A Computional History of Gender

In two centuries of fiction

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

Introduction

- **Antecedent:** Underwood, Ted, David Bamman, Sabrina Lee. “The Transformation of Gender in English-Language Fiction”. *Journal of Cultural Analytics*, 3, 2, 2018.

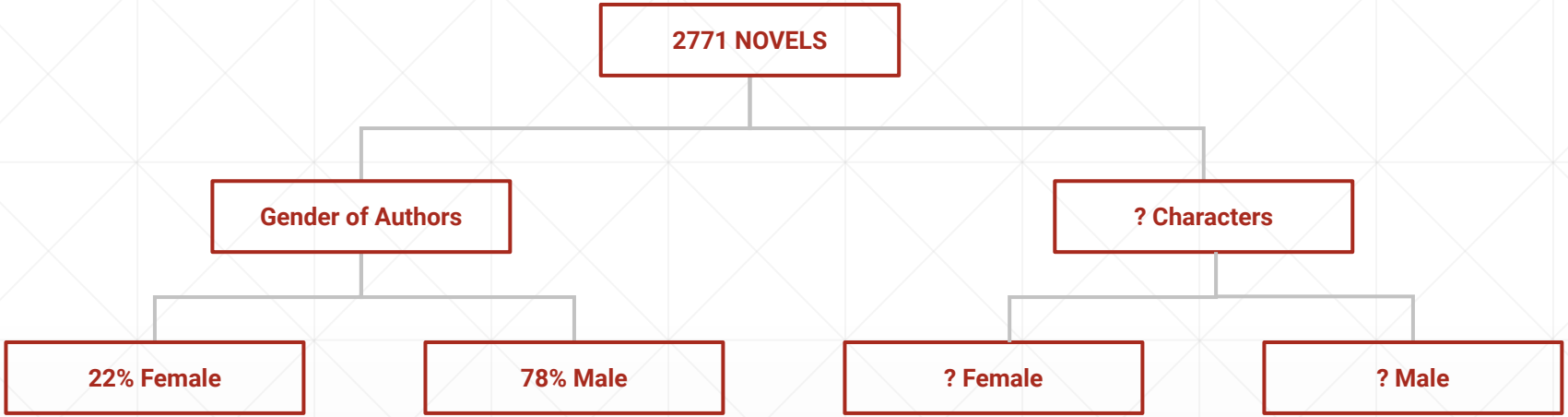
- **Main task:**
  - Predict Character’s Gender over more than two centuries of fiction.
  - Try to reproduce the results of the paper in another corpus and in another language.

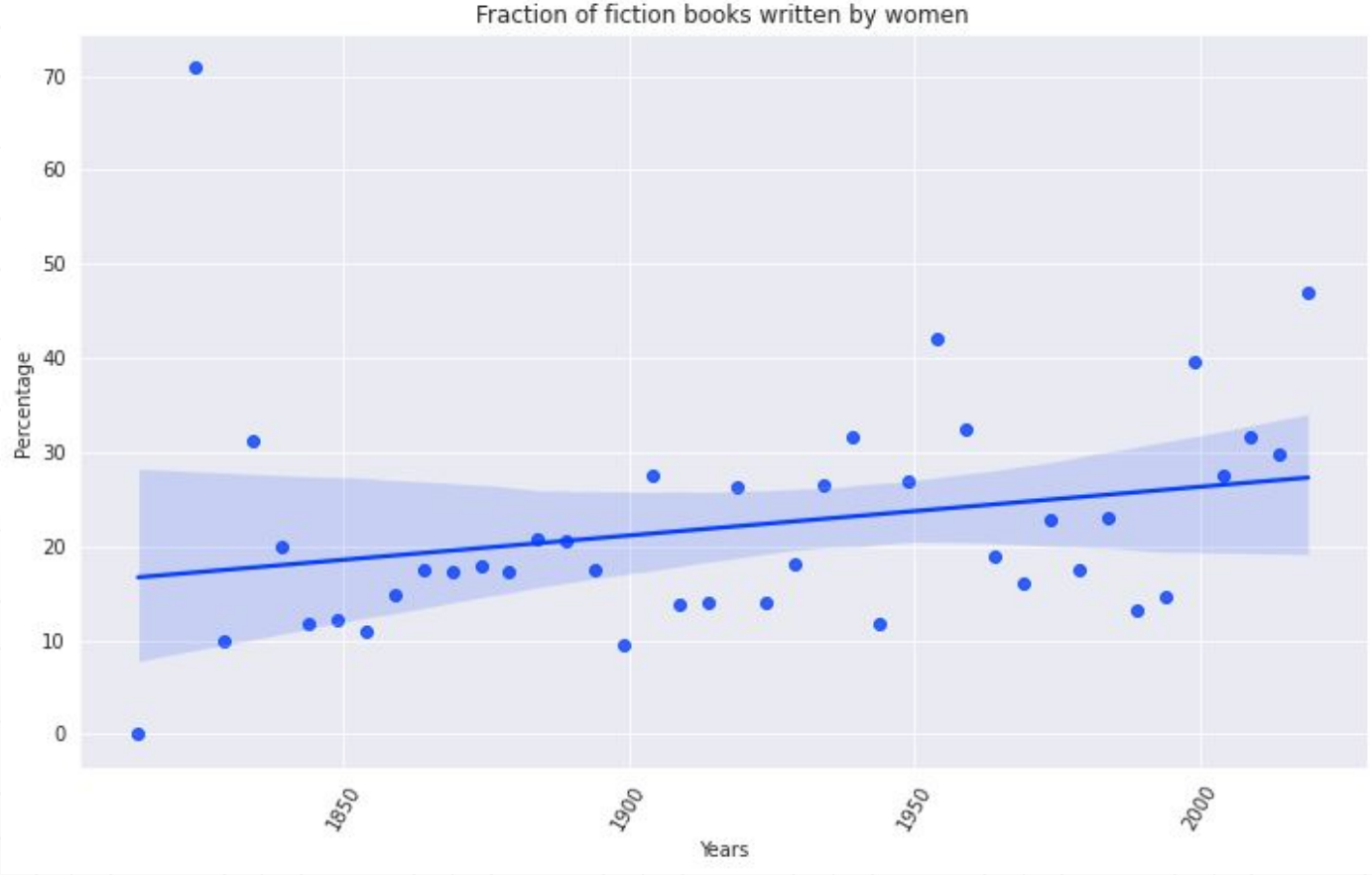
Outline:

