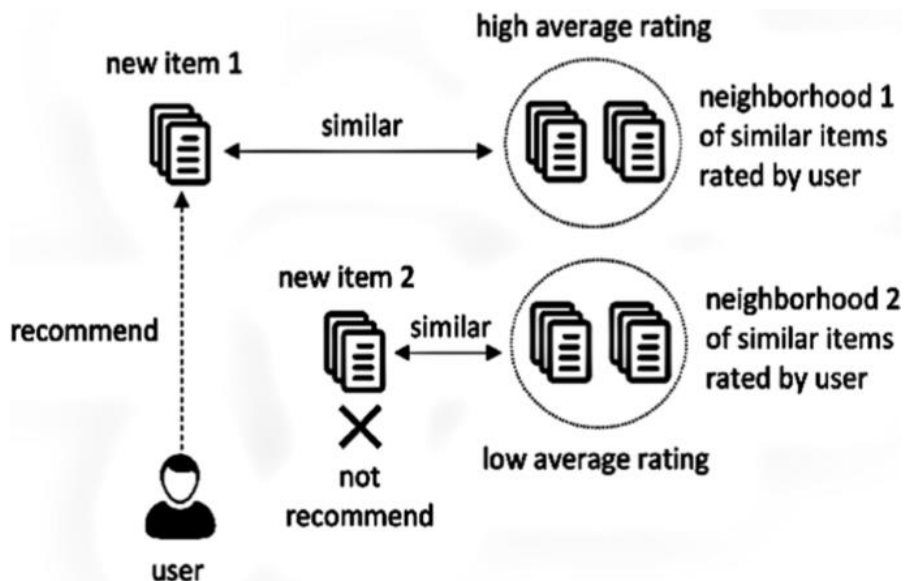


Figure 3 : Architecture Collaborative Filtering

Dans l'approche basée sur les éléments, un quartier d'éléments est construit à partir de tous les éléments similaires que l'utilisateur a évalués précédemment. Ensuite, l'évaluation de cet utilisateur pour un nouvel élément différent est prédite en calculant la moyenne pondérée de toutes les évaluations présentes dans un quartier d'article similaire, comme le montre :



Les systèmes basés sur des modèles utilisent divers algorithmes d'exploration de données et d'apprentissage automatique pour développer un modèle permettant de prédire l'évaluation de l'utilisateur pour un élément non classé. Ils ne s'appuient pas sur l'ensemble du jeu de données lors du calcul des recommandations, mais extraient des entités du jeu de données pour calculer un modèle. D'où le nom de technique basée sur les modèles. Ces techniques nécessitent également deux étapes pour la prédiction :

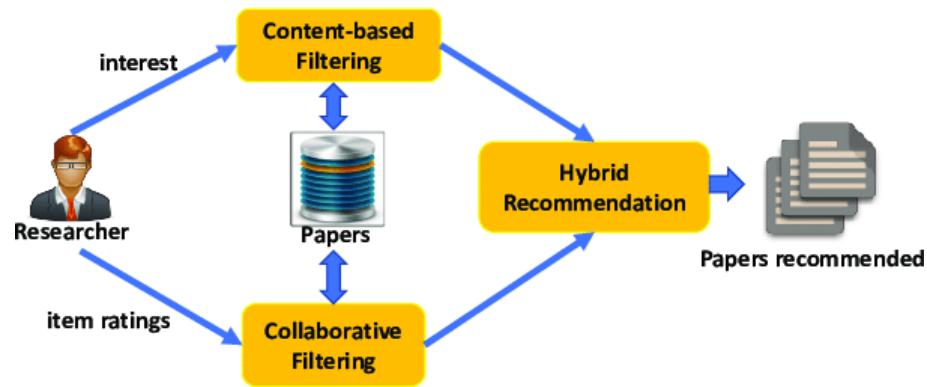
- la première étape consiste à construire le modèle
- la deuxième étape consiste à prédire les évaluations à l'aide d'une fonction (f) qui prend en entrée le modèle défini lors de la première étape et le profil utilisateur.

Les techniques basées sur des modèles ne nécessitent pas d'ajouter le profil utilisateur d'un nouvel utilisateur dans la matrice d'utilité avant d'effectuer des prédictions. Nous pouvons faire des recommandations même aux utilisateurs qui ne sont pas présents dans le modèle. Les systèmes basés sur des modèles sont plus efficaces pour les recommandations de groupe. Ils peuvent rapidement recommander un groupe d'éléments à l'aide du modèle pré-entraîné.

La précision de cette technique repose en grande partie sur l'efficacité de l'algorithme d'apprentissage sous-jacent utilisé pour créer le modèle. Les techniques basées sur des modèles sont capables de résoudre certains problèmes traditionnels des systèmes de recommandation tels que la parcimonie et l'évolutivité en utilisant des techniques de réduction de dimensionnalité et des techniques d'apprentissage des modèles.

iii. Filtrage hybride :

Une technique hybride est une agrégation de deux ou plusieurs techniques employées ensemble pour remédier aux limites des techniques de recommandation individuelles. L'incorporation de différentes techniques peut être réalisée de différentes manières. Un algorithme hybride peut incorporer les résultats obtenus à partir de techniques distinctes, ou il peut utiliser le filtrage basé sur le contenu dans une méthode collaborative ou utiliser une technique de filtrage collaboratif dans une méthode basée sur le contenu. Cette incorporation hybride de différentes techniques se traduit généralement par une augmentation des performances et une précision accrue dans des nombreuses applications de recommandation. Certaines des approches d'hybridation sont le méta-niveau, l'augmentation des caractéristiques, la combinaison de caractéristiques, l'hybridation mixte.



IV. Les Framework et les outils

i. HTML :

Le (HyperText Markup Language), généralement abrégé HTML ou dans sa dernière version HTML5, est le langage de balisage conçu pour représenter les pages web. C'est un langage permettant d'écrire de l'hypertexte, d'où son nom. HTML permet également de structurer sémantiquement la page, de mettre en forme le contenu, de créer des formulaires de saisie, d'inclure des ressources multimédias dont des images, des vidéos, et des programmes informatiques [3].



ii. CSS :

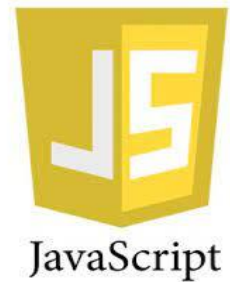
Les feuilles de style en cascade, généralement appelées CSS de l'anglais (Cascading Style Sheets), forment un langage informatique qui décrit la présentation des documents HTML et XML. Les standards définissant CSS sont publiés par le World Wide Web Consortium(W3C). Introduit au milieu des années 1990, CSS devient couramment utilisé dans la conception de sites web et bien pris en charge par les navigateurs dans les années 2000 [4].



iii. JavaScript :

JavaScript est un langage de programmation de haut niveau utilisé principalement pour créer des interactions dynamiques sur les sites web. Il s'exécute dans les navigateurs web, permettant de rendre les pages web interactives avec des fonctionnalités telles que des formulaires interactifs, des animations, des effets visuels et des jeux. JavaScript est essentiel pour améliorer l'expérience utilisateur en permettant aux pages web de réagir

en temps réel aux actions de l'utilisateur. Il est souvent combiné avec HTML et CSS pour créer des sites web complets. JavaScript est largement utilisé dans le développement web moderne et est essentiel pour créer des sites web dynamiques et interactifs [5].



iv. Python :

Python est un langage de programmation informatique. Son usage n'est pas limité au développement web. Il peut être utilisé pour tout type de programmation et de développement logiciel. On s'en sert notamment pour le développement back end d'applications web ou mobile, et pour le développement de logiciels et d'applications pour PC. Il permet également d'écrire des scripts système, afin de créer des instructions pour un système informatique. Par ailleurs, Python est le langage informatique le plus populaire pour le traitement Big Data, l'exécution de calculs mathématiques ou le Machine Learning. De manière générale, il s'agit du langage de prédilection pour la Data Science[6].



v. Visual studio code:

Visual Studio Code (VS Code) est un éditeur de code source léger, gratuit et open-source, développé par Microsoft. Il offre une interface conviviale et de nombreuses fonctionnalités pour le développement de logiciels, y compris la coloration syntaxique, l'auto-complétions, le débogage, le contrôle de version et la gestion des extensions. VS Code est apprécié pour sa polyvalence, ses performances élevées et sa large communauté de développeurs qui créent des extensions pour étendre ses fonctionnalités de base. C'est un outil populaire et largement utilisé dans le domaine du développement de logiciels [7].



vi. Plotly:

La bibliothèque Python de Plotly est une bibliothèque open-source interactive. Cela peut être un outil très utile pour la visualisation des données et la compréhension des données simplement et facilement. Les objets Plotly Graph sont une interface de haut niveau pour plotly qui est facile à utiliser. Il peut tracer différents types de graphiques et de diagrammes tels que des nuages de points, des graphiques linéaires, des graphiques à barres, des boîtes à moustaches, des histogrammes, des diagrammes circulaires[8].



vii. Django:

Django est un puissant framework web Python conçu pour un développement rapide avec son approche propre et pratique. Il adopte le modèle MVT (Model-View-Template), une variante du modèle MVC (Model-View-Controller) bien connu. Django se concentre sur la simplification de la création de sites Web sophistiqués basés sur des bases de données en promouvant la réutilisation et la flexibilité des composants. Il fournit des fonctionnalités intégrées pour l'authentification des utilisateurs, la gestion des URL, la création de modèles et les mises à jour du schéma de base de données, ce qui en fait un choix populaire parmi les développeurs Web qui cherchent à créer des applications en ligne dynamiques[9].



viii. Scikit-Learn

C'est une librairie Pythonne qui donne accès à des versions efficaces d'un grand nombre d'algorithmes courants. Elle offre également une API propre et uniformisée. Par conséquent, un des gros avantages de Scikit-Learn est qu'une fois que vous avez compris l'utilisation et la syntaxe de base de Scikit-Learn pour un type de modèle, **le passage à un nouveau modèle ou algorithme est très simple**. La librairie ne permet pas seulement de faire de la modélisation, elle peut assurer également des étapes de preprocessing[10].



V. Conclusion

-Dans ce chapitre, nous avons étudié certains outils que nous utilisons lors de la réalisation de ce projet. Puis nous passons à la partie méthodologie, Qu'est-ce que ces méthodes ?

Chapitre 3 : Méthodologie Proposé

I. Introduction :

Dans ce chapitre, on va présenter et notre méthodologie d'étude de ce projet, aussi nous présentons le fonctionnement détaillé de notre modèle de recommandation des livres tel que le fonctionnement mathématique, et la base d'architecture de ce modèle et en explique toutes les fonctionnalités nécessaires pour réaliser ce dernier.

