Une Histoire Computationnelle du Genre dans la Fiction

Apprentissage Machine & Littérature

Ismail El Hadrami, Ottilie Candau, Marc Noujaim, Milica Prugic, Pedro Cabrera Ramirez, Jean Barré

30 november 2022

PSL Intensive Week DHAI

Outline

Introduction

Introduction

Études Littéraires Computationnelles

- Apprentissage Machine & Fouille de textes pour modéliser des concepts dans de grands corpus littéraires.
- Un concept clé : Distant Reading - Lecture Distante - Franco Moretti.
- Le projet : Focus sur la notion de genre dans la fiction.

Principale question de recherche

Quels sont les enjeux dans la représentation du genre dans la fiction sur ces deux derniers siècles de production littéraire ?

- Évaluez les signes genrés que les écrivains utilisent pour décrire des personnages.
- Les hommes fictifs sont-ils très différents des femmes fictives ?
- Dans quelle mesure les signes publics du genre influencent la caractérisation en général ?

Reproduire des résultats de recherches

Underwood, Ted, David Bamman, Sabrina Lee. “The Transformation of Gender in English-Language Fiction”. Journal of Cultural Analytics, 3, 2, 2018. doi : https ://doi.org/10.22148/16.019

Principaux résultats :

1. Un modèle prédictif entraîné avec des mots comme caractéristiques et des étiquettes féminines et masculines perd en précision entre les années 1980 et aujourd’hui
2. Le temps accordé aux personnages féminins est 3 fois moins important dans le cas d’un auteur masculin
3. Suivre les mots individuels liés au genre