I. Data presentation and annotation
II. Features and predictions
III. Visualisations and results analysis

I. Data presentation and annotation

Corpus statistics




Fraction of fiction books written by women
Percentage
0
10
20
30
40
50
60
70
1850
1900
1950
2000
Years

BookNLP

- Entity recognition (PER, FAC, TIME, ORG, LOC)
- Character name clustering (e.g., "Tom", "Tom Sawyer", "Mr. Sawyer", "Thomas Sawyer" -> TOM_SAWYER)
- Coreference resolution

Annotations

- The task consisted of defining genders of characters in the chosen 83 novels
- 10 characters - most frequent ones
- 10 surrounding tokens - mention of PER included
- The different labels are: Male, Female and Neutral
- Data that was used is provided by French BookNLP
- Binary prediction and the least neutrality possible

Challenges faced in the annotation

- “One man + One woman” were characterized as one character
- Je, j’, moi, mes …
- Vous, Ils…
- “Le gros chat”

II. Features and predictions

Pipeline

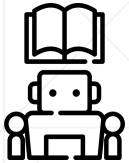
Data annotated

80% for training

Step 1 :

Non annotated data

Feature extraction

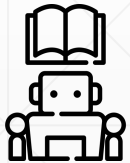


Model training

Classification

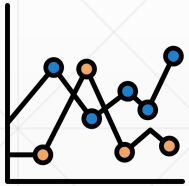
Step 2 :

Non annotated data



Trained model

Annotated data



Beautiful little plots

Feature extraction

- **Bag of words :** Using the most common words and their frequency for each character.

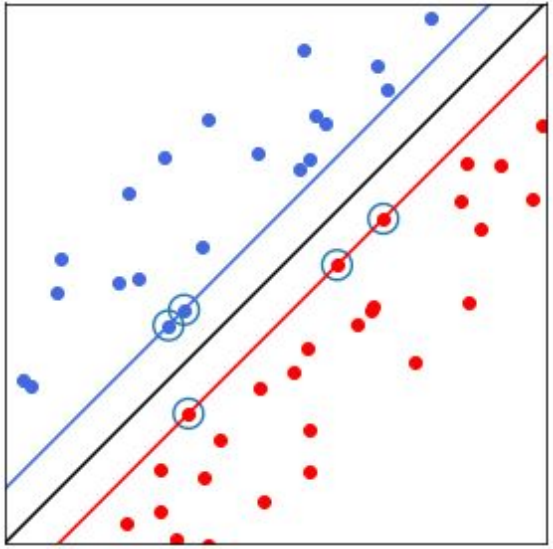
- **TF-IDF :** Measure of originality of a word by comparing the number of times a word appears in a doc with the number of docs the word appears in

- **Doc2Vec :** an NLP tool for representing documents as a vector and is a generalizing of the Word2Vec method.