II. Méthodologies proposées :

i. Le modèle de recommandation des livres :

a) Motivation :

La surcharge d'informations. Dans le domaine de la littérature, où des milliers de nouveaux livres sont publiés chaque année, il devient de plus en plus difficile pour les lecteurs de découvrir des œuvres pertinentes et enrichissantes. C'est dans ce contexte que les systèmes de recommandation de livres jouent un rôle crucial, donc le système de recommandation de livres vise à résoudre ce problème en offrant aux utilisateurs une expérience de découverte personnalisée et efficace. Donc l'idée de notre système c'est analyser votre donnée que nous donne dans la partie inscription, un formulaire qui collecte votre information tel que le nom, prénom et email et plus important pour nous c'est d'exprimer votre intérêt concernant les livres et exprimer le genre des livres de votre intérêt dans la lecture, on récupère le genre et un titre similaire pour ce genre de votre intérêt par exemple le titre d'un livre.

ii. Fonctionnement Mathématiques :

Avant de plonger dans l'algorithme Word2Vec, nous devons d'abord comprendre ce qu'est un « *Embedding* » de mots et pourquoi il est important.

Un « *embedding* » de mots est une manière de convertir un morceau de texte en un format numérique que nos machines peuvent lire.

Avant l'apparition de Word2Vec, plusieurs algorithmes étaient utilisés pour représenter les textes sous forme de nombres. L'un des algorithmes les plus courants est « *Count Vectorizer* », qui nous fournit une représentation localiste du texte. Prenons l'exemple suivant :

Supposons que nous ayons 5 mots dans notre vocabulaire et que nous voulions représenter "Happy" et "Cheerful" avec un vecteur. Voici le résultat du Count Vectorizer :

Happy	0	0	1	0	0
Cheerful	0	1	0	0	0

Dans le tableau ci-dessus, nous aurons 1 si un mot est présent dans le texte, sinon la valeur sera zéro. Nous pouvons également le faire pour l'ensemble de la phrase en entrant le compte du mot dans ce document particulier et en gardant le reste à zéro. Mais il y a deux problèmes avec cette méthode

- Bien que "heureux" et "joyeux" soient similaires, il n'y a pas de notion de similarité entre ces deux vecteurs car ce sont des vecteurs orthogonaux.
- Lorsque nous avons un corpus important, ces vecteurs sont très dispersés et deviennent très inefficaces lorsque nous avons des millions de mots dans notre vocabulaire. Il existe des algorithmes améliorés comme TF-IDF, mais ils rencontrent toujours les deux problèmes ci-dessus.

Ces lacunes ont été résolues par l'algorithme Word2Vec. Il repose sur l'idée de la sémantique distributionnelle, ce qui signifie que nous pouvons comprendre le sens d'un mot en comprenant les mots qui l'entourent (le contexte).

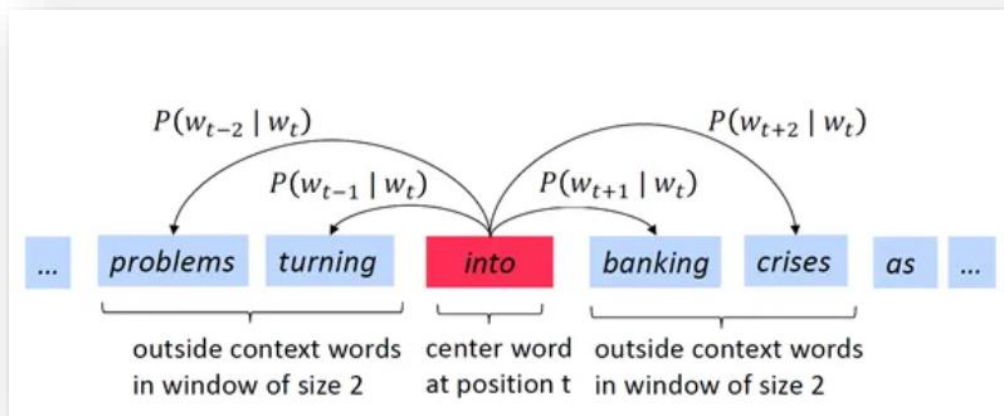
Ainsi, pour apprendre la représentation d'un mot, nous allons essayer d'apprendre le contexte dans lequel un mot particulier apparaît et espérer que le vecteur de ce mot sera similaire au vecteur des mots de contexte. Ce sera une représentation distribuée car maintenant, le sens du mot est réparti sur toutes les dimensions. Cela résoudra les deux problèmes que nous avons discutés ci-dessus[11].

Maintenant, la question est comment pouvons-nous faire cela ?

➤ **Algorithme Skip-gram :**

L'idée principale derrière l'algorithme est que tout d'abord, nous initialiserons de manière aléatoire le vecteur pour chaque mot du vocabulaire.

Ensuite, nous passerons par chaque position t et nous définirons le mot central à cette position comme c et son mot de contexte comme o . Pour identifier les mots de contexte, nous définirons une fenêtre de taille m , ce qui signifie que notre modèle regardera les mots en position $t-m$ à $t+m$ comme le contexte.



Une fois que nous avons tous les mots de contexte à la position t , nous essayerons de maximiser la vraisemblance des mots de contexte donnés le mot central, c'est-à-dire que nous calculerons la probabilité que notre modèle prédise les mots de contexte donnés le mot central et nous essayerons de maximiser cette probabilité. Cette vraisemblance peut être représentée par cette formule :

$$L(\theta) = \prod_{t=1}^T \prod_{\substack{-m \leq j \leq m \\ j \neq 0}} P(w_{t+j} | w_t; \theta)$$

Pour mettre cette équation sous une forme telle qu'elle soit facile à dériver et à en faire un problème de minimisation, nous prendrons simplement le log de l'équation et la multiplierons par $* 1$ pour calculer le *-ve log-likelihood*.

Maintenant, la multiplication se transformera en une sommation car nous avons pris le log.

$$J(\theta) = -\frac{1}{T} \log L(\theta) = -\frac{1}{T} \sum_{t=1}^T \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \log P(w_{t+j} | w_t; \theta)$$

- La question est comment pouvons-nous calculer la probabilité du mot de contexte étant donné le mot central ?

Pour ce faire, nous représenterons chaque mot par deux ensembles de vecteurs, Uw et Vw . Nous utiliserons Uw lorsque w est le mot de contexte et Vw lorsque w est le mot central. En utilisant ces deux vecteurs, notre équation de probabilité pour le mot central o et le mot de contexte c ressemblera à ceci :

$$P(O = o | C = c) = \frac{\exp(\mathbf{u}_o^\top \mathbf{v}_c)}{\sum_{w \in \text{Vocab}} \exp(\mathbf{u}_w^\top \mathbf{v}_c)}$$

Comme nous pouvons le voir, il s'agit d'un exemple de la fonction *softmax*. Dans le numérateur, nous avons le produit scalaire du mot o et c qui capture la similarité entre ces deux vecteurs. Plus la similarité est grande, plus la probabilité est grande. Le dénominateur nous donne un moyen de normaliser les valeurs de probabilité sur l'ensemble du vocabulaire afin qu'elles s'additionnent toutes à 1.

Maintenant que nous avons notre fonction de perte en place, vient maintenant la partie de l'entraînement du modèle. Si nous regardons notre équation de vraisemblance ci-dessus, nous pouvons voir que nous n'avons qu'un seul paramètre de modèle dans notre équation θ . C'est un long vecteur qui contient à la fois les vecteurs v et u de longueur d pour tous les mots.

$$\Theta = \begin{bmatrix} \mathbf{V}_{w1} \\ \mathbf{V}_{w2} \\ \vdots \\ \mathbf{V}_{wn} \\ \mathbf{U}_{w1} \\ \mathbf{U}_{w2} \\ \vdots \\ \mathbf{U}_{wn} \end{bmatrix} \in \mathbb{R}^{2dV}$$

Nous utiliserons l'algorithme de descente de gradient pour descendre la colline et changer tous les poids lentement pour maximiser notre vraisemblance. Nous obtiendrons la direction dans laquelle nous devons nous déplacer pour changer les poids en prenant la dérivée de notre fonction de perte par rapport à \mathbf{U} et \mathbf{V} .

Commençons par la dérivée de $\mathbf{J}(\theta)$ par rapport à \mathbf{V}_c :

Puisque $\mathbf{J}(\theta)$ est un rapport, en prenant un log, le dénominateur sera au-dessus avec un signe négatif. Ainsi, nous pouvons représenter notre dérivée comme ceci :

$$\frac{\partial \mathbf{J}(\theta)}{\partial v_c} = \frac{\partial}{\partial v_c} (\log (\exp(u_o^T v_c))) - \frac{\partial}{\partial v_c} (\log \sum_{w=1}^v \exp (u_w^T v_c))$$

Divisons notre équation en deux parties et résolvons-les individuellement.

Pour la première partie, puisque $\log(\exp(x))$ est égal à x , nous pouvons l'écrire comme ceci :

$$\frac{\partial}{\partial v_c} (\log (\exp(u_o^T v_c))) = \frac{\partial}{\partial v_c} (u_o^T v_c) = u_o$$

Si ce n'est pas clair comment cela devient \mathbf{U}_o , essayes de prendre la dérivée élément par élément. Si nous prenons la dérivée par rapport à $\mathbf{V}_c \mathbf{I}$, le résultat sera $\mathbf{U}_o \mathbf{I}$ car seul ce terme aura V_{c1} . En le faisant pour tous les éléments, nous obtiendrons un vecteur de $\mathbf{U}_o \mathbf{I}$ à $\mathbf{U}_o n$ qui est équivalent à \mathbf{U}_o .

Passons maintenant à la deuxième partie de l'équation. En prenant la dérivée de $\log(x)$ et en déplaçant la dérivée à l'intérieur de la sommation, il ne reste que ce terme :

$$\frac{1}{\sum_{w=1}^v \exp(u_w^T v_c)} \sum_{x=1}^v \frac{\partial}{\partial v_c} \exp(u_x^T v_c)$$

Prendre la dérivée du terme $\exp(x)$ et réarranger le signe de sommation nous donnera ce terme :

$$\sum_{x=1}^v \frac{\exp(u_x^T v_c)}{\sum_{w=1}^v \exp(u_w^T v_c)} * u_x$$

Si nous examinons attentivement le terme ci-dessus, nous pouvons voir que le terme à l'intérieur de la sommation est le même que celui du terme de probabilité que nous avons défini ci-dessus et peut donc être écrit comme ceci :

$$\sum_{x=1}^v P(x|c) * u_x$$

En combinant le tout, nous pouvons écrire :

$$\frac{\partial J(\theta)}{\partial v_c} = -u_o + \sum_{x=1}^v P(x|c) * u_x$$

Il y a quelque chose de très intéressant dans cette équation.

- La première partie est la représentation actuelle du mot de contexte.
- La deuxième partie est une attente de ce à quoi devrait ressembler le mot de contexte selon notre modèle car nous prenons la représentation actuelle de tous les mots de contexte et les multiplions par leur probabilité dans le modèle actuel.