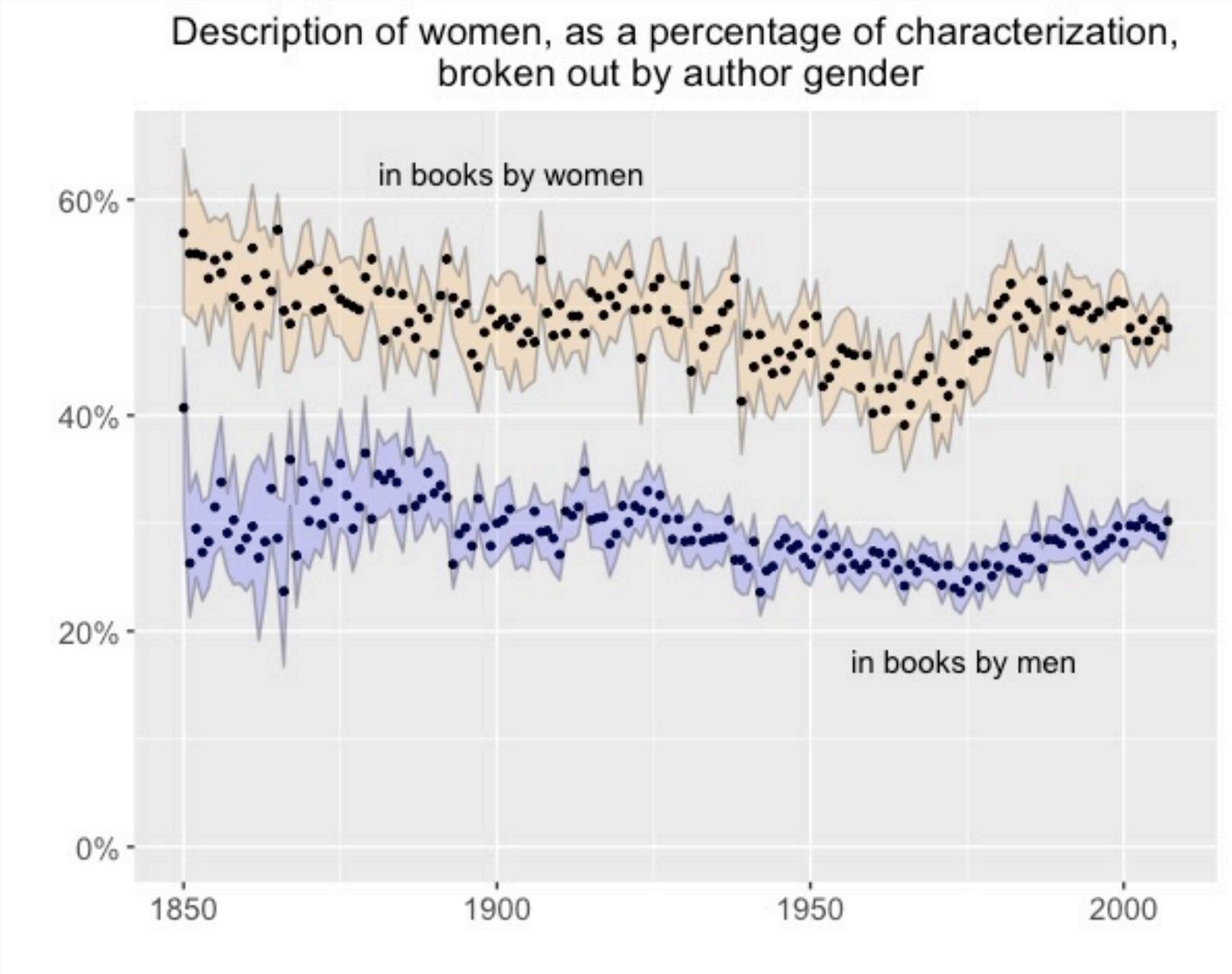

Figure 1 : Différenciation du temps d'écran en fonction du sexe de l'auteur

Principale Tâche : Prédire le genre des personnages

Principale Tâche : Prédire le genre des personnages

Prédiction du genre en fonction des mots qui caractérisent les personnages

- Data Annotation
- Data manipulation - Pandas
- Feature Engineering - NLP - Spacy
- Supervised Machine Learning - SKLearn
- Data Visualization - Matplotlib & Seaborn

Corpus Chapitres

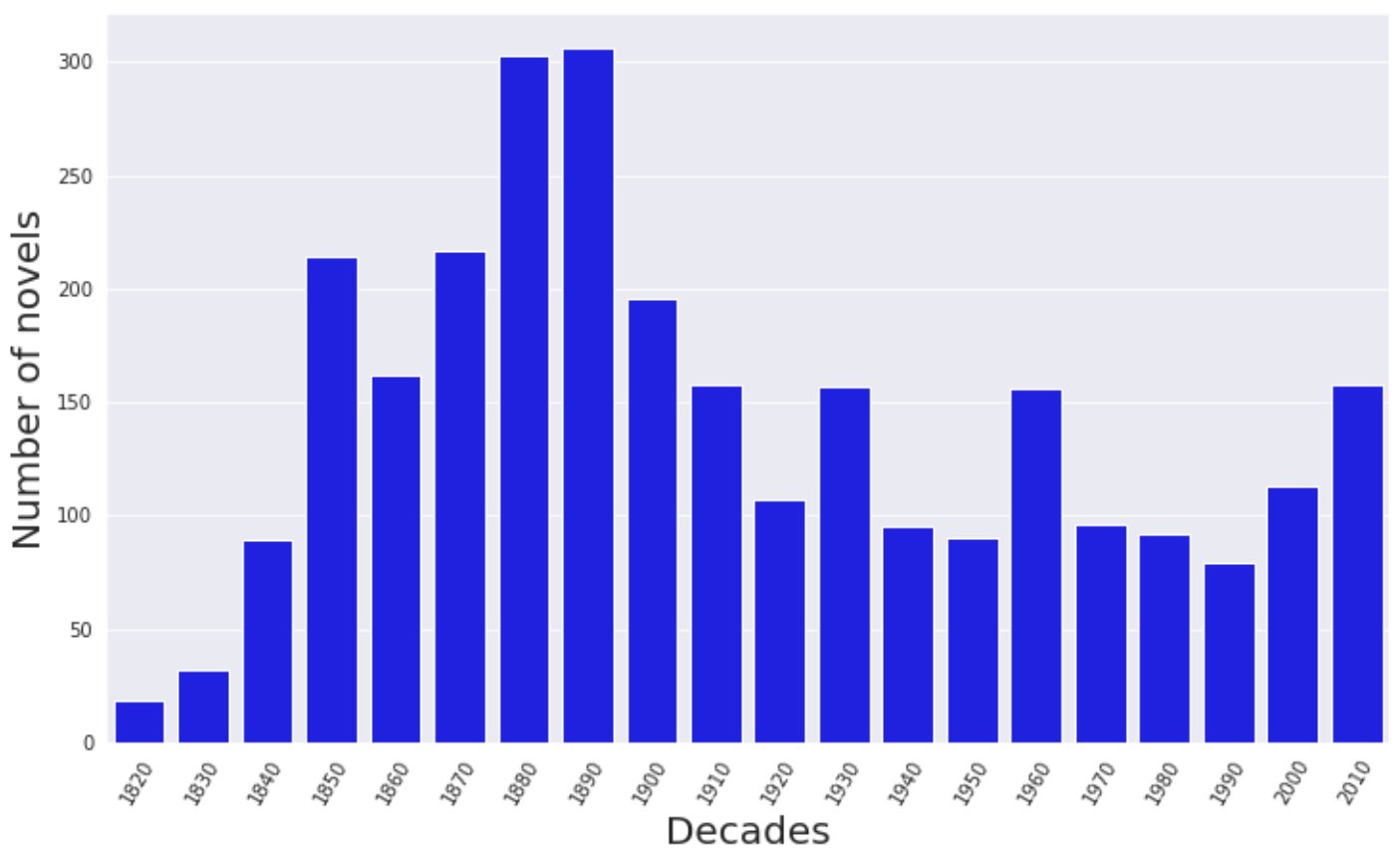

Figure 2 : Distribution des textes sur le temps

Corpus Chapitres

Figure 3 : Pourcentage de livres de fiction écrits par des femmes

French BookNLP

NLP pipeline scaling to books

- Entity recognition (PER, FAC, TIME, ORG, LOC)
- Clustering Names
- Co-reference resolution

Méthode

Annotation

- Les données utilisées sont fournies par BookNLP.
- 10 personnages - les plus fréquents
- 10 tokens environnants
- Les différentes étiquettes sont : Homme, Femme et Neutre
- La tâche consistait à définir les genres des personnages dans 83 romans choisis aléatoirement.