Model

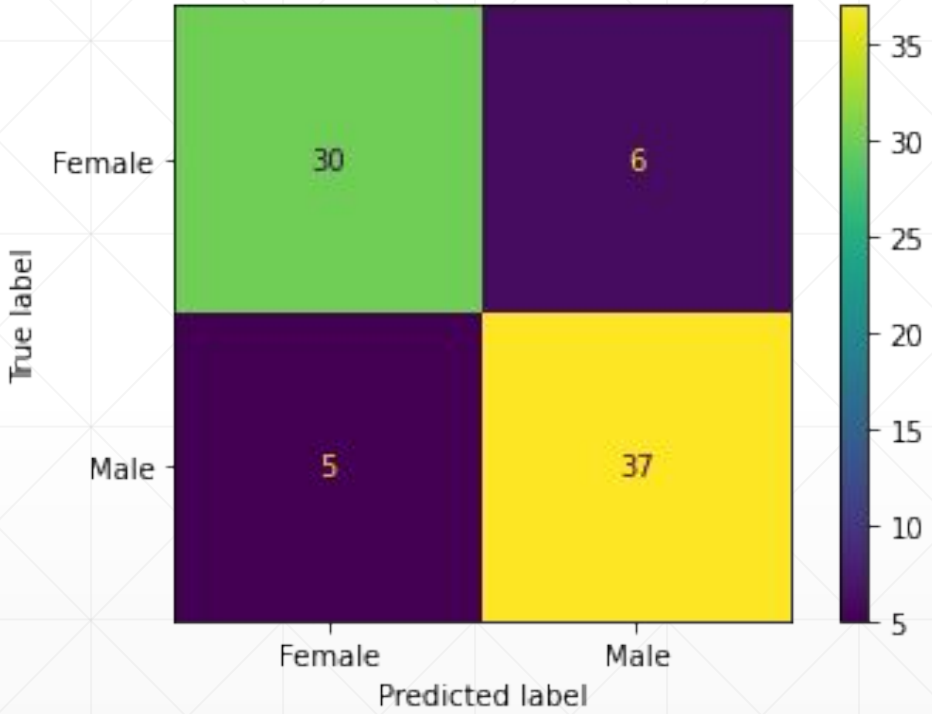
We use a Support Vector Machine (SVM) for the gender classification with an ‘rbf’ kernel :

- Support vectors are the data points that lie closest to the decision surface (or hyperplane)
- We draw the hyperplane that optimizes the margin between the support vectors
- We use different kernels when the data are not linearly separable