Donc, fondamentalement, nous soustrayons les représentations réelles et attendues pour obtenir la direction dans laquelle je devrais me déplacer et changer mon vecteur de poids V_c afin de maximiser la vraisemblance.

En avançant de la même manière, nous pouvons également calculer la dérivée de $J(\theta)$ par rapport à U_w . Il y aura deux cas pour U_w , un lorsque w est le mot de contexte et un lorsque w n'est pas le mot de contexte. Ce seront les sorties des dérivées dans les deux cas :

➤ $w \neq o$:

$$\frac{\partial J(\theta)}{\partial u_w} = \sum_{x=1}^v P(x|c) * v_c$$

➤ $w = o$:

$$\frac{\partial J(\theta)}{\partial u_w} = -v_c + \sum_{x=1}^v P(x|c) * v_c$$

Une fois que nous avons les deux dérivées, nous pouvons les utiliser dans notre équation de SGD pour mettre à jour les poids.

Cependant, il y a un problème dans cette approche. Comme nous pouvons le voir, dans le dénominateur, nous devons prendre l'exponentielle du produit scalaire de tous nos mots et cela prend beaucoup de temps lorsque nous avons un vocabulaire énorme. Nous devons entraîner des millions de poids ce qui n'est pas réalisable.

Ainsi, pour augmenter le temps d'entraînement, une nouvelle méthode a été utilisée appelée échantillonnage négatif. Dans cette méthode, nous mettrons à jour uniquement un petit pourcentage de poids en une étape. Nous sélectionnerons quelques mots *-ve*, c'est-à-dire les mots qui ne sont pas dans la fenêtre de contexte et nous changerons nos poids de telle sorte qu'ils maximisent la probabilité des vrais mots de contexte et minimisent la probabilité des mots

aléatoires apparaissant autour du mot central. Cela change la fonction de perte et maintenant nous essayons de maximiser l'équation suivante :

$$J_{neg-sample}(\mathbf{o}, \mathbf{v}_c, \mathbf{U}) = -\log(\sigma(\mathbf{u}_o^T \mathbf{v}_c)) - \sum_{k=1}^K \log(\sigma(-\mathbf{u}_k^T \mathbf{v}_c))$$

Où K est le nombre d'échantillons négatifs.

Pour sélectionner des mots aléatoires, nous utilisons une distribution unigramme où les mots les plus fréquents sont plus susceptibles d'être sélectionnés.

Pour maximiser le terme ci-dessus, nous devons à nouveau prendre la dérivée de la fonction de perte par rapport aux poids, dans ce cas, ce seront $\mathbf{U}\mathbf{w}$, $\mathbf{U}\mathbf{k}$ et $\mathbf{V}\mathbf{c}$. En faisant cela de la même manière que nous l'avons fait ci-dessus, nous obtiendrons les trois équations suivantes :

$$\frac{\partial J(\theta)}{\partial v_c} = -\sigma(-\mathbf{u}_o^T \mathbf{v}_c) \mathbf{u}_o + \sum_{k=1}^K \sigma(\mathbf{u}_k^T \mathbf{v}_c) \mathbf{u}_k$$

$$\frac{\partial J(\theta)}{\partial u_o} = -\sigma(-\mathbf{u}_o^T \mathbf{v}_c) \mathbf{v}_c$$

$$\frac{\partial J(\theta)}{\partial u_k} = \sum_{k=1}^K \sigma(\mathbf{u}_k^T \mathbf{v}_c) \mathbf{v}_c$$

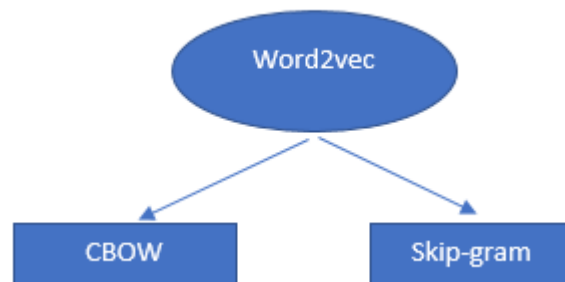
Avec toutes les dérivées calculées, maintenant nous pouvons mettre à jour nos vecteurs de poids petit à petit et obtenir une représentation vectorielle qui fera pointer les mots apparaissant dans un contexte similaire dans la même direction.

iii. Architecture :

Word2Vec est un modèle d'apprentissage automatique largement utilisé pour apprendre des représentations vectorielles de mots à partir des grands corpus de texte non étiqueté. Cette technique a été développée par une équipe de recherche de Google dirigée par Tomas Mikoloven 2013[12].

L'idée centrale derrière Word2Vec est de représenter chaque mot dans un espace vectoriel continu, où les mots ayant des significations similaires sont placés à proximité les uns des autres. Word2Vec utilise un réseau de neurones artificiels pour apprendre ces représentations de mots à partir de données textuelles.

Word2Vec utilise principalement deux algorithmes différents pour apprendre des représentations vectorielles de mots : *Continuous Bag of Words* (CBOW) et Skip-Gram. Ces deux algorithmes sont des variantes de réseaux de neurones peu profonds (shallow neural networks).



a) Continuous Bag of Words (CBOW):

Le modèle prédit le mot central à partir d'un contexte donné, c'est-à-dire un ensemble de mots environnants. Ainsi, CBOW estime la probabilité d'un mot donné en fonction de son contexte. L'architecture du réseau de neurones consiste en une couche d'entrée, une couche cachée et une couche de sortie.

Les poids de la couche d'entrée vers la couche cachée représentent les embeddings des mots, qui sont appris pendant l'entraînement.

Architecture de CBOW:

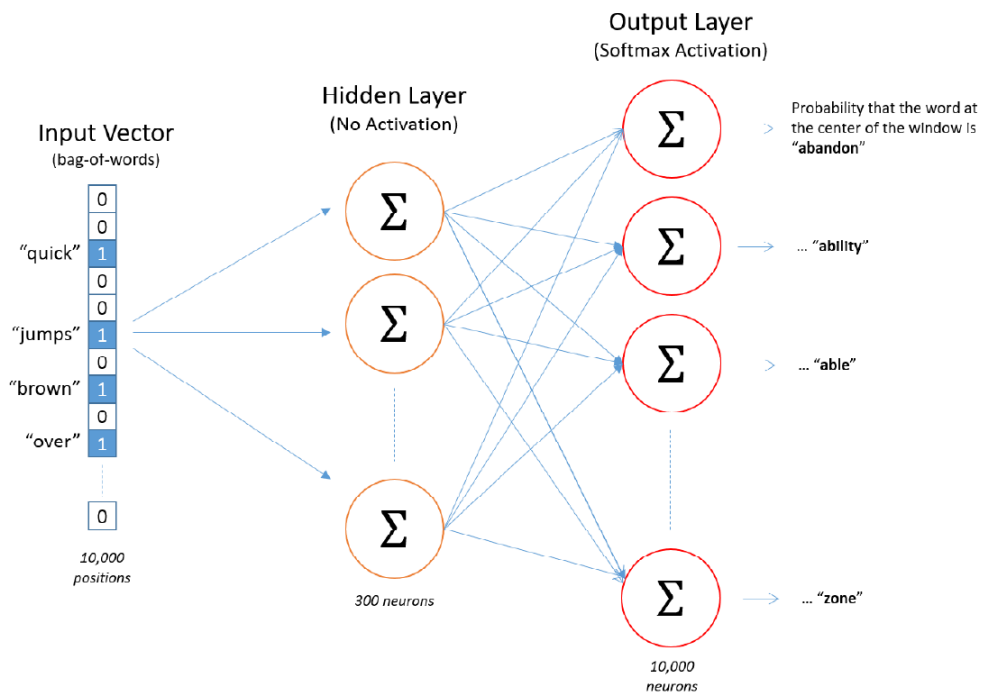


Figure 4 : Architecture de CBOW

b) Skip-Gram :

Skip-Gram est un modèle prédit les mots environnants à partir d'un mot central donné. Cela signifie que le modèle estime la probabilité des mots contextuels donnés le mot central. L'architecture du réseau est similaire à celle de CBOW, mais la couche d'entrée et la couche de sortie sont inversées. Les incorporation (*embedding*) de mots sont appris de la même manière que dans CBOW, mais dans ce cas, les vecteurs de mots sont utilisés pour prédire les mots environnants.

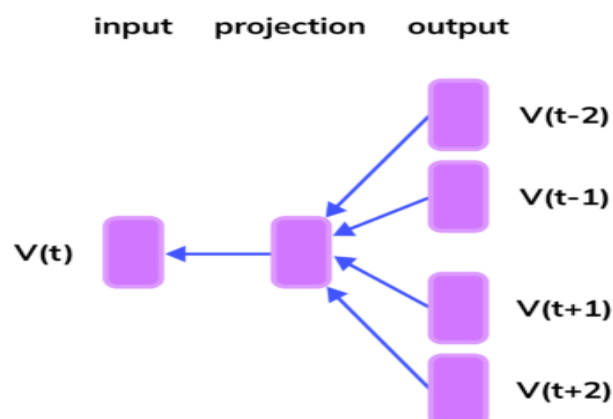


Figure 5 : Architecture de Skip-Gram

ii. Statistiques sur la violence contre les femmes :

1. Introduction :

-La visualisation des données joue un rôle crucial dans la compréhension et la communication des informations complexes. Dans le contexte de la violence contre les femmes, il est essentiel de comprendre les tendances et les facteurs qui contribuent à ce phénomène. Les graphiques interactifs basés sur *Plotly* offrent une approche dynamique pour explorer ces données, en permettant aux utilisateurs de visualiser les niveaux de violence en fonction de différentes variables démographiques telles que l'âge, l'éducation, l'emploi, etc. Cette visualisation interactive fournit un moyen efficace d'analyser et de présenter les données, aidant ainsi à sensibiliser les efforts pour lutter contre la violence à l'égard des femmes[13].

2. Analyse des données

-La violence contre les femmes et les filles est un problème mondial qui nécessite une compréhension approfondie pour être combattu efficacement. Dans notre ensemble de données, nous avons collecté des informations sur la violence contre les femmes et les filles dans 70 pays. Nous examinons les données démographiques pour comprendre qui est le plus touché par la violence, ainsi que les réponses aux questions sur les raisons de la violence pour obtenir des informations plus détaillées.

Voici un aperçu des différentes questions démographiques, des raisons de la violence et des statistiques sur le nombre maximum et moyen de cas de violence enregistrés :

Questions démographiques :

- Âge :
 - Les groupes d'âge étudiés sont les suivants : 15-24 ans, 25-34 ans, 35-49 ans.
- Éducation :
 - Pas d'éducation, Primaire, Secondaire, Supérieur.
- Emploi :
 - Sans emploi, Employé pour de l'argent, Employé pour des biens.
- Statut matrimonial :
 - Marié ou vivant ensemble, Veuf, Divorcé, Séparé, Jamais marié.
- Résidence :
 - Rural, Urbain.

