PSL UNIVERSITÉ PARIS

A Computational History of Gender

Machine Learning & Literature

Jean Barré

21 novembre 2022

PSL Intensive Week DHAI

Outline

1. Introduction
    - Main Research Question
    - Paper takeaways

2. Main Task : Gender Prediction
    - Our Corpus
    - French BookNLP

3. Proposed sub-tasks

Introduction

Introduction

Computational Literary Studies

- Machine Learning & Text Mining to model concepts in large literary corpus.
- A key concept : Distant Reading - Franco Moretti.
- Our Project : Trace the history of gender roles in 19th century French language fiction

Main Research Question

What is at stake in the representation of gender in fiction over the last two hundred years?

- Evaluate the signs of gender writers use in producing characters
- Were fictive men quite different from fictive women?
- How strongly public signs of gender shaped characterization in general?

Paper takeaways

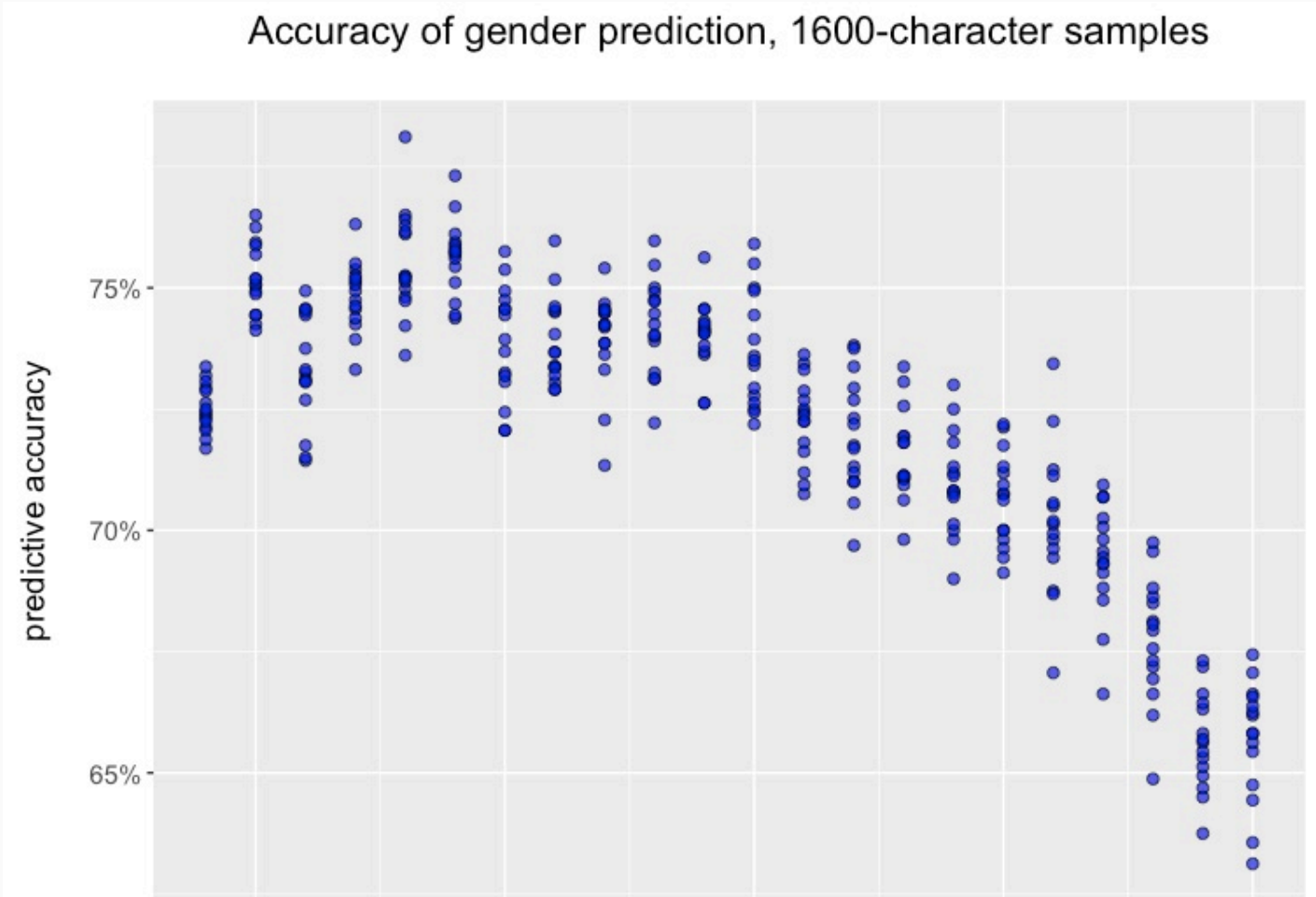
Lets try reproduce one of the following results