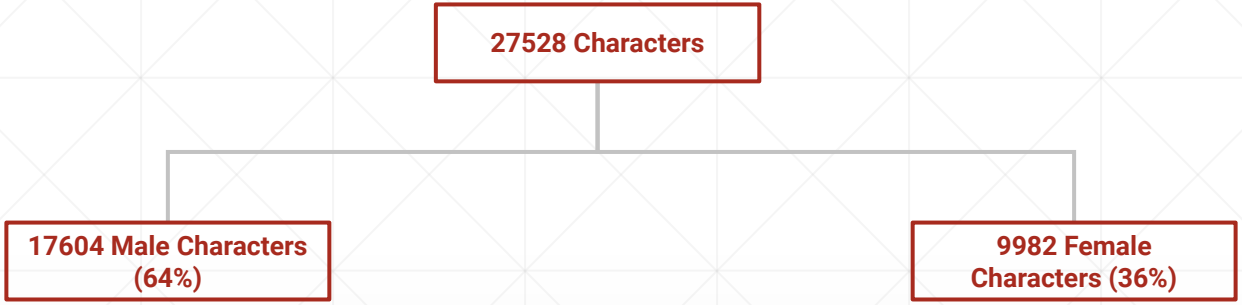


Results

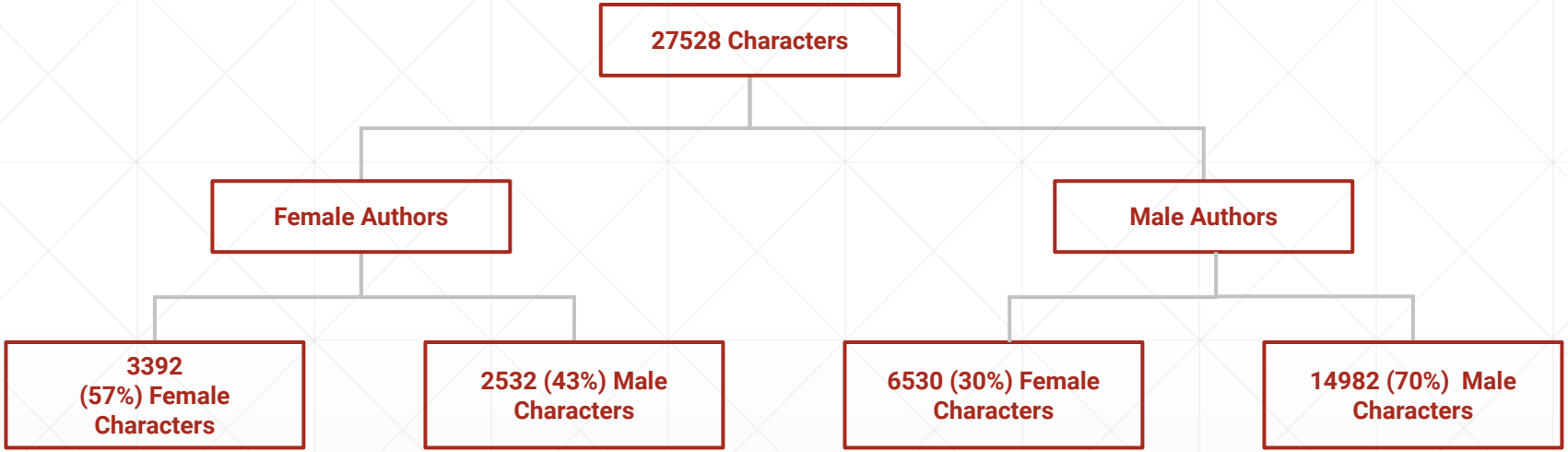
- Accuracy according to the 3 features extraction methods
  - BoW : 53%
  - TF-IDF : 66 %
  - Doc2Vec : 85 %