Raisons de la violence :

-Les raisons pour lesquelles la violence peut être infligée aux femmes et aux filles comprennent des facteurs tels que :

- Brûler la nourriture ;
- Refuser d'avoir des relations sexuelles ;
- Se disputer avec le partenaire ;
- Sortir sans en informer le partenaire ;
- Négliger les enfants.

Statistiques sur la violence :

- Nombre maximum de cas de violence : Le nombre maximal de cas de violence enregistrés pour chaque groupe démographique ou chaque raison de violence.
- Nombre moyen de cas de violence : Le nombre moyen de cas de violence enregistrés pour chaque groupe démographique ou chaque raison de violence.

En analysant ces données, nous pouvons obtenir des informations précieuses sur les tendances et les corrélations entre les différentes variables. Par exemple, nous pourrions constater que les femmes âgées de 15 à 24 ans et sans éducation sont plus susceptibles d'être victimes de violence. De même, nous pourrions identifier des motifs de violence récurrents, tels que le refus d'avoir des relations sexuelles ou les disputes avec le partenaire.

3. Visualisation des Racines de la Violence

a) Contexte

Cette visualisation explore les raisons profondes de la violence contre les femmes, en se basant sur les valeurs médianes et maximales. Elle met en lumière des motifs surprenants tels que la brûlure des aliments, le refus de relations sexuelles, les disputes, le fait de sortir sans permission et la négligence des enfants, qui émergent comme des déclencheurs clés de la violence.

b) Analyse

Les raisons courantes de la violence affichent des valeurs médianes élevées et des pourcentages absolus maximum, soulignant l'urgence d'adresser ces problèmes sociétaux profondément enracinés. Ces résultats soulignent le besoin critique d'interventions ciblées et de systèmes de soutien pour lutter efficacement contre la violence à l'égard des femmes.

c) Visualisation :

Données Utilisées : Les données de violence contre les femmes sont regroupées par catégorie démographique, notamment l'âge, l'éducation, l'emploi, le statut matrimonial et la résidence.

Méthode de Visualisation : Les données sont agrégées et présentées sous forme de graphiques polar (radar) pour chaque catégorie démographique. Chaque graphique représente soit la médiane, soit le maximum de la violence pour une réponse démographique spécifique

Fonctionnement : L'utilisateur peut sélectionner une catégorie démographique (âge, éducation, etc.) être un type de statistique (médiane ou maximum). En fonction de ces sélections, un graphique Plotly correspondant est chargé et affiché sur la page. Cette approche interactive offre une exploration des données de violence contre les femmes, permettant une comparaison visuelle des niveaux de violence selon différentes variables démographiques.

Cette combinaison de visualisations permet une compréhension approfondie des causes et des manifestations de la violence contre les femmes, soulignant l'importance cruciale de l'action pour mettre fin à ce problème mondial.

4. Visualisation à travers des Cartes

a) Contexte :

La visualisation de la violence contre les femmes à travers des cartes choroplèthes interactives offre une approche détaillée et engageante pour comprendre la répartition de la violence à l'échelle mondiale. En se basant sur des données regroupées par pays et par indicateur démographique tel que le niveau d'éducation, l'âge, etc., ces cartes permettent d'identifier les tendances et les disparités régionales de la violence contre les femmes. Ce contexte est essentiel pour sensibiliser et mobiliser les efforts de lutte contre la violence à l'égard des femmes.

b) L'analyse :

Les cartes choroplèthes offrent une représentation visuelle claire et détaillée de la médiane de la violence contre les femmes pour chaque indicateur démographique dans chaque pays. Ces cartes permettent de mettre en évidence les pays où la violence est plus prévalente pour certains groupes démographiques, ce qui peut aider à orienter les politiques et les interventions de manière plus ciblée. L'analyse de ces cartes permet de mieux comprendre la complexité de la violence contre les femmes et d'identifier les zones où des mesures de prévention et de soutien sont les plus nécessaires.

c) *Visualisation*

Données Utilisées : Les données utilisées pour cette visualisation proviennent de différentes sources telles que les rapports nationaux sur la violence contre les femmes, les enquêtes démographiques et de santé, ainsi que les données recueillies par des organisations internationales et des ONG. Ces données sont regroupées par pays et par indicateur démographique, tels que le niveau d'éducation, l'âge, le statut matrimonial, etc. Par exemple, pour l'indicateur du niveau d'éducation, les données peuvent être divisées en catégories telles que "pas d'éducation", "éducation primaire", "éducation secondaire", "éducation supérieure"...

Méthode de Visualisation : Les données sont agrégées et analysées pour calculer la médiane de la violence pour chaque indicateur démographique dans chaque pays. Ensuite, ces valeurs sont représentées graphiquement sur des cartes choroplèthes. Dans une carte choroplèthe, chaque pays est coloré en fonction de la valeur de la médiane de la violence pour l'indicateur démographique sélectionné. Les couleurs sont généralement choisies en fonction d'une échelle de couleur prédéfinie pour représenter différentes gammes de valeurs.

Fonctionnement : Lorsque l'utilisateur accède à la visualisation, il est invité à sélectionner un indicateur démographique à partir d'une liste déroulante, tels que le niveau d'éducation ou l'âge. En fonction de cette sélection, une carte choroplèthe correspondante est chargée et affichée sur la page. L'utilisateur peut ainsi explorer visuellement les données de violence contre les femmes pour différents indicateurs démographiques et pays. Cette approche interactive permet une compréhension plus approfondie des tendances et des disparités régionales de la violence contre les femmes, ce qui peut aider à orienter les politiques et les interventions de manière plus efficace.

III. Conclusion

Ce chapitre explique comment le système de recommandation de livres basé sur Word2Vec a été créé, ainsi que comment les données sur la violence contre les femmes ont été utilisées. Le chapitre suivant expliquera comment mettre en pratique ce système, en fournissant des instructions étape par étape pour le faire.

Chapitre 4 : Expérimentation Analyse de résultats

I. Introduction

Ce chapitre se concentre sur l'expérimentation et l'analyse des résultats de notre système de recommandation des livres. Nous commencerons par présenter l'ensemble de données utilisé et la manière dont nous l'avons collecté. Ensuite, nous décrirons en détail notre approche expérimentale, en mettant en évidence les différents paramètres et métriques que nous avons pris en compte. Enfin, nous discuterons de l'implémentation de notre plateforme.

II. Le système de recommandation

i. Introduction

Le système de recommandation est un élément clé de notre plateforme dédiée à la lutte contre les abus et les violences envers les femmes. Pour recommander des livres pertinents aux utilisateurs, nous avons choisi d'utiliser l'algorithme Word2Vec. Cette approche, basée sur le traitement du langage naturel, nous permet de comprendre les relations sémantiques entre les mots et d'identifier les livres qui correspondent le mieux aux intérêts et aux besoins des utilisateurs

ii. Dataset et la collection des données

Dans notre modèle de recommandation on travaille sur la trois bases données collectés depuis la grande siteweb des données kaggle :

Les trois dataset collecter sont :

- Dataset des livres : Books.csv
- Datast des ratings : Ratings.csv
- Dataset de utilisateur : Users.csv

a. Dataframe des livres

Notre dataframe des livres contient 271360 ligne et 8 colonnes, ces 8 colonnes sont :

ISBN : ce colonne contient id de chaque livre

Book-Title : ce colonne contient titre de chaque livre

Book-Author: ce colonne contient le nom de auteur

Year-Of-Publication : ce colonne contient la date de publication de chaque livre

Publisher : ce colonne contient source de publication

URL-S : ce colonne contient le lien de l'image de couverture de notre livre en petite taille

URL-M : ce colonne contient le lien de l'image de couverture de notre livre en grande taille

	ISBN	Book-Title	Book-Author	Year-Of-Publication	Publisher	Image-URL-S	Image-URL-M
0	0195153448	Classical Mythology	Mark P. O. Morford	2002	Oxford University Press	http://images.amazon.com/images/P/0195153448.0...	http://images.amazon.com/images/P/0195153448.0...
1	0002005018	Clara Callan	Richard Bruce Wright	2001	HarperFlamingo Canada	http://images.amazon.com/images/P/0002005018.0...	http://images.amazon.com/images/P/0002005018.0...
2	0060973129	Decision in Normandy	Carlo D'Este	1991	HarperPerennial	http://images.amazon.com/images/P/0060973129.0...	http://images.amazon.com/images/P/0060973129.0...
3	0374157065	Flu: The Story of the Great Influenza Pandemic...	Gina Bari Kolata	1999	Farrar Straus Giroux	http://images.amazon.com/images/P/0374157065.0...	http://images.amazon.com/images/P/0374157065.0...
4	0393045218	The Mummies of Urumchi	E. J. W. Barber	1999	W. W. Norton & Company	http://images.amazon.com/images/P/0393045218.0...	http://images.amazon.com/images/P/0393045218.0...

b. Dataframe des notations

Notre dataframe des notations contient 1149780 lignes et 3 colonnes, ces 3 colonnes sont :

User-ID : ce colonne contient id de chaque utilisateur

ISBN : ce colonne contient id de chaque livre

Book-Rating : ce colonne contient notation de chaque utilisateur

	User-ID	ISBN	Book-Rating
0	276725	034545104X	0
1	276726	0155061224	5
2	276727	0446520802	0
3	276729	052165615X	3
4	276729	0521795028	6

c. Dataframe des utilisateurs

Notre dataframe des utilisateur contient 278858 lignes et 3 colonnes, ces 3 colonnes sont :

User-ID : ce colonne contient id de chaque utilisateur

Location : ce colonne contient la ville de chaque utilisateur

Age : ce colonne contient l'âge de chaque utilisateur

	User-ID	Location	Age
0	1	nyc, new york, usa	NaN
1	2	stockton, california, usa	18.0
2	3	moscow, yukon territory, russia	NaN
3	4	porto, v.n.gaia, portugal	17.0
4	5	farnborough, hants, united kingdom	NaN

Après cette partie de définir chaque dataset on combiner tous ces trois dataset dans une seule dataset on basons en premier de combiner la dataset des livre et des notation on basons sur la colonne ISBN, après en combiner la nouveaux dataset avec la dataset des utilisateur on basons sur la colonnes User-Id pour obtenir notre dernière base de donnée nommée final_df.

iii. Partie d'analyse des données

Dans cette partie en utilise les déférents méthode de la bibliothèque numpy et pandas et matplotlib pour visualiser et voir les données de notre dataset les livres basons sur les notations des utilisateur ou bien la date de publication :