Extraction de données & Modèle

Feature extraction

- Sac de mots : utilisation des mots les plus courants et de leur, fréquence pour chaque caractère.
- TF-IDF : Mesure de l'originalité d'un mot en comparant le nombre de fois qu'un mot apparaît dans un document avec le nombre de documents dans lesquels il apparaît.
- Doc2Vec : outil NLP permettant de vectoriser du texte

Estimateur

- Machine à Vecteur de support

Résultats

Efficacité du modèle

1. BoW : 53%
2. TF-IDF : 66%
3. Doc2Vec : 85% - CV : 79.2%

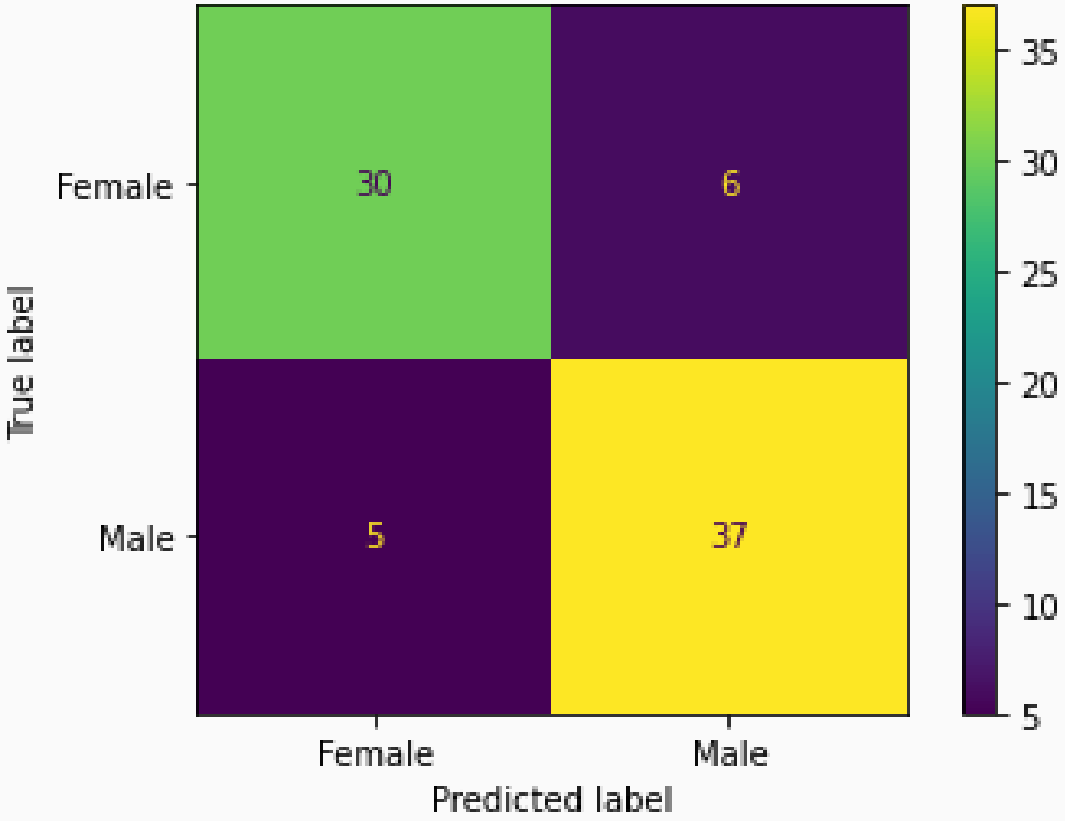

Figure 4 : Matrice de confusion du modèle

Efficacité du modèle

	precision	recall	f1-score	support
Female	0.857143	0.833333	0.845070	36.000000
Male	0.860465	0.880952	0.870588	42.000000
accuracy	0.858974	0.858974	0.858974	0.858974
macro avg	0.858804	0.857143	0.857829	78.000000
weighted avg	0.858932	0.858974	0.858811	78.000000