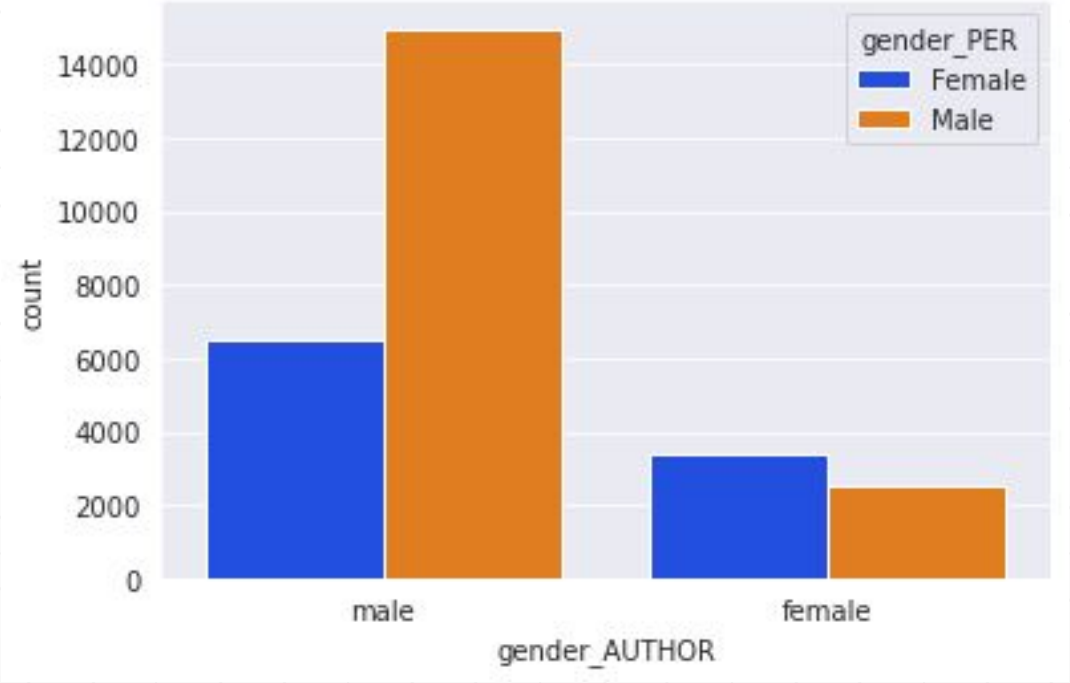
Results Statistics



Results statistics

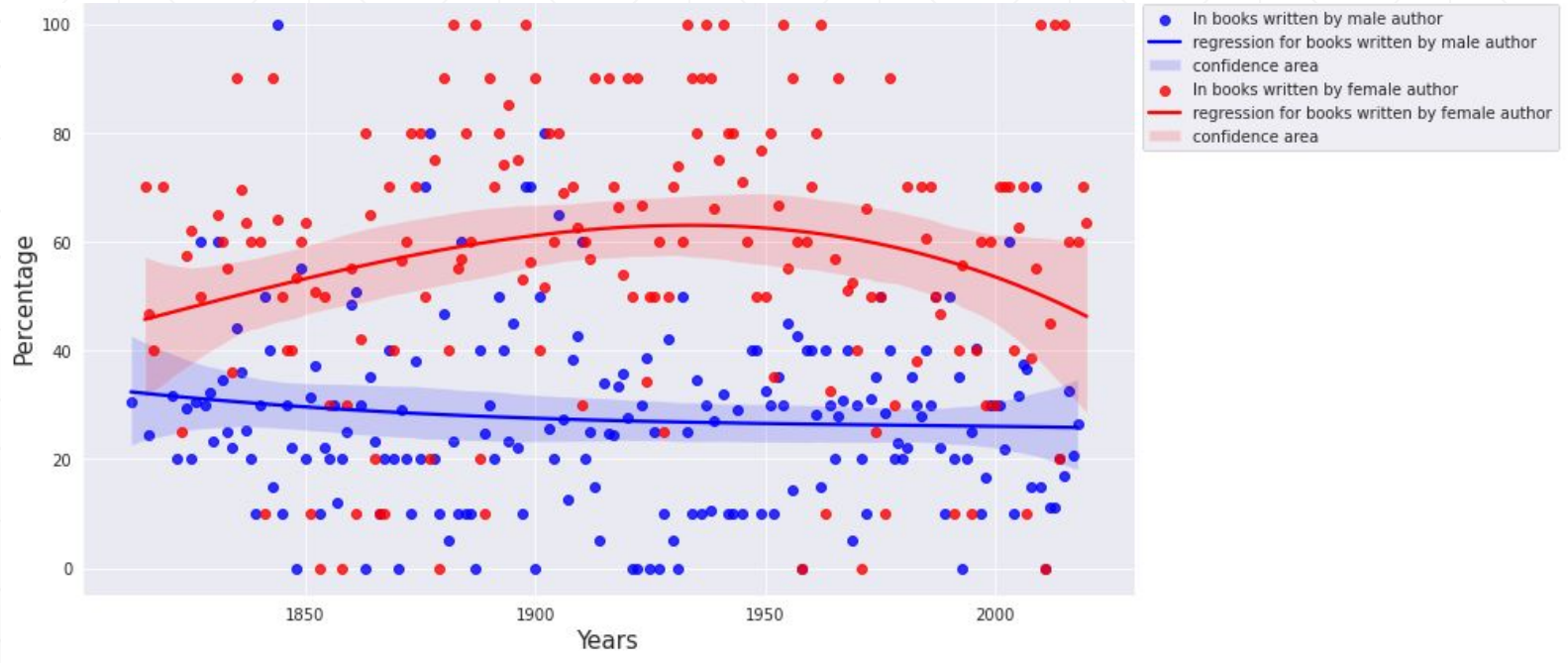


Results Statistics

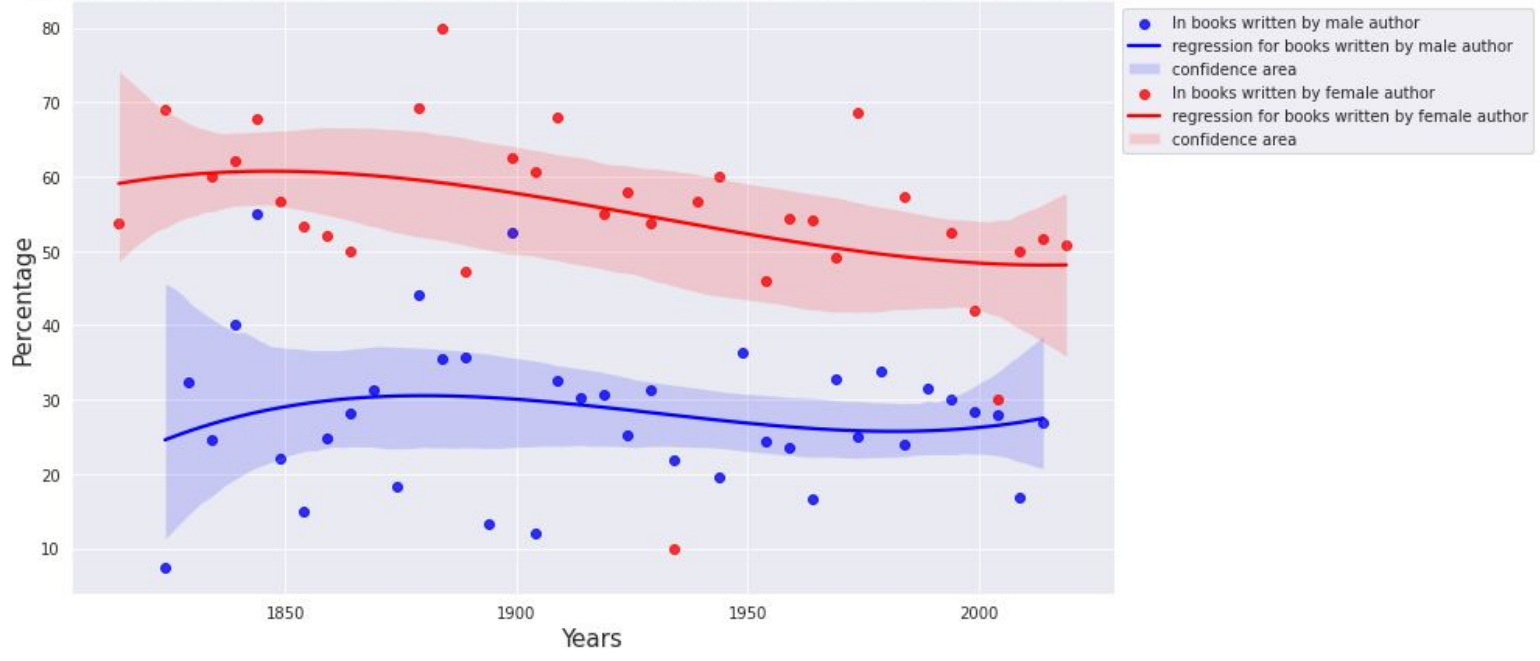


