A Computational History of Gender in French Fiction

Machine Learning & Literature

Jean Barré

30, november 2022

PSL Intensive Week DHAI

Outline

1. Introduction
    - Main research question
    - Reproduce research results
2. Main Task : Predict character's gender
    - Corpus
    - French BookNLP
3. Method
    - Annotation
    - Operationalization of genre
4. Results
    - Model Accuracy
    - Results visualizations

Introduction

Introduction

Computational Literary Studies

- Machine learning & Text mining to model concepts in large literary corpora.
- A key concept : Distant Reading - Franco Moretti.
- The project : Focus on the notion of gender in fiction.

Main research question

What is at stake in the representation of gender in fiction over the last two centuries of literary production ?

- Evaluate the gendered signs writers use to describe characters.
- Are fictional men very different from fictional women ?
- To what extent do public signs of gender influence characterization in general ?

Reproduce research results

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

Main Results :

1. A predictive model trained with words as characteristics and female and male labels loses accuracy between the 1980s and today
2. The time given to female characters is 3 times less in the case of a male author
3. Track individual words related to gender

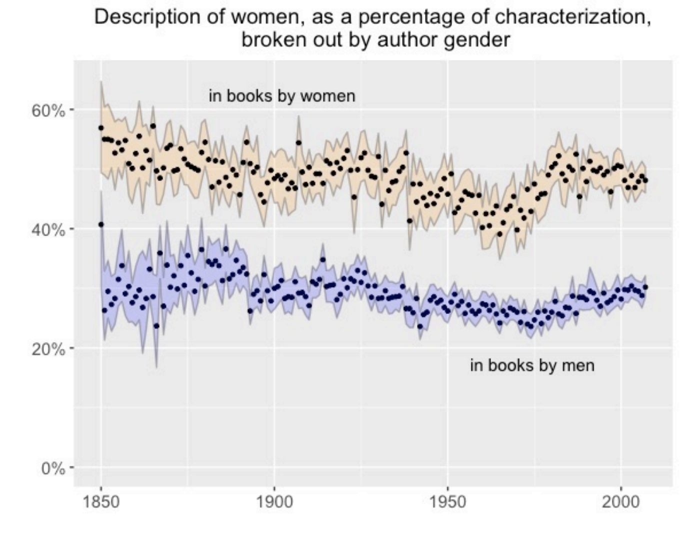


Figure 1 : Screen-time differentiation broken out by author’s gender

Main Task : Predict character's gender

Main Task : Predict character’s gender

Gender prediction based on words that characterize the characters

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

Corpus Chapitres

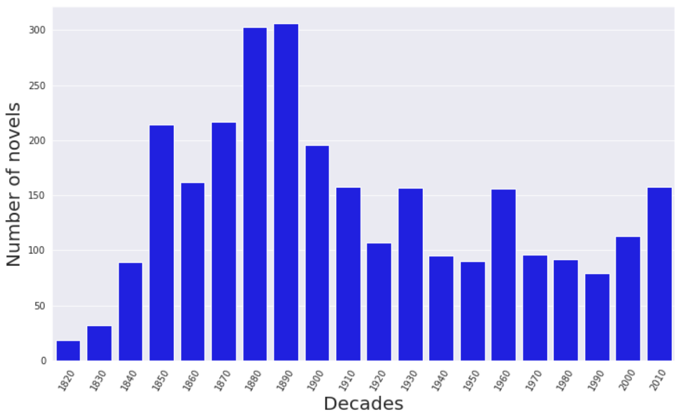


Figure 2 : Time distribution of the texts

Corpus Chapitres

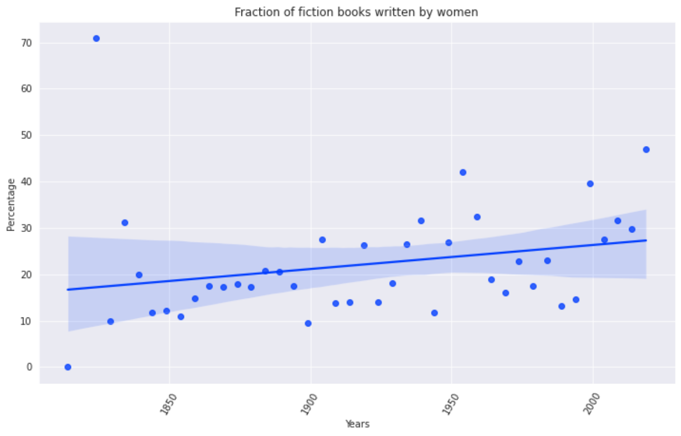


Figure 3 : Percentage of books written by female writers

French BookNLP

NLP pipeline scaling to books

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

Method

Annotation

- The data used is provided by BookNLP.
- 10 most frequent characters
- 10 surrounding tokens
- 3 tags : Male, Female and Neutral
- The task was to define the genres of characters in 83 randomly selected novels.