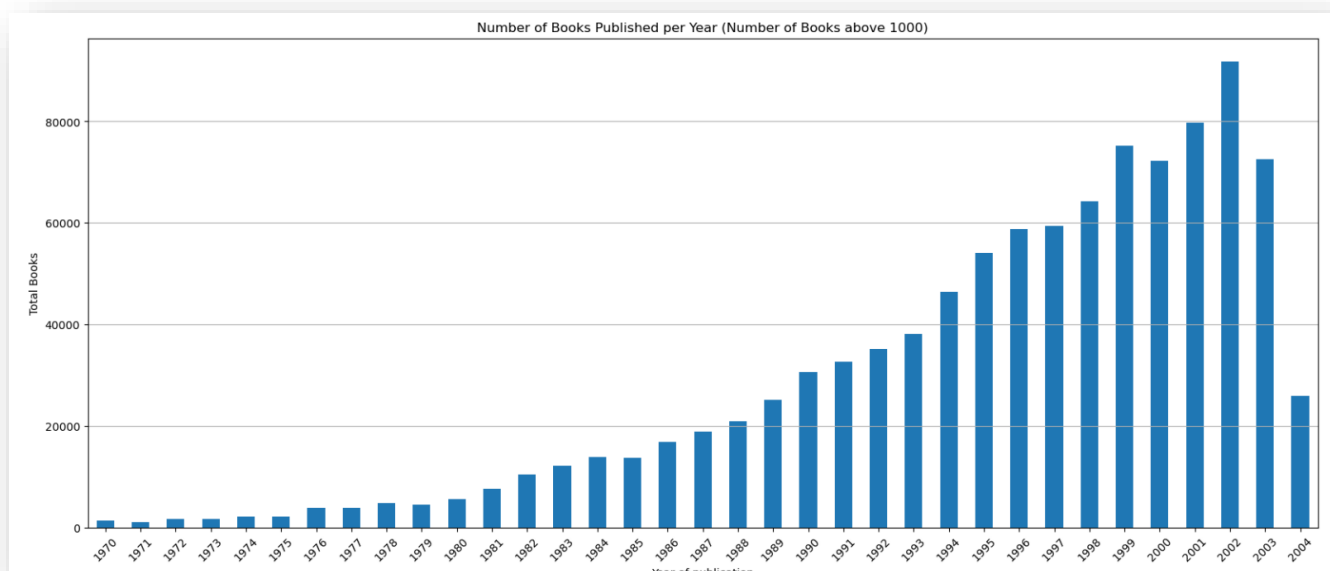


Figure 6 : Affichage des données basons sur les notations

iv. Partie de prétraitement des données :

Dans cette étapes de prétraitement visent à nettoyer et à préparer les données pour une analyse ou un modèle ultérieur, plusieurs opérations sont effectuées pour prétraiter les données contenues dans **final_df** (notre DataFrame) avant de les utiliser pour une analyse ultérieure ou pour construire notre modèle. Voici les étapes de ce qui on fait :

1. Gestion des valeurs manquantes :

- Les valeurs manquantes dans les colonnes 'Book-Author' et 'Publisher' sont remplies avec le mode (la valeur la plus fréquente) de chaque colonne.
- Les valeurs manquantes dans la colonne 'Age' sont remplies avec la médiane de cette colonne.

2. Suppression de certaines colonnes :

- Les colonnes 'Image-URL-L' et 'Location' sont supprimées car elles ne sont pas nécessaires pour l'analyse ou le modèle.

3. Conversion de types de données :

- Les colonnes 'Age', 'Book-Rating', et 'User-ID' sont converties en entiers.

4. Gestion des doublons :

- Les lignes en double dans le DataFrame sont supprimées.

5. Filtrage des données :

- Les entrées avec un 'Book-Rating' de 0 sont supprimées, probablement car elles sont considérées comme des données incorrectes ou non valides.

6. Catégorisation des évaluations de livres :

- Les évaluations de livres sont divisées en catégories ('Low', 'Medium', 'High') en fonction de leurs valeurs.

v. Partie de prédiction des genres

En utilisons un classificateur Naive Bayes pour la prédiction du genre. Voici les étapes de ce qu'en fait :

1. **Nettoyage du texte** : Une fonction **clean_text** est définie pour nettoyer les titres de livres en enlevant les caractères spéciaux, les nombres et en les convertissant en minuscules.
2. **Prédiction de genre** : Une fonction **predict_genre** est définie pour prédire le genre d'un livre à partir de son titre. Cette fonction nettoie d'abord le titre, puis vérifie quels mots du titre sont présents dans le vocabulaire du modèle Word2Vec. Ensuite, elle calcule la représentation vectorielle moyenne des mots présents dans le titre et cherche le genre le plus similaire en utilisant la similarité cosinus entre le vecteur moyen du titre et les vecteurs de genre.
3. **Évaluation de la prédiction** : Un titre de livre est choisi pour la prédiction de genre, et la fonction **predict_genre** est utilisée pour prédire son genre.

4. **Ajout de la colonne "Genre" au dataframe :** La fonction **predict_genre** est appliquée à chaque titre de livre dans un dataframe **final_df**, et les résultats sont stockés dans une nouvelle colonne "Genre".

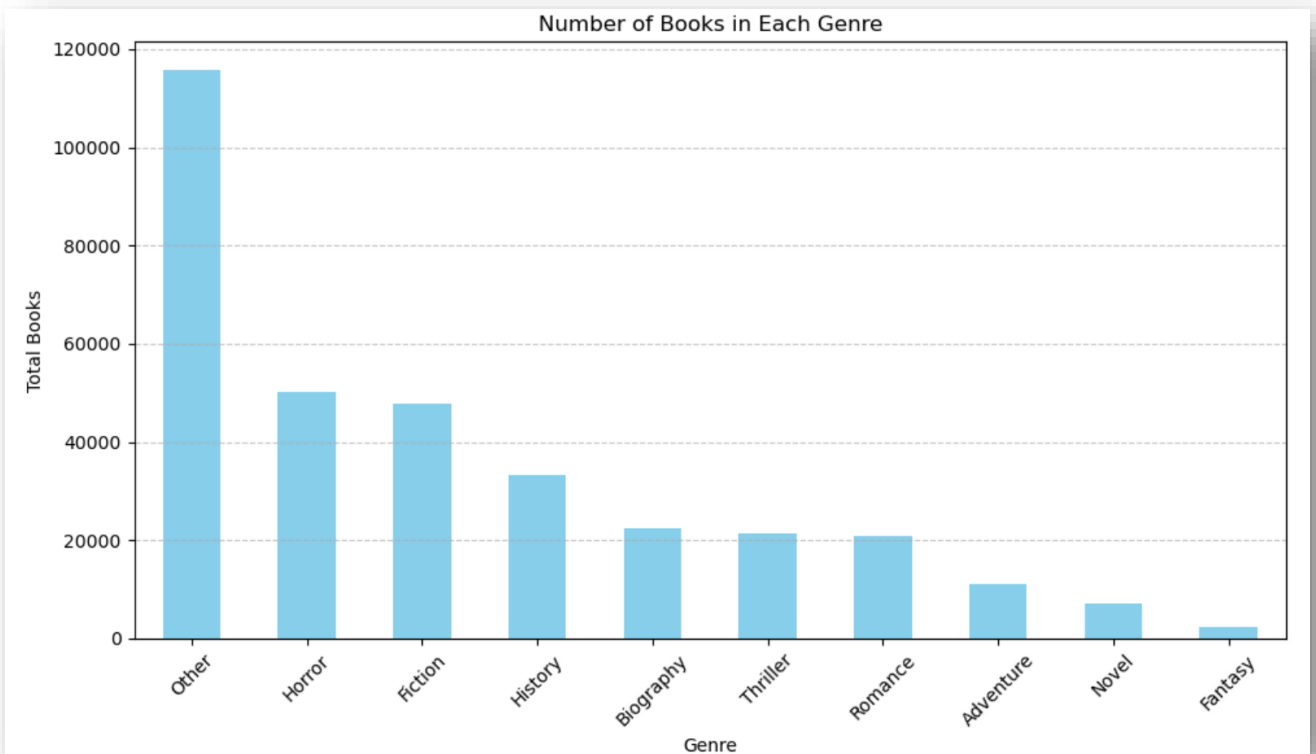
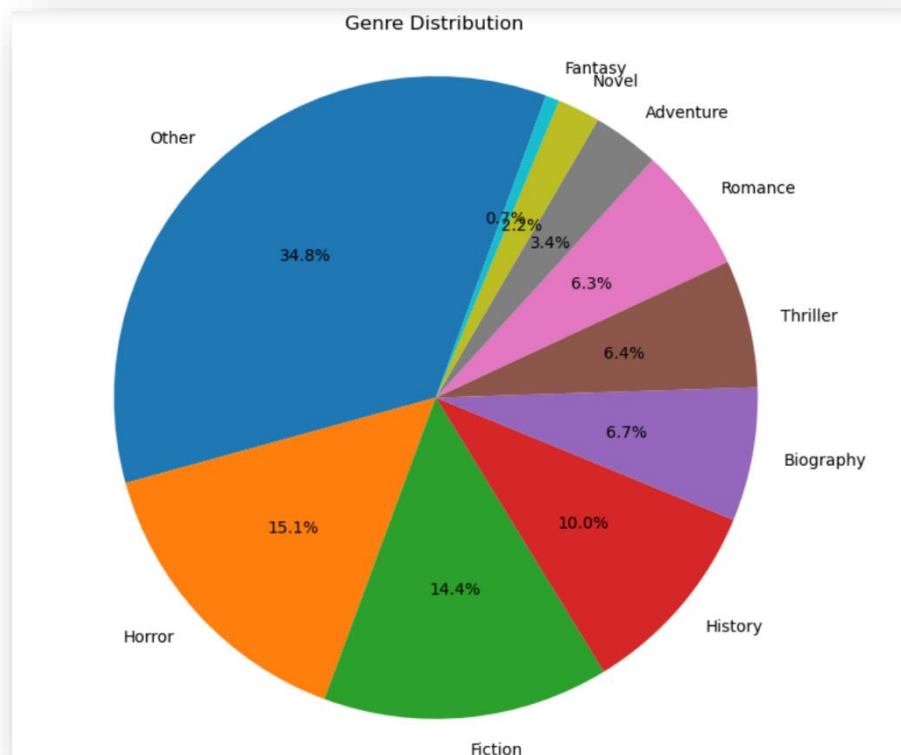


Figure 7 : Affichage des livres basons sur le genre

5. **Visualisation des genres de livres :** Un diagramme à barres est tracé pour montrer le nombre de livres dans chaque genre.
6. **Classification avec Naive Bayes :** Les données sont divisées en ensembles d'entraînement et de test, puis vectorisées en utilisant TF-IDF. Un classificateur Naive Bayes multinomial est initialisé et entraîné sur les données d'entraînement.
7. **Prédiction et évaluation de la performance du classificateur :** Le classificateur entraîné est utilisé pour prédire les genres des livres dans l'ensemble de test, et la précision du modèle est calculée pour les ensembles d'entraînement et de test.

```
Accuracy train: 0.7501543511979158
Accuracy test: 0.7183321286591977
```



vi. La création de la fonction de recommandation :

La création de la fonction **get_recommendations**, cette fonction utilise Word2Vec pour mesurer la similarité entre les mots des titres de livres et générer des recommandations de livres similaires en fonction d'un titre de livre donné et d'un genre donné:

La fonction prend quatre arguments :

- **Titre:** Le titre du livre à partir duquel vous souhaitez trouver des recommandations.
- **genre:** Le genre du livre.
- **data:** Le DataFrame contenant les données sur les livres à partir desquelles les recommandations seront générées.
- **word2vec_model:** Le modèle Word2Vec entraîné sur un corpus de texte, utilisé pour calculer les similarités entre les mots.