III. Visualizations and results analysis

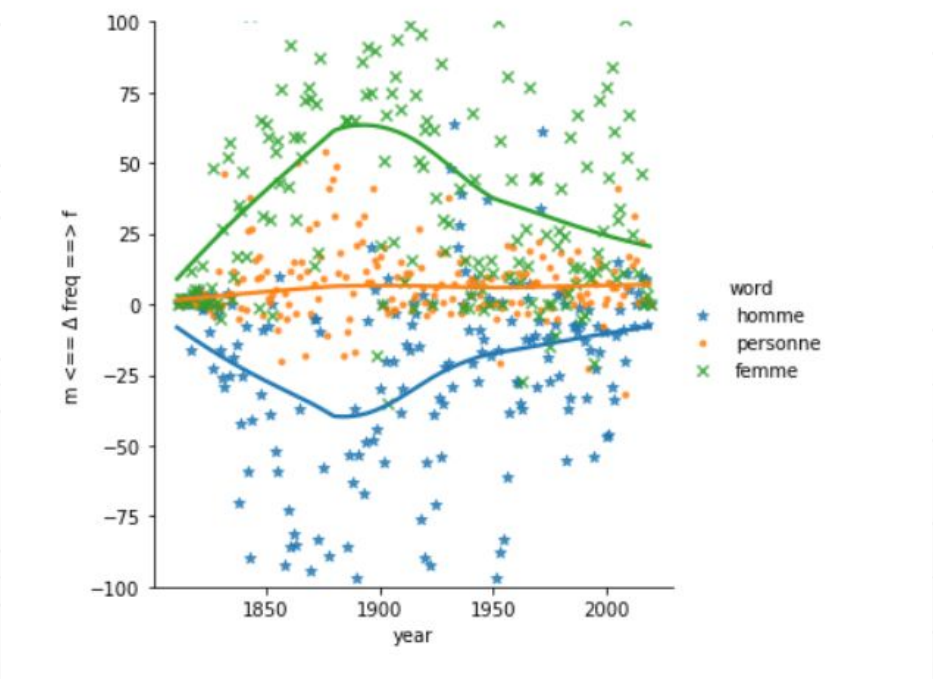
Proportion of characterization of women by women and men - for each year