- Percentage of words used in novel for the characterization of female character decrease from 1800 to 1970
- Female author write about female character more and with larger percentage of words characterizing female characters
- A predictive model trained with word as features and female & male labels declines in accuracy from 1980s to nowadays
- Track individual words related to gender
- Screentime given to female characters is 3 times less important in case of a male author

Main Task : Gender Prediction

Main Task : Gender Prediction

- Prediction of gender based on adjectives, nouns, verbs that are characterizing the characters
- Relatively easy task (?), but what-if we exclude words that are explicitly gendered ?



Main Task : Gender Prediction

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

Corpus Chapitres

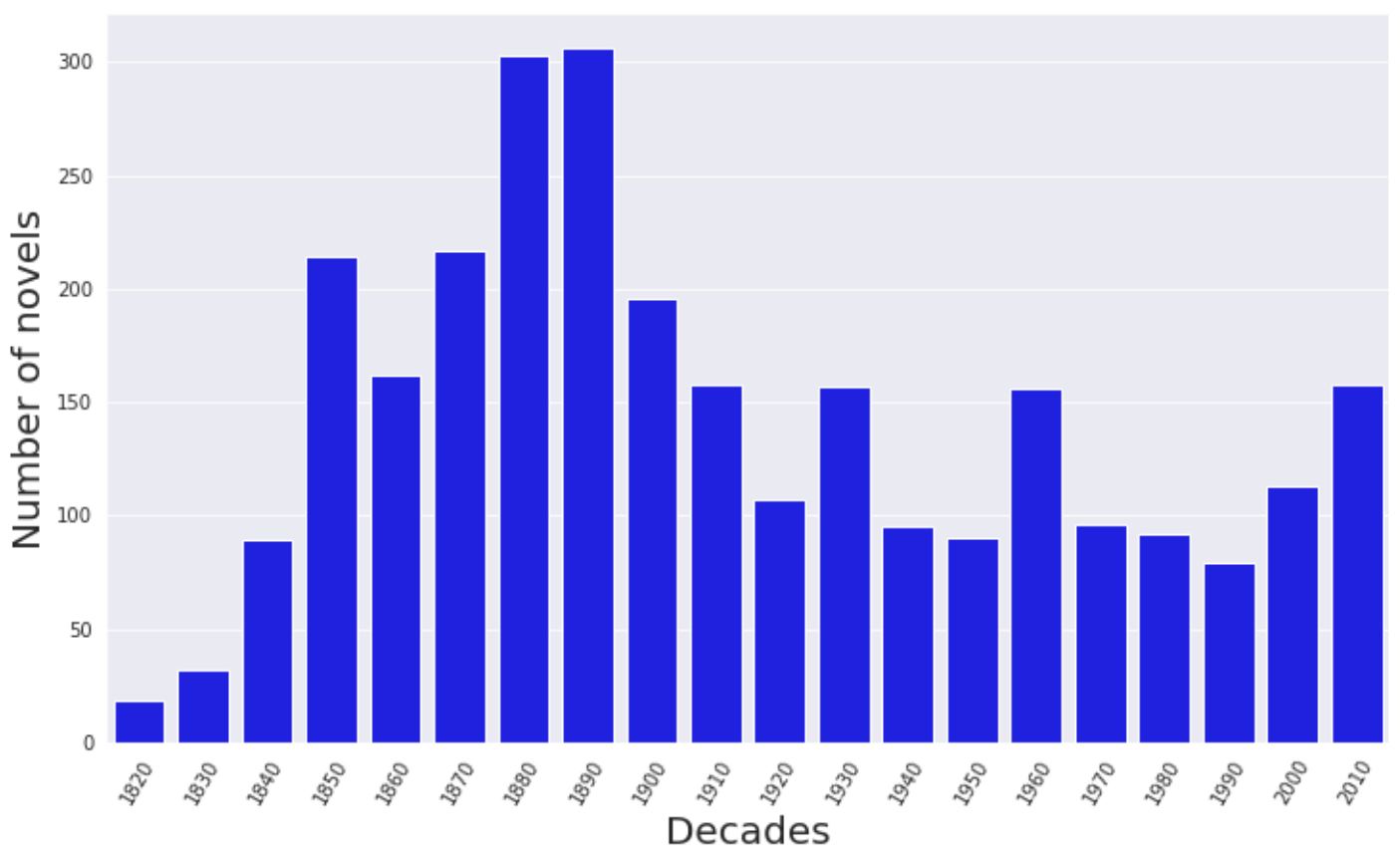


Figure 2 : Distribution of texts over time

Corpus Chapitres

Figure 3 : Percentage of fiction books written by women

French BookNLP

NLP pipeline scaling to books

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

Proposed sub-tasks

Proposed sub-tasks

- Lexical Investigation

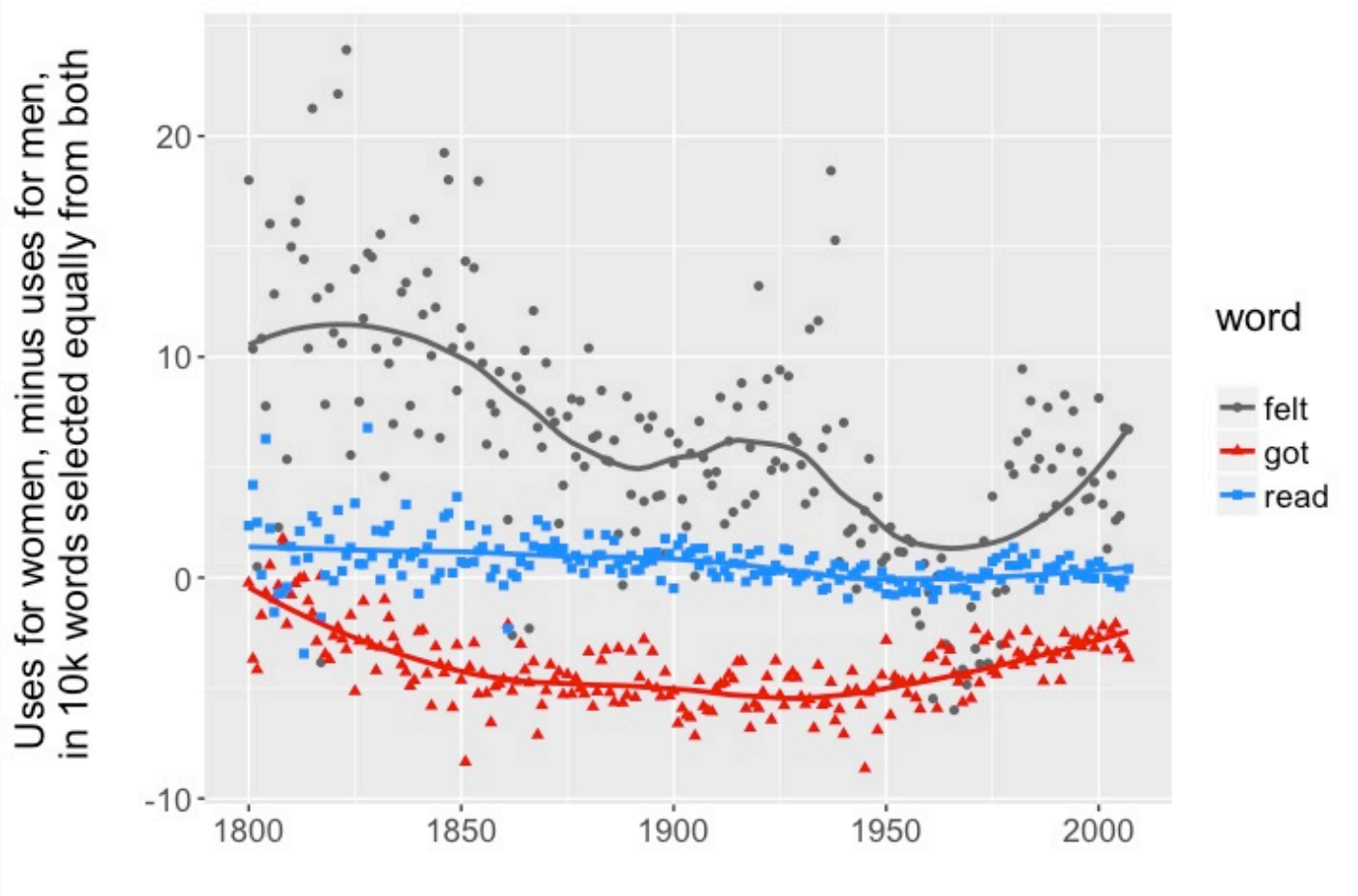