La fonction commence par combiner le titre du livre et le genre en une seule chaîne de texte, ensuite, elle tokenize cette chaîne de texte en une liste de mots, pour chaque mot dans la liste, la fonction vérifie si ce mot a un vecteur dans le modèle Word2Vec. Si c'est le cas, elle récupère

ce vecteur, ensuite, elle calcule le vecteur moyen pour tous les mots présents dans le modèle Word2Vec, la fonction parcourt ensuite toutes les lignes du dataset **data** et calcule la similarité cosinus entre le vecteur moyen du livre donné et les vecteurs moyens de chaque autre livre dans le Dataset, les paires (livre, similarité) sont stockées dans une liste appelée **similarities**, cette liste est triée par similarité, de la plus grande à la plus petite.

Les 10 premières recommandations uniques (les livres les plus similaires et avec une note supérieure à 5) sont extraites de cette liste triée et stockées dans un DataFrame **df_recommendations**, enfin, ce DataFrame de recommandations est renvoyé en sortie de la fonction.

Remarque : ne peut pas faire une mesure pour évaluer la performance d'un modèle de Word2Vec car Word2Vec n'est pas un modèle de classification. Word2Vec est utilisé pour générer des représentations vectorielles de mots à partir de données textuelles, et son objectif est de capturer les relations sémantiques et syntaxiques entre les mots

III. Experimentation

Dans le cadre de notre projet visant à développer un système de recommandation de livres utilisant Word2Vec et des techniques de classification, nous avons réalisé une série d'expérimentations pour évaluer les performances de nos modèles. Nous nous sommes particulièrement concentrés sur l'amélioration du classificateur Naive Bayes en utilisant une recherche aléatoire pour optimiser ses hyperparamètres.

Après avoir effectué une recherche aléatoire en trouve les meilleurs hyperparamètres pour notre classificateur Naive Bayes, nous avons obtenu un alpha optimal de 0.2158. En utilisant cet alpha et en réentraînant le classificateur, nous avons obtenu les résultats suivants :

- Accuracy train:0.7970
- Accuracy test : 0.7405

Ces résultats montrent une amélioration significative par rapport au modèle de base non optimisé, qui avait une précision de test de 0.7162. Cela démontre l'efficacité de la recherche aléatoire pour l'optimisation des hyperparamètres.

Voici les métriques pour chaque genres des livre:

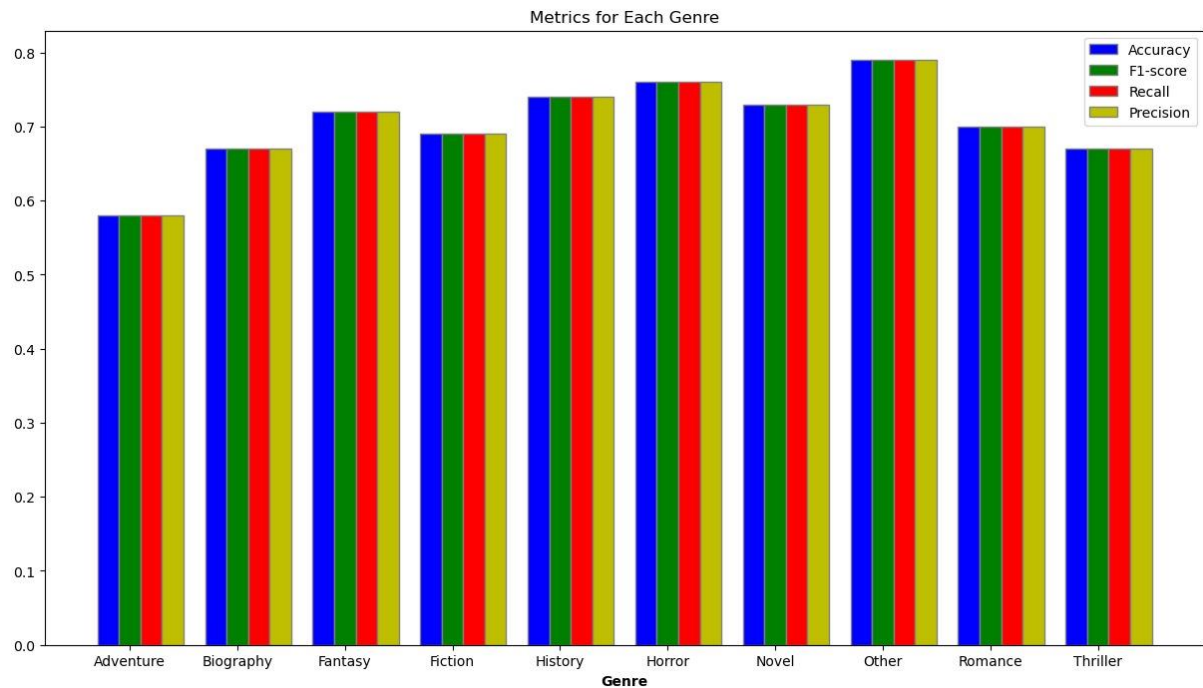


Figure 8 : Métriques pour chaque genre

Précision : La précision mesure le nombre de prédictions correctes faites par le modèle parmi toutes les prédictions positives qu'il a faites. Une précision élevée indique que le modèle a moins de faux positifs, c'est-à-dire qu'il prédit correctement les étiquettes positives.

Rappel : Le rappel mesure le nombre de prédictions correctes faites par le modèle parmi toutes les instances qui étaient réellement positives. Un rappel élevé indique que le modèle identifie efficacement les instances positives.

F1-score : Le F1-score est la moyenne harmonique de la précision et du rappel. Il donne une mesure de la performance globale du modèle en tenant compte à la fois de la précision et du rappel.

Accuracy : L'accuracy, ou exactitude, mesure la proportion d'instances correctement classées parmi toutes les instances.

Nos expérimentations ont montré que l'optimisation des hyperparamètres du classificateur Naive Bayes peut conduire à une amélioration significative des performances. De plus, l'utilisation de Word2Vec pour représenter les titres de livres semble être une approche efficace pour la classification. Ces résultats nous encouragent à continuer à explorer ces techniques pour améliorer notre système de recommandation des livres.

IV. Implémentation

i. Login page

Login page est la toute première interface affichée, deux lien affichée pour l’inscription en notre site si tu n’as pas d’un compte, et autre pour la connexion à votre compte.

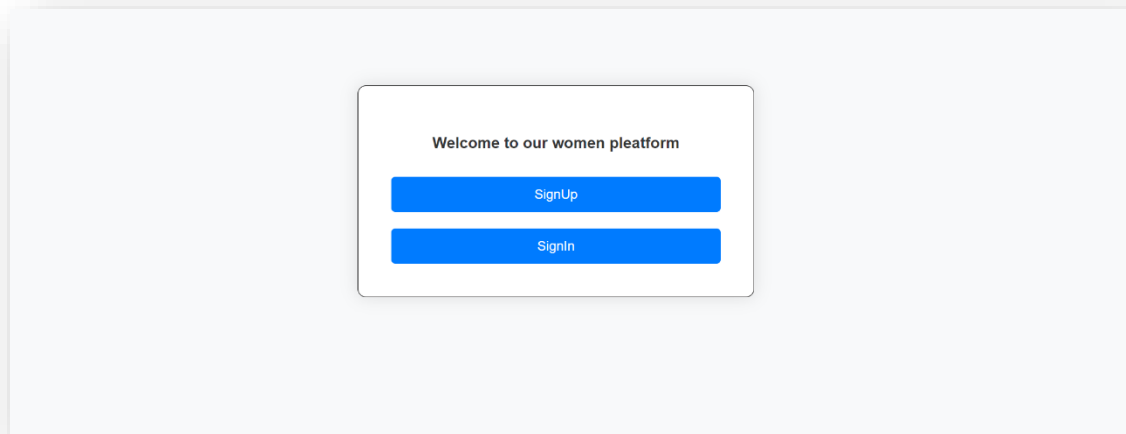


Figure 9 : Login page

ii. Page inscription

Une interface qui donner accès pour s’inscrire dans notre site web comme un utilisateur.

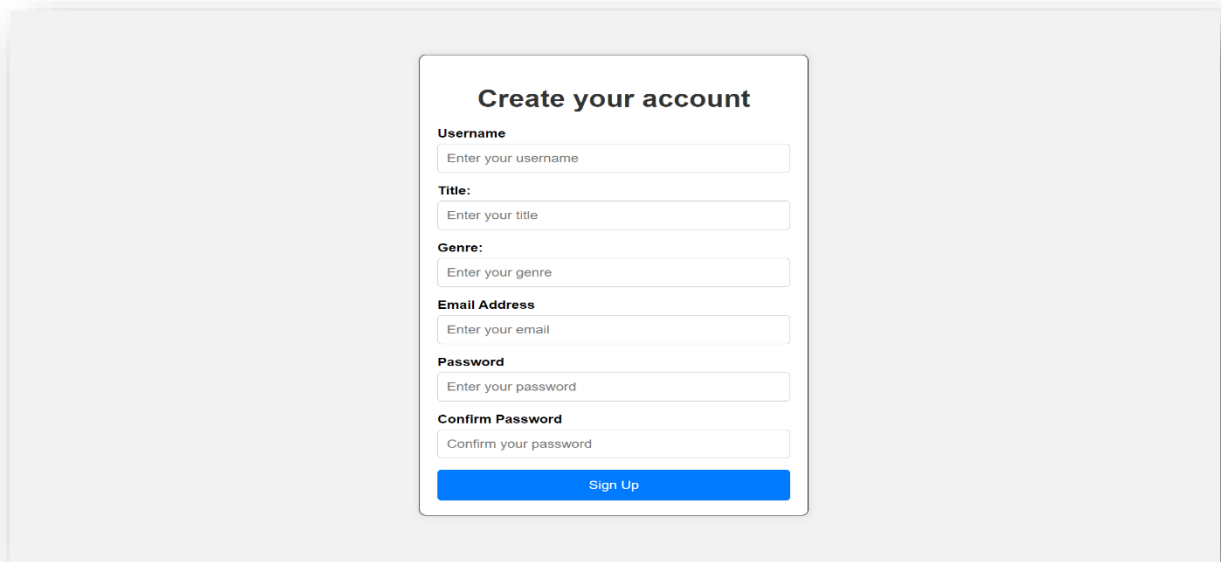
The image shows a registration form titled "Create your account" in bold black text. The form is contained within a white rectangular box with a thin gray border, set against a light gray background. The form includes several input fields, each with a label above it: "Username" (with placeholder text "Enter your username"), "Title:" (with placeholder text "Enter your title"), "Genre:" (with placeholder text "Enter your genre"), "Email Address" (with placeholder text "Enter your email"), "Password" (with placeholder text "Enter your password"), and "Confirm Password" (with placeholder text "Confirm your password"). At the bottom of the form, there is a blue rectangular button with the text "Sign Up" in white.