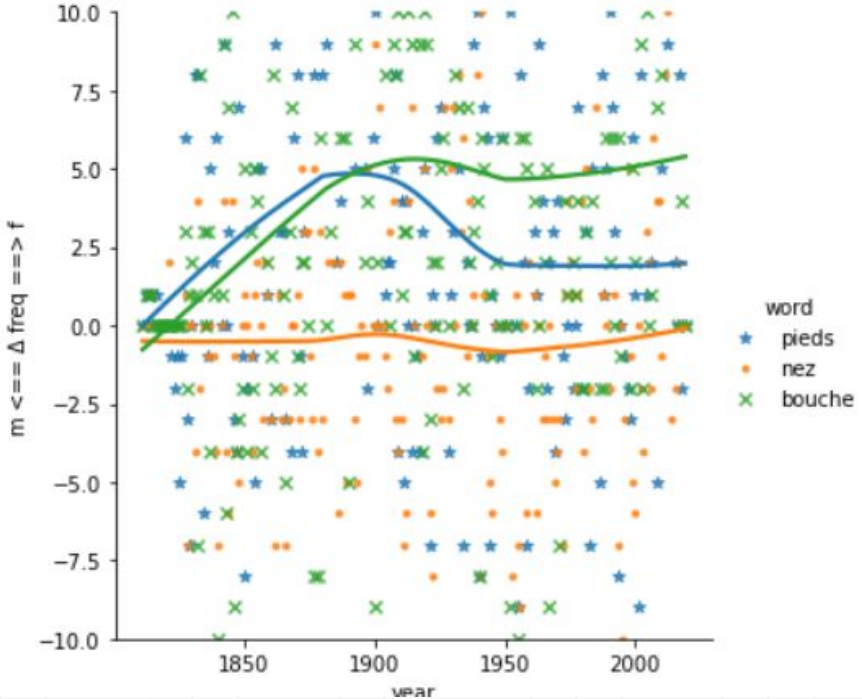
Proportion of characterization of women by women and men - for every 5 years