Data Extraction and Statistical Model

Feature extraction

- Bag of words : use of the most common words and their frequency for each character.
- TF-IDF : Measures the originality of a word by comparing the number of times a word appears in a document with the number of documents in which it appears.
- Doc2Vec : NLP tool allowing to vectorize text

Estimator

- Support Vector Machine

Results

Model Accuracy

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

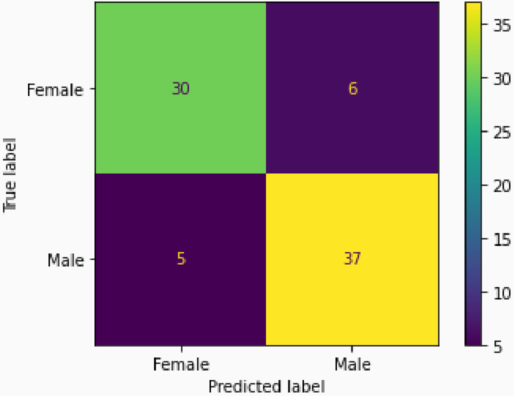
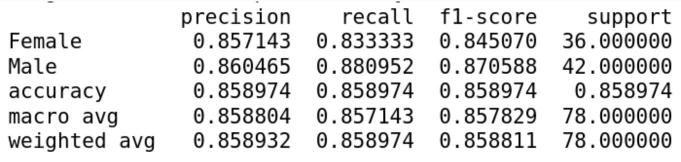