Figure 10 : Page inscription

iii. Page de connexion

Page qui donne l'accès à notre plateforme

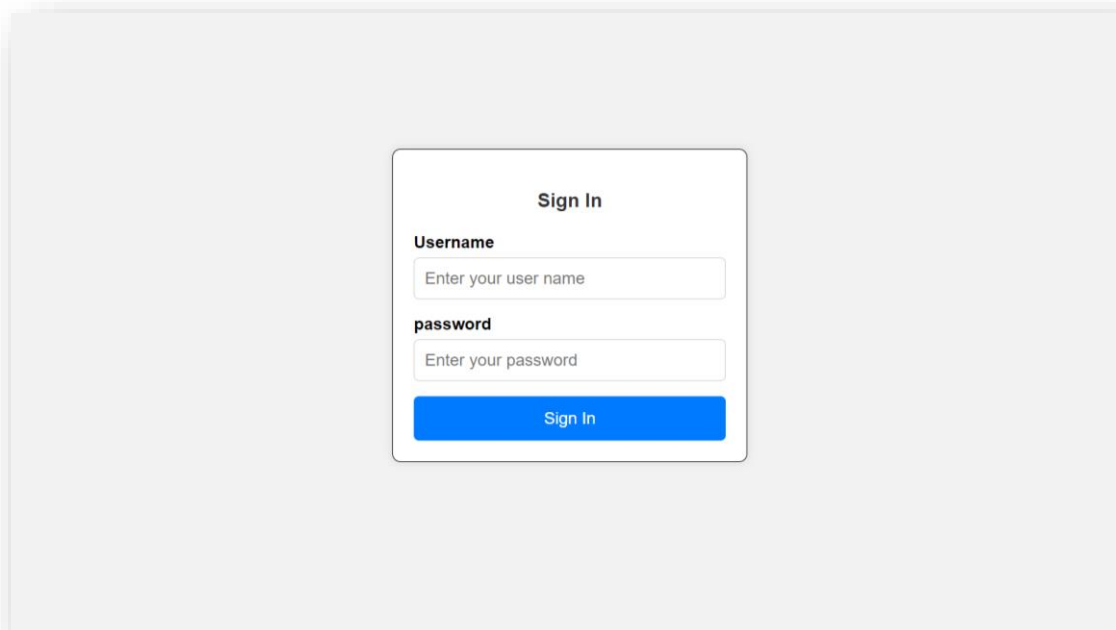
A screenshot of a 'Sign In' form. The form is centered on a light gray background. It has a title 'Sign In' at the top. Below the title, there are two input fields: one for 'Username' with the placeholder text 'Enter your user name', and one for 'password' with the placeholder text 'Enter your password'. Below these fields is a blue button with the text 'Sign In'.

Figure 11 : Page de connexion

iv. Page d'accueil

la première interface affichée après l'inscription dans notre site web, il donne des statistiques sur la violence des femmes, et des liens rapide en haut de la page de même nos services, dans la derrière élément de cette interface est une liste des livres recommander par notre système.

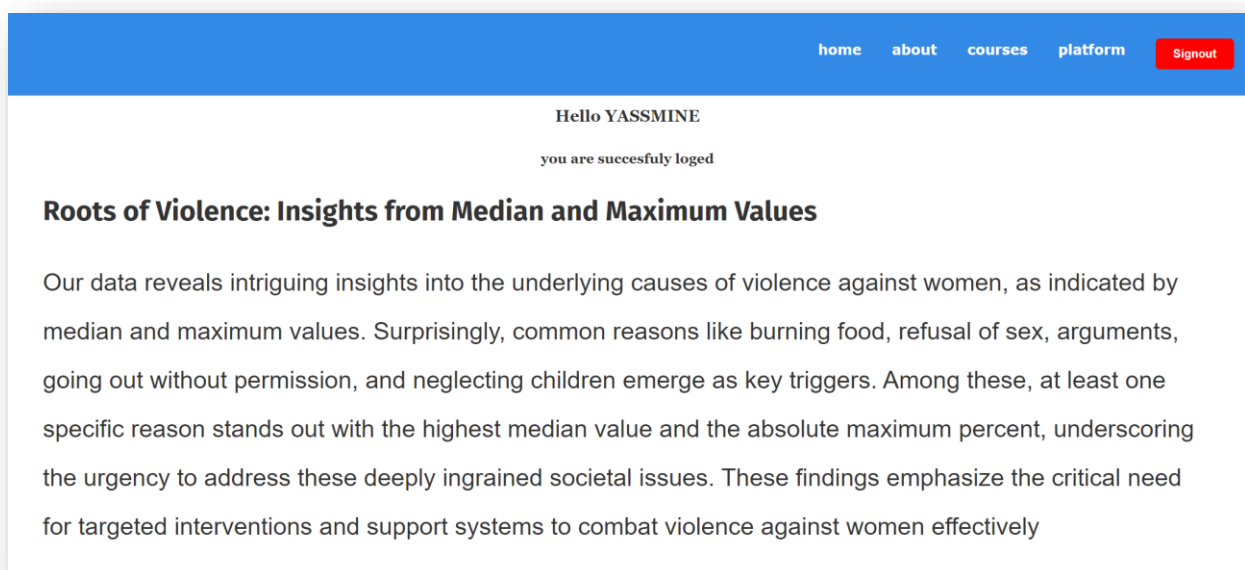
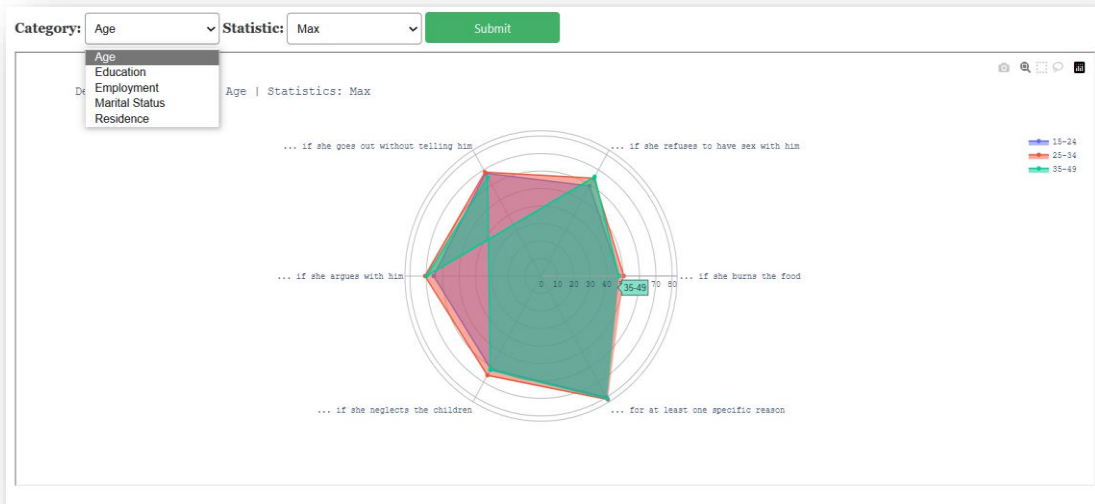
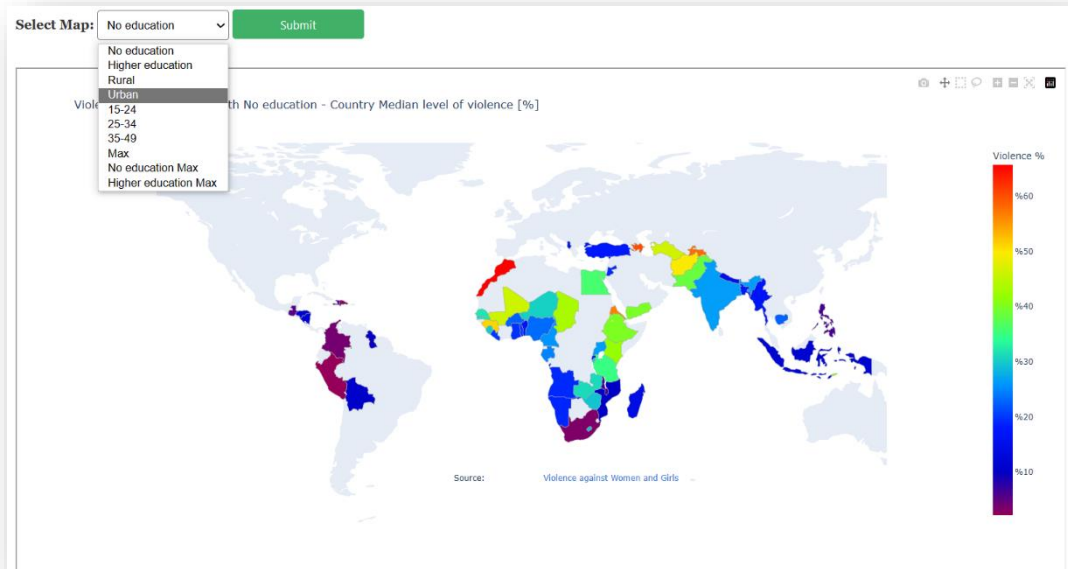
A screenshot of a web application's home page. At the top is a blue navigation bar with links for 'home', 'about', 'courses', and 'platform', and a red 'Signout' button. Below the navigation bar, the text 'Hello YASSMINE' and 'you are succesfully logged' is displayed. The main content area features a heading 'Roots of Violence: Insights from Median and Maximum Values' followed by a paragraph of text discussing data on violence against women.

Figure 12 : Page d'accueil

Méthodologie



home about courses platform

Course 1

Title : Rented Rooms: A Collection of Short Fiction

Auteur : Linda A. Lavid

Course 2

Title : Juniper Tree (Modern Fiction S.)