Figure 5 : Métriques d'évaluation pour une classification binaire

Visualisation des résultats 1/4

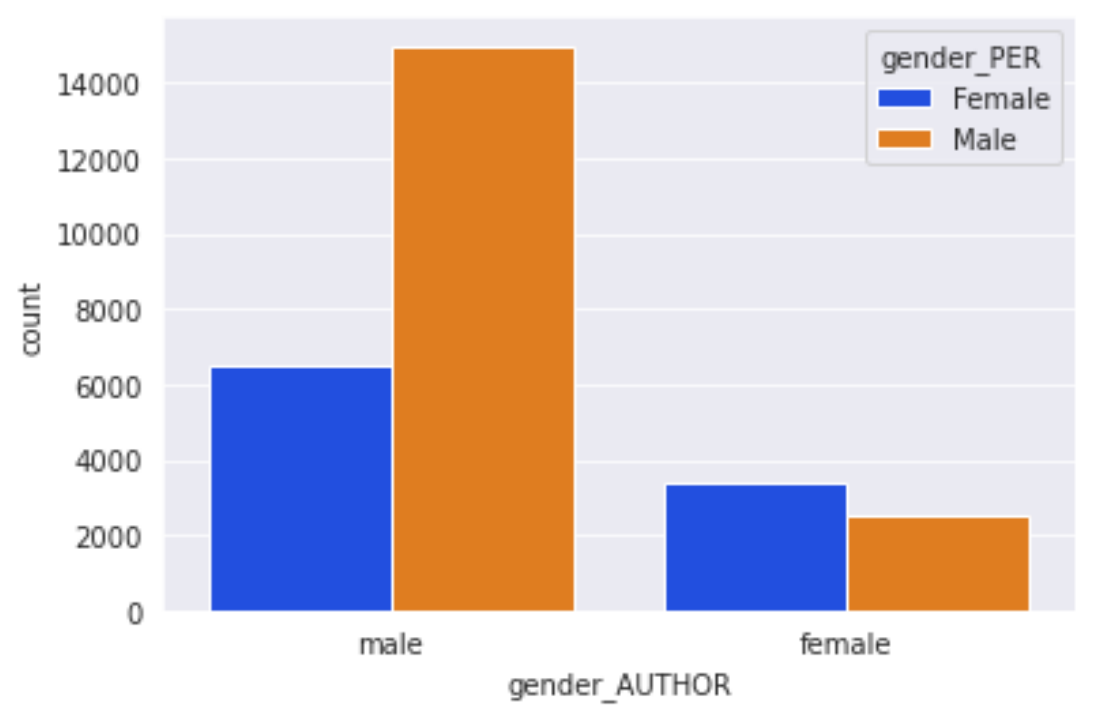

Figure 6 : Proportion de la caractérisation des femmes par des auteurs et des autrices

Female Authors - 57% Female Characters, 43% Male Characters

Male Authors - 30% Female Characters, 70% Male Characters

Visualisation des résultats 2/4

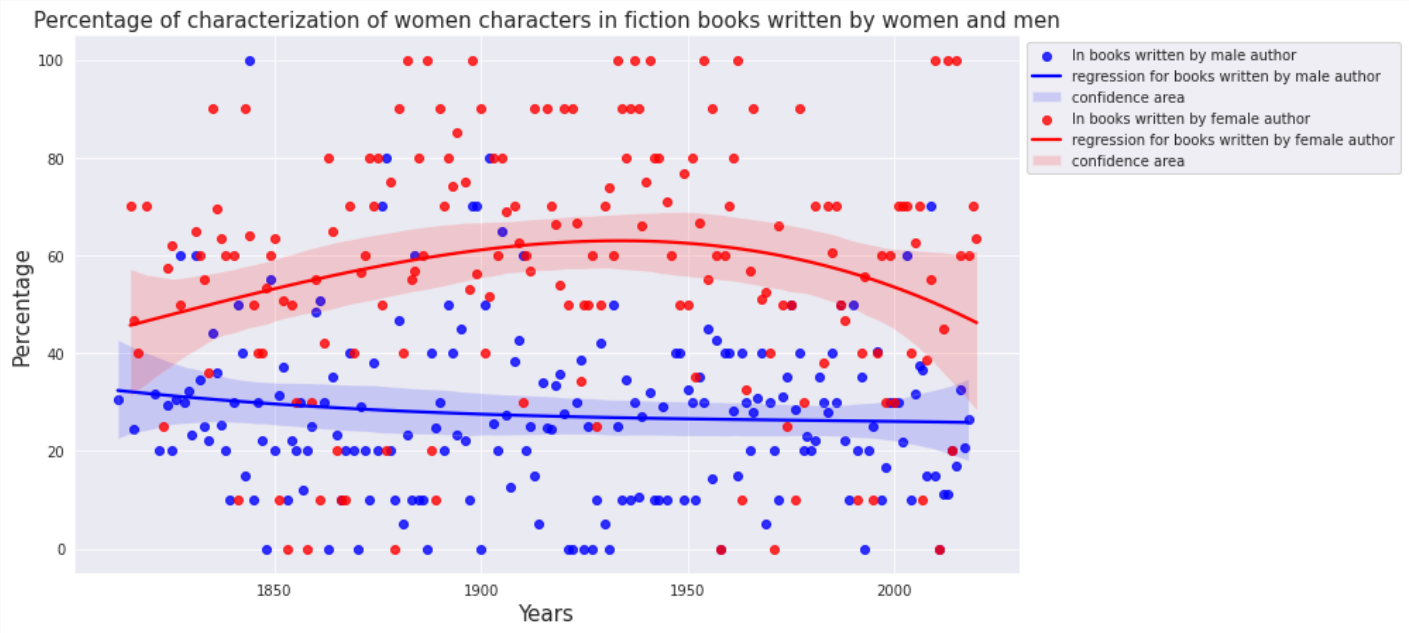

Figure 7 : Proportion de la caractérisation des femmes par des auteurs et des autrices, moyenne par année