Figure 4 : Confusion Matrix

Model Accuracy



```
              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 : Evaluation metrics for a binary classification

Results visualizations 1/4

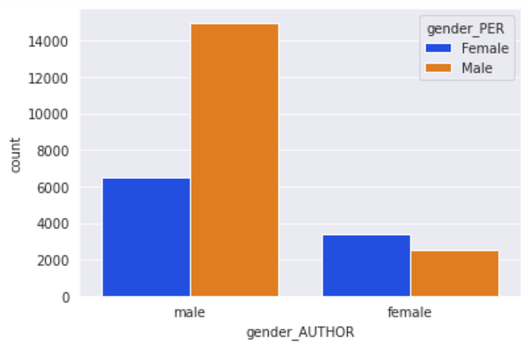


Figure 6 : Proportion of the characterization of women by male and female authors

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

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

Results visualizations 2/4

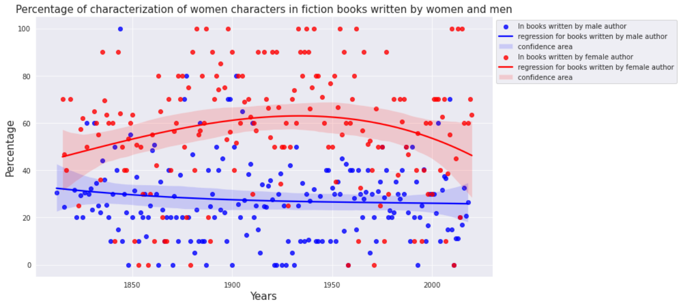


Figure 7 : Proportion of the characterization of women by male and female authors, on average every year

Results visualizations 3/4

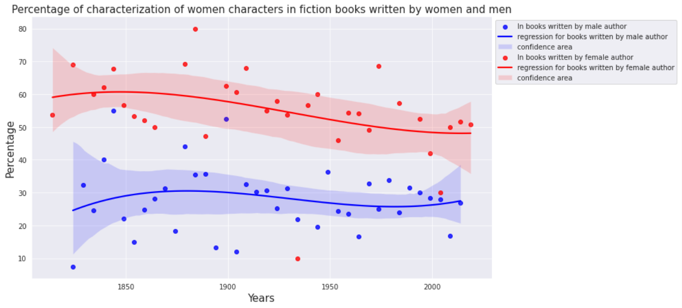


Figure 8 : Proportion of the characterization of women by male and female authors, on average every five years

Results visualizations 4/4

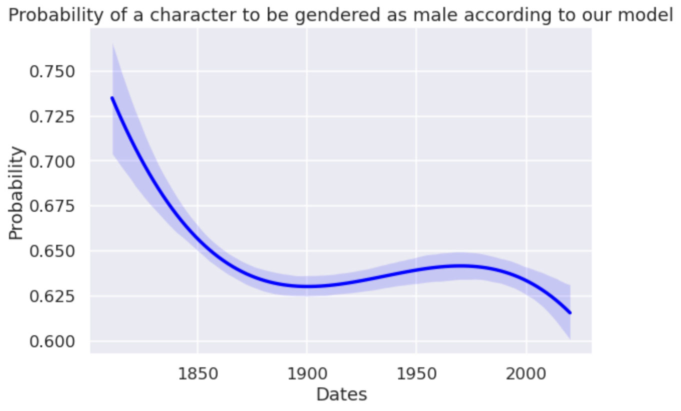


Figure 9 : Probability to be characterized as male for our model

Some gendered words in fiction 1/3

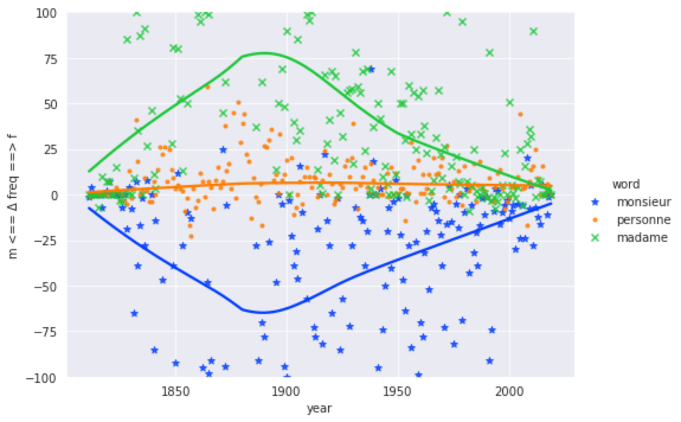


Figure 10 : How men and women are characterized by obvious words : homme et femme

Some gendered words in fiction 2/3

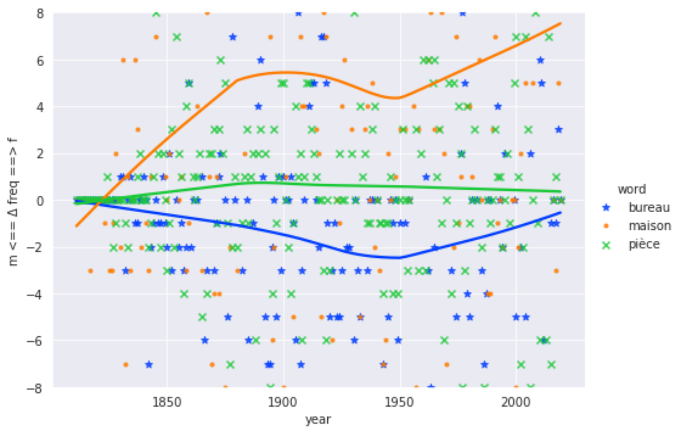


Figure 11 : How men and women are characterized in fictional space

Some gendered words in fiction 3/3

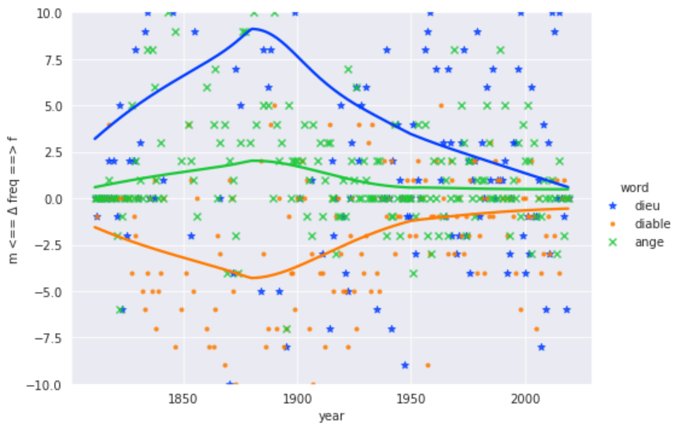


Figure 12 : How men and women are characterized by religious words

Conclusion

Conclusion

- We were able to assess the extent to which literary characterization is related to gender stereotypes.
- There are individual words/lexical fields related to gender stereotypes.
- The proportion of characterization of female characters depends strongly on the gender of the author.
- Male authors write half as much about female characters as female authors.

Q & A

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