Auteur : B. Comyns

Course 3

Title : A Model Crime: A True Fiction

Auteur : Curtis Gathje

Course 4

Title : Bestial Noise: The Tin House Fiction Reader

Auteur : Rob Spillman

Course 5

Title : Ceremony (Contemporary American Fiction Series)

Auteur : Leslie Marmon Silko

v. Page des plateformes

Une interface qui donne l'accès a les autres plateformes de femme

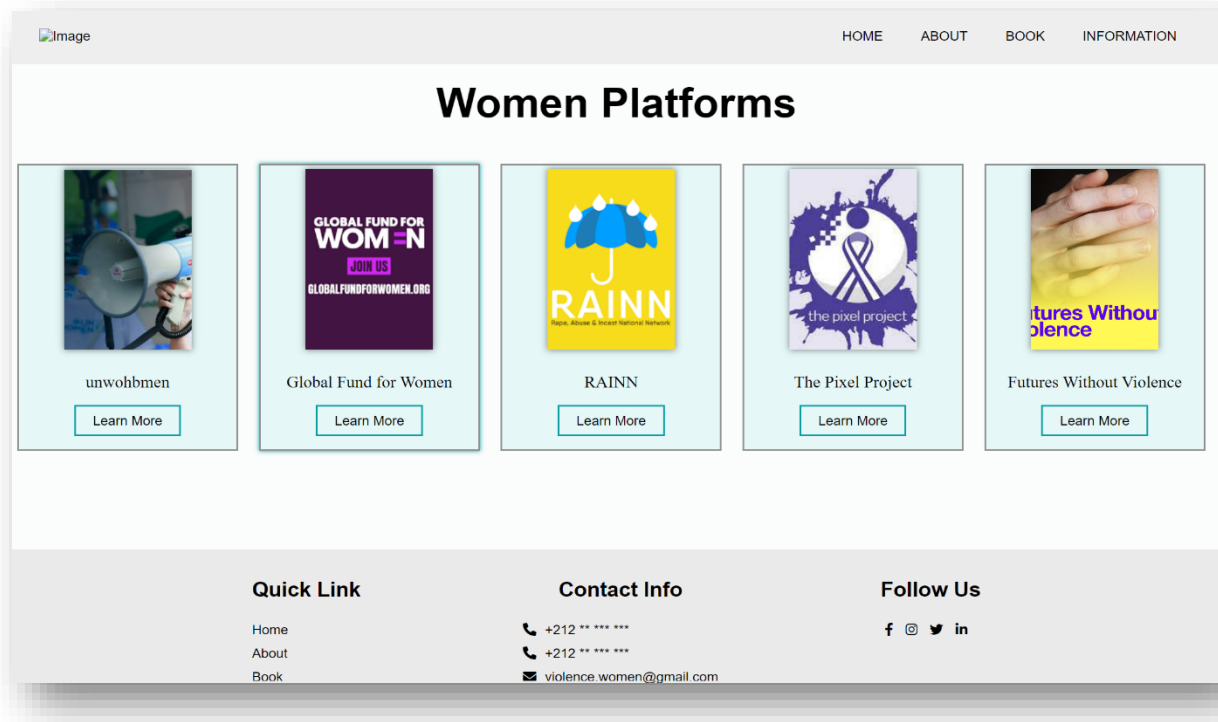


Figure 13 : Page des plateformes

vi. Page des livres

Une interface qui donne l'accès a notre livre recommander pour toute les utilisateurs

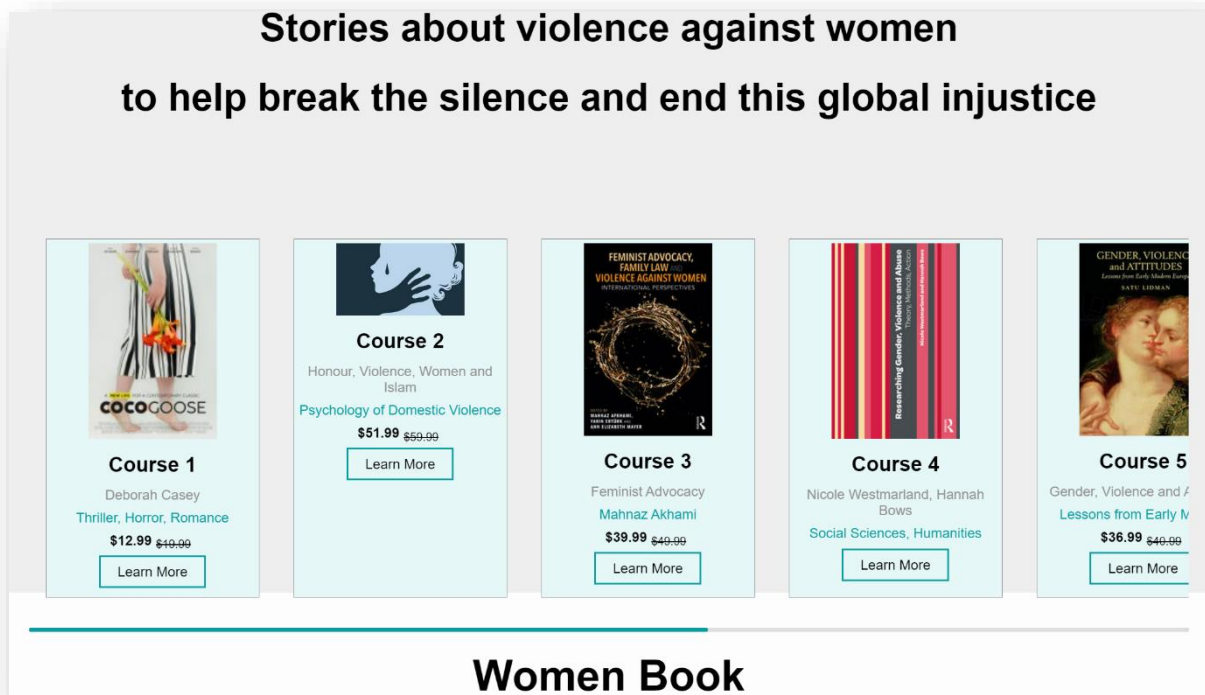


Figure 14 : Page des livres

V. Conclusion

Dans ce chapitre, nous avons examiné la structure de notre dataset utilisé par le système de recommandation des livres, et nous avons évalué les performances de notre modèle de recommandation. Ensuite, nous avons abordé l'implémentation de notre plateforme. Ensuite dans le dernier chapitre, nous concluons notre rapport par une conclusion générale et des perspectives pour l'avenir.

Chapitre 5 : Conclusion générale

I. Introduction

La plateforme de réduction de la violence contre les Femmes vise à aider les femmes et à réduire la violence à leur égard en offrant des ressources éducatives, de soutien et de développement personnel. Pour renforcer son impact et répondre aux besoins évolutifs de son public, la plateforme envisage d'intégrer une composante de développement de carrière basée sur l'IA. Ce rapport explore les perspectives d'avenir pour la plateforme, en mettant en lumière les avantages et les implications de cette initiative.

II. Les perspectives en future

Développement de compétences professionnelles : nous visons à intégrer des outils d'apprentissage automatique pour recommander des cours et des ressources de développement professionnel adaptés aux compétences et aux objectifs de carrière des femmes.

Orientation professionnelle personnalisée : L'utilisation de l'intelligence artificielle pour fournir des conseils de carrière personnalisés en fonction des intérêts, des compétences et du contexte professionnel de chaque utilisatrice.

Soutien psychologique et développement personnel : L'intégration des ressources pour le bien-être mental et émotionnel, ainsi que des outils de développement personnel pour renforcer la confiance en soi et la résilience des femmes.

Partenariats avec des ONG et des institutions : Faire des collaboration avec des organisations non gouvernementales et des institutions spécialisées dans la lutte contre la violence contre les femmes pourrait enrichir votre plateforme en offrant un soutien professionnel et des ressources supplémentaires.

Services de conseil en ligne : L'intégration des services de conseil en ligne avec des professionnels qualifiés pourrait offrir un soutien immédiat et confidentiel aux femmes qui en ont besoin.

Évolution de l'IA : Continuer à améliorer l'aspect d'intelligence artificielle de la plateforme pour des recommandations plus précises et personnalisées en fonction des besoins spécifiques des utilisatrices.

Communauté en ligne : Créer une communauté en ligne où les femmes peuvent partager leurs expériences, se soutenir mutuellement et trouver un sentiment de solidarité.

Accès à des services juridiques : Faciliter l'accès à des services juridiques pour aider les femmes à comprendre leurs droits et à obtenir une assistance juridique en cas de besoin.

III. Conclusion

Le présent rapport présente une synthèse du travail que nous réalisons tout le long de mon projet de fin d'étude, ce projet de fin d'étude permis de mettre en œuvre notre savoirs et notre connaissances théoriques, et de les enrichir par la pratique, ainsi que la maîtrise d'un ensemble des nouvelles technologies et d'améliorer notre connaissance langages de développements. Plus, Cette expérience a été une opportunité pour approfondir dans le travail en équipe, assimiler ses concepts et manipuler ses outils. Durant la période de projet de fin d'étude notre mission était de concevoir et réaliser une plateforme de réduction de la violence contre les femmes vise à aider les femmes et à réduire la violence à leur égard en offrant des ressources éducatives. D'autre part, j'ai eu la chance d'intégrer dans un milieu de travail professionnel, et de travailler avec une équipe sous un encadrant dans un environnement très convivial. Ce qui nous permis de confronter les difficultés du monde du travail. On peut dire enfin que les objectifs que nous fixons au départ sont en grandes parties atteintes. En fait, ce projet a été une véritable expérience de travail. Nous sommes entièrement satisfaite du travail que nous accompli et Nous sommes fière d'avoir contribué à faire progresser ce projet. J'espère sincèrement que cette expérience au sein de Ecole Supérieure de Technologie nous prépare de manière optimale pour mon intégration future dans le domaine professionnel.

Référence

- [1] : API Twitte, <https://tweetdelete.net/fr/ressources/what-is-twitter-api-all-the-details-about-this-interface/>
- [2] : Système de recommandation, <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00592-5>
- [3] : HTML, <https://developer.mozilla.org/fr/docs/Web/HTML>
- [4]: CSS, <https://www.w3schools.com/css/>, <https://getbootstrap.com/docs/5.3/getting-started/introduction/>
- [5] : JAVA script, <https://www.tutorialspoint.com/javascript/index.htm>.
- [6] : Python, <https://www.python.org/doc/essays/blurb/>
- [7] : Visual Studio Code, <https://code.visualstudio.com/docs/editor/whyvscode>
- [8] : Plotly, <https://domino.ai/data-science-dictionary/plotly>
- [9]: Django, <https://developer.mozilla.org/en-US/docs/Learn/Server-side/Django>
- [10] : Scikit-Learn, <https://datascientest.com/tout-savoir-sur-scikit-learn>
- [11] : Modèle de recommandation, <https://medium.com/analytics-vidhya/maths-behind-word2vec-explained-38d74f32726b>
- [12] : Word2vec, <https://medium.com/analytics-vidhya/maths-behind-word2vec-explained>
- [13] : Data Visualization, <https://www.geeksforgeeks.org/python-plotly-tutorial/%20>