Visualisation des résultats 3/4

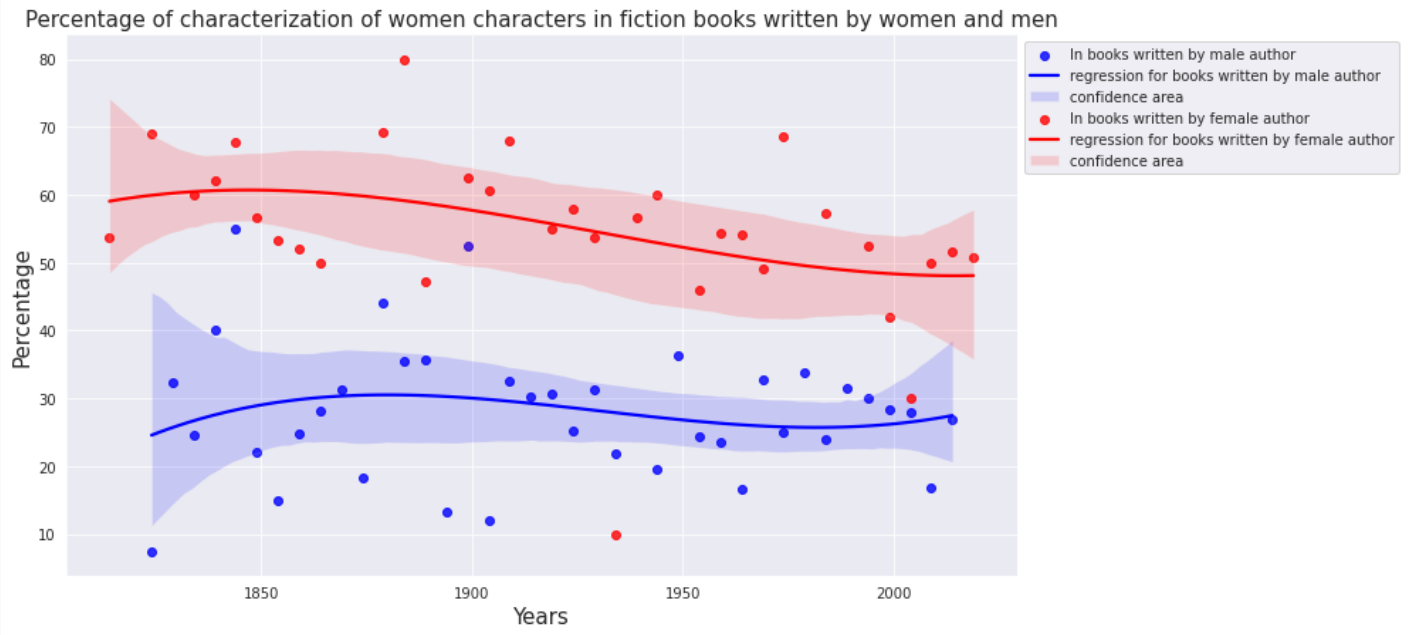

Figure 8 : Proportion de la caractérisation des femmes par des auteurs et des autrices, moyenne tous les cinq ans