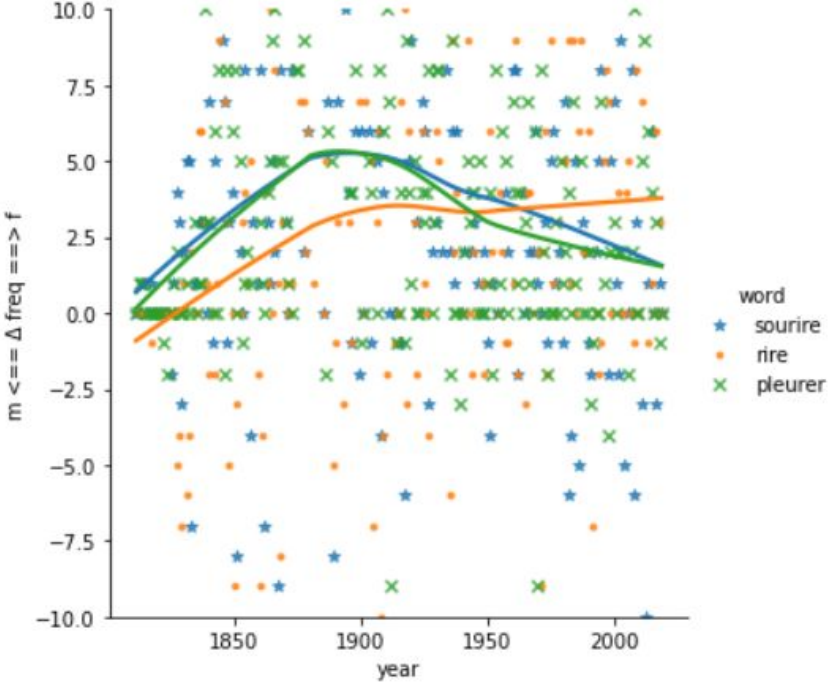


How men and women are characterized by obvious words : homme and femme

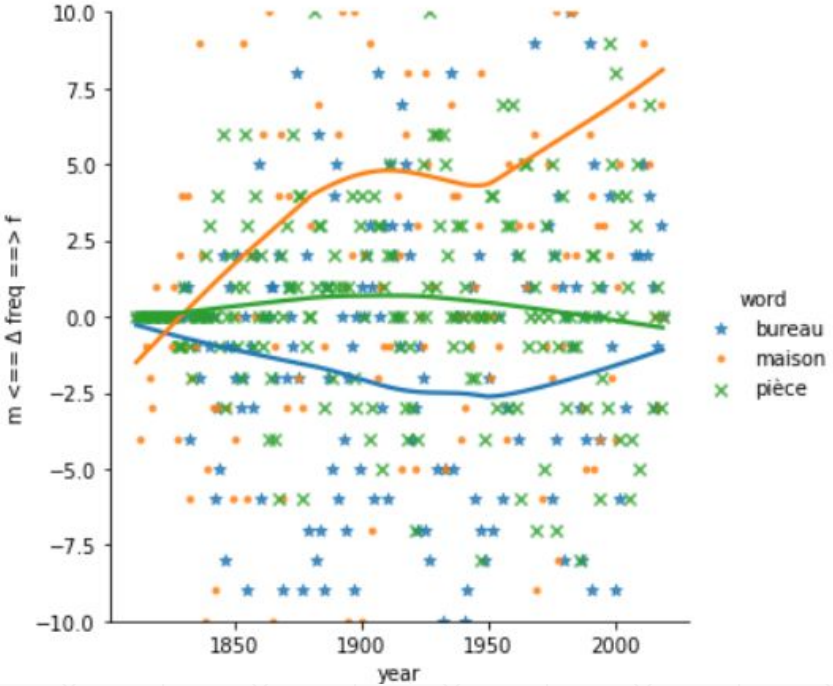


How men and women are characterized : body and emotions



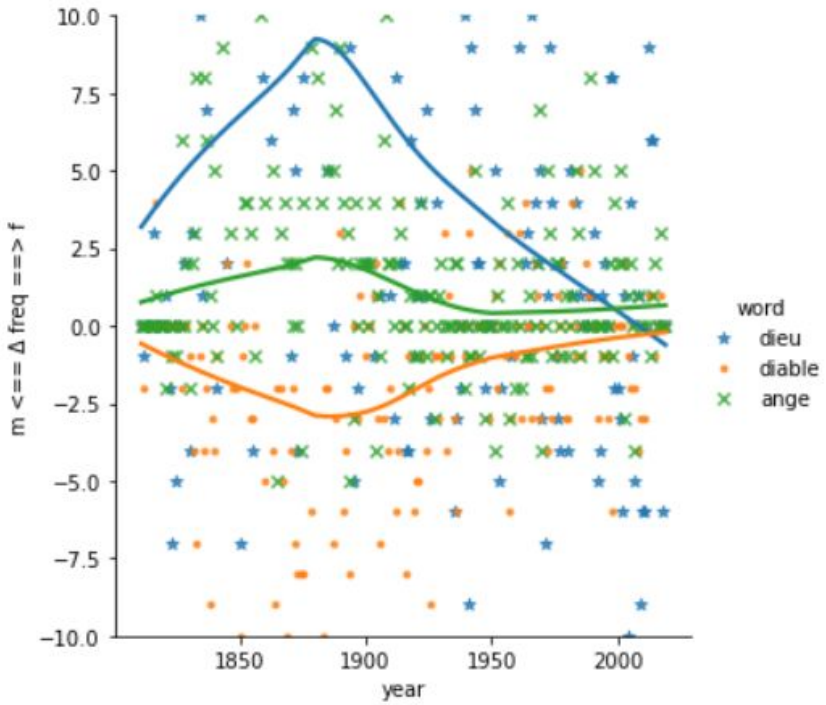


How men and women are characterized by places



- Stronger differentiation that challenges intuition and that is possibly explained by literary trends.
- Appearance of the auto-fiction as a genre from the 1950's and on.
- Writers like Annie Ernaux (*La Place*), Christine Angot (L'Inceste), Marie Darrieussecq (Le Mal de mer).
- Interest on the infancy of the female characters in the irruption of the female centered-coming of age novel.

How men and women are characterized by religious words



- Different periods with stronger attribution of certain lexical fields to a specific gender.
- Possibly explainable thanks to literary tropes.
- “Alors elle laissa retomber sa tête, croyant entendre dans les espaces le chant des harpes séraphiques et apercevoir en un ciel d’azur, sur un trône d’or, au milieu des saints tenant des palmes vertes, Dieu le père tout éclatant de majesté, et qui d’un signe faisait descendre vers la terre des anges aux ailes de flamme pour l’emporter dans leur bras”.

Madame Bovary, Flaubert

Conclusion

- We were able to evaluate how much literary characterization is related with gender stereotypes
- There are individual words / lexical fields related to gender stereotypes
- Proportion of characterization for female characters highly depends on the author’s gender
- Male authors write half less about female characters than female authors

Thank you!