Figure 4 : Lexical Investigation

Proposed sub-tasks

- Topic Modeling

Topic 0

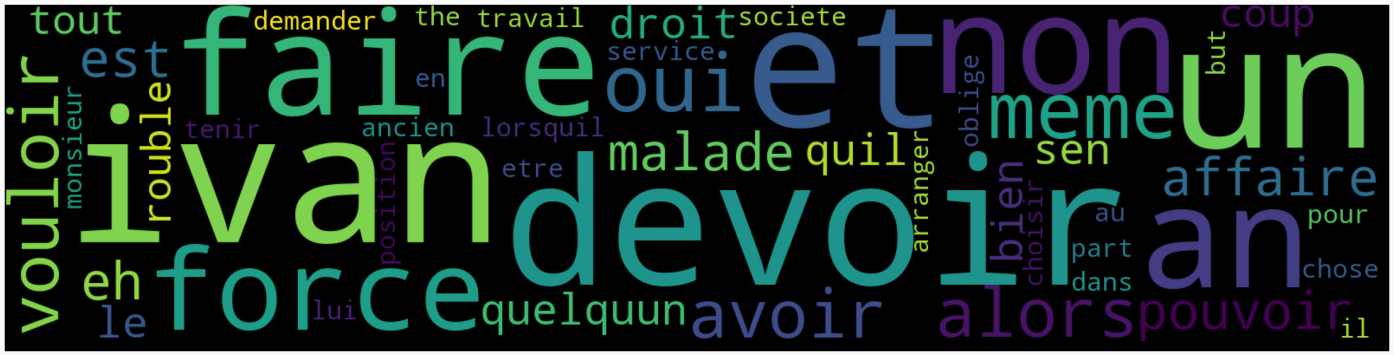
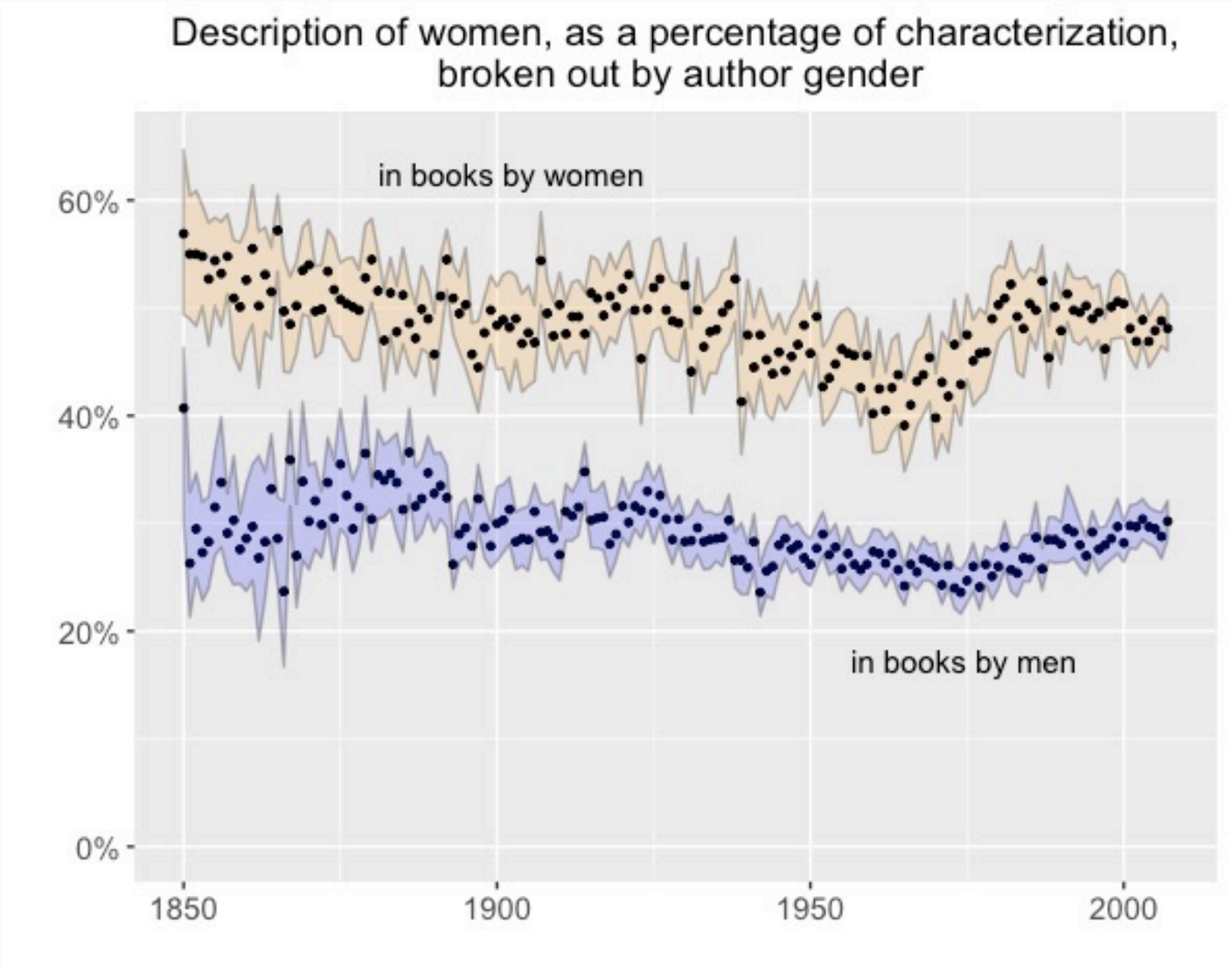


Figure 5 : Word-cloud of a topic w/ Bert-topics

Proposed sub-tasks

- Evaluate screen-time differentiation according to author's gender



Questions ?

You can find my github w/ slides, data & notebooks here :
https://github.com/crazyjeannot