Visualisation des résultats 4/4

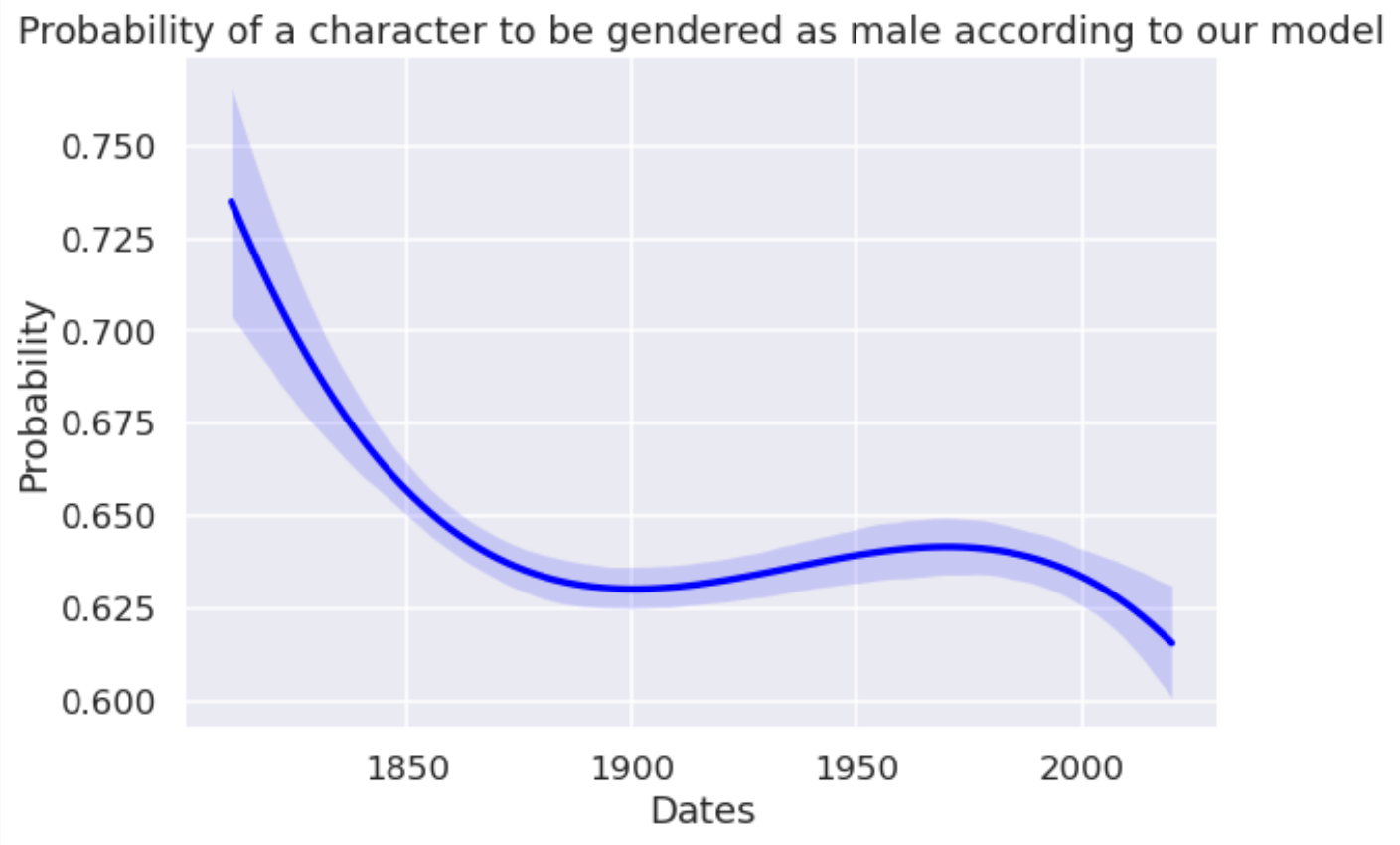

Figure 9 : Probabilité d'être caractérisé comme un homme pour notre modèle

Quelques mots genrés dans la fiction 1/3

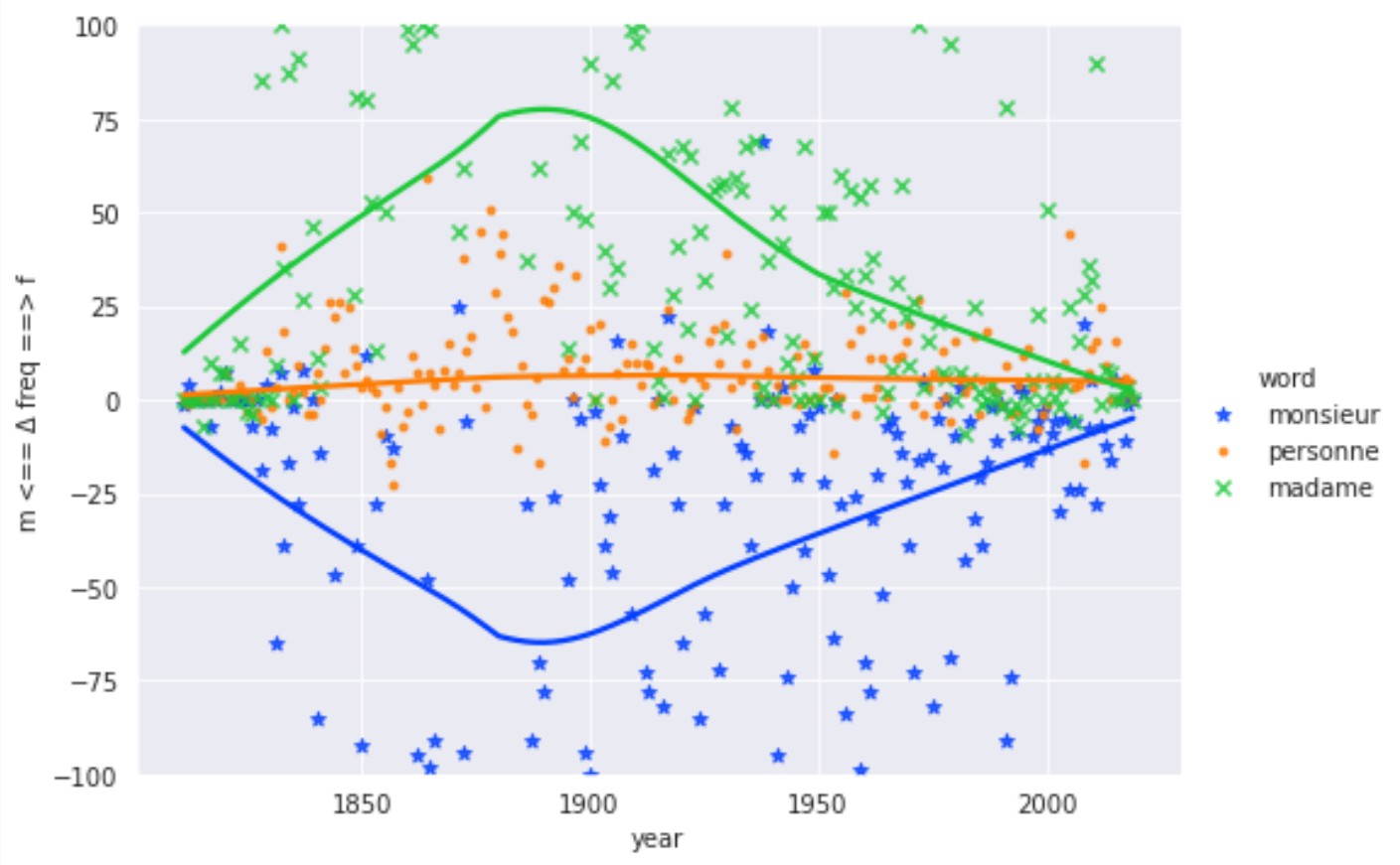

Figure 10 : Comment les hommes et les femmes sont caractérisés par mots évidents : homme et femme

Quelques mots genrés dans la fiction 2/3

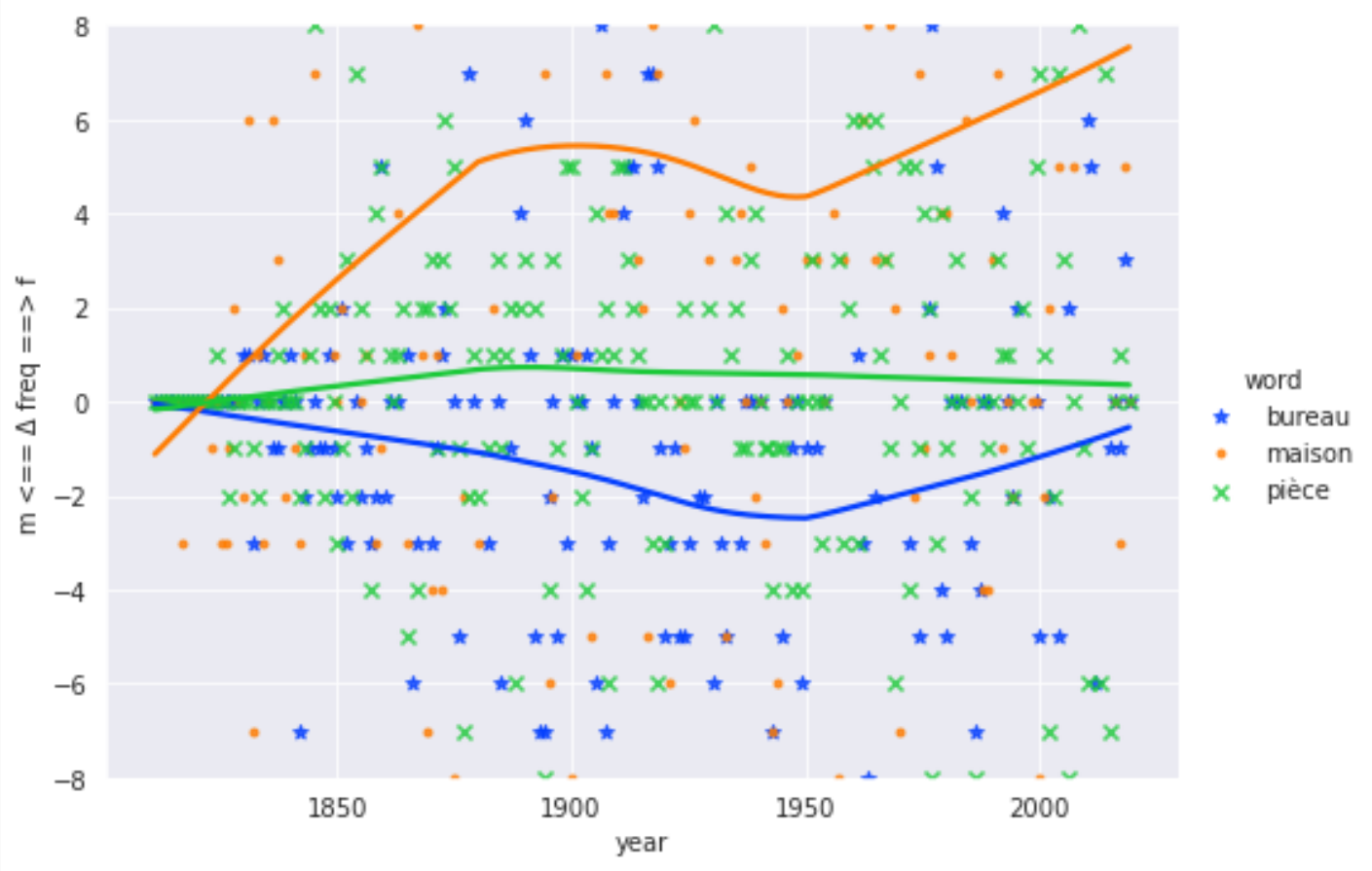

Figure 11 : Comment les hommes et les femmes sont caractérisés dans l'espace

Quelques mots genrés dans la fiction 3/3

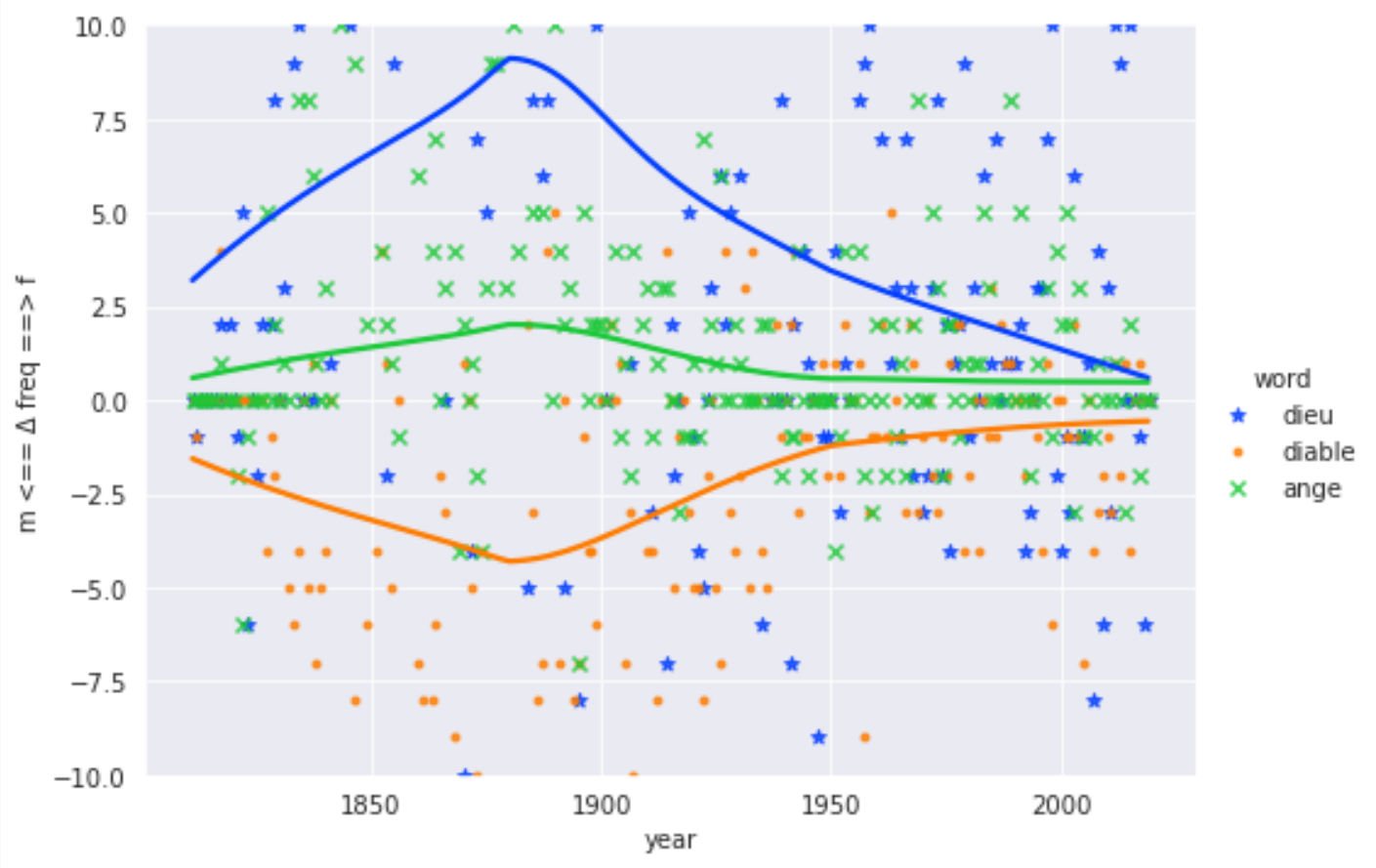

Figure 12 : Comment les hommes et les femmes sont caractérisés par des mots religieux

Conclusion

Conclusion

- Nous avons pu évaluer dans quelle mesure la caractérisation littéraire est liée aux stéréotypes de genre.
- Il existe des mots individuels / champs lexicaux liés aux stéréotypes de genre.
- La proportion de caractérisation des personnages féminins dépend fortement du sexe de l'auteur.
- Les auteurs masculins écrivent moitié moins sur les personnages féminins que les auteurs féminins.

Q & A

Code, data, slides sur github :
https://github.com/crazyjeannot/dhai_intensive_week