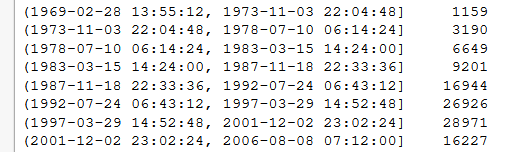
# Generic

nb lines: 114528 lines

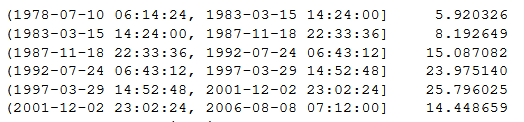
# Year of publication

## Lots of books published between 1983 and 2006



[…]

## Years intervals representing at least 5% of books in our dataset



# Languages

## Cleaning

Some languages were ***coded in several words*** :

* "Catalan; Valencian" => ca
* "English, Middle (1100-1500)" en
* "Multiple languages" en
* "zh-CN" zh
* "zh-TW" zh
* Replaced by their ***ISO 639 coding*** :

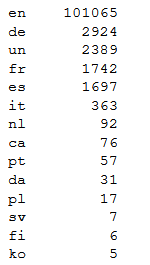


[…]

Then some languages are coded with ***words instead of ISO 639 coding***

* Replaced by their ISO 639 coding

## Mainly english books

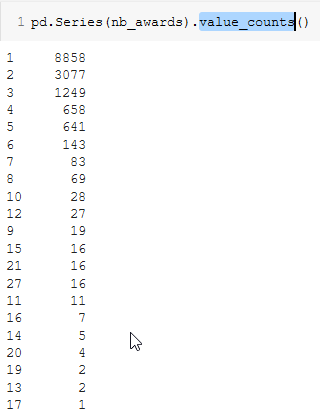


[…]

# Awards

## Books with lost of awards (before cleaning)

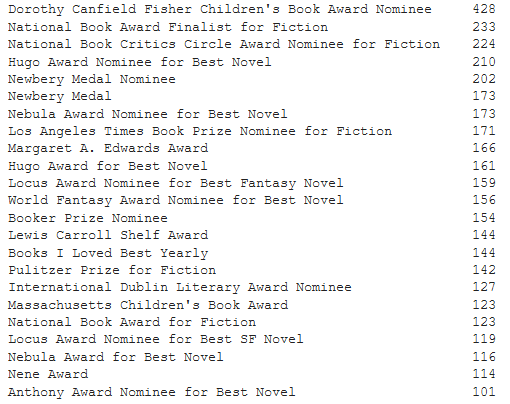
The maximum number of awards for a same book is 27 !



* Harry Potter and the Sorcerer's Stone (at indice 2809) !!!!

## Kind of awards (before cleaning)

* list saved in ***"awards\_names.csv"***



[…]

## Cleaning

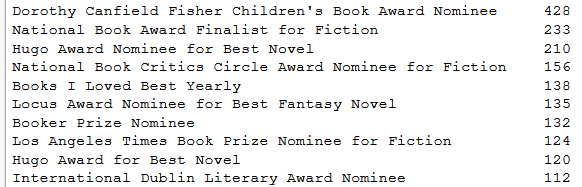
For each book:

* split according to ***","*** in order to obtain the list of awards
* then split according to ***"("*** in order to keep only the name (and not the date)

A list of all cleaned awards has been created in ***"awards\_names.csv"***

I ***kept only one award name per book***, the most famous awards in each cells (compared to the previously created list).

## Kind of awards (after clening)



[…]

* It remains 613 kind of awards.

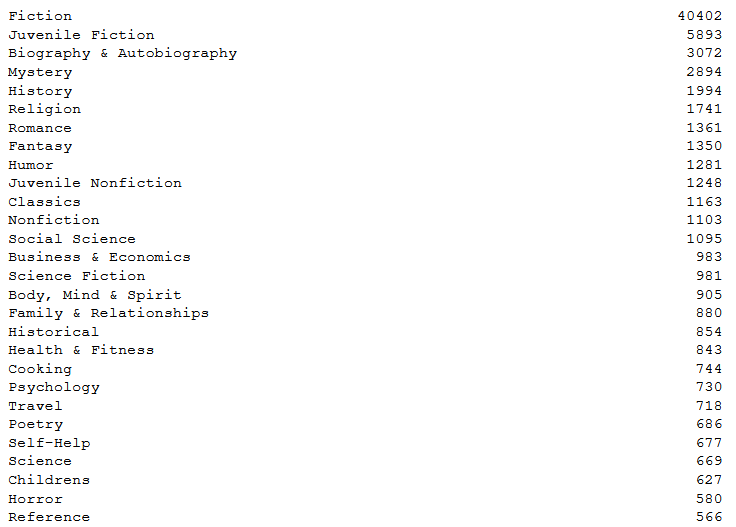
# Categories

## Before cleaning

5010 categories

* Categories with only one element saved in ***"cat\_one\_book.csv"***

### Lots of Fiction : 35.28 % of all books



[…]

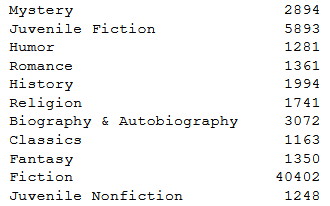
### Repartition of the number of elements insed categories : lots of categories with less than 80 books inside



[…]

### To keep 95% of the most important categories, we could keep the 3989 first categories (on 5010)

### Categories representing at least 1% of all the books



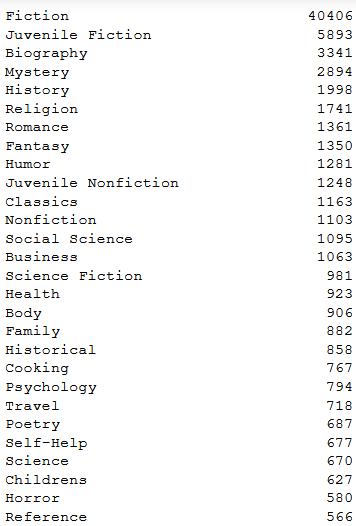
## Cleaning

The cleaning has been performed in several steps :

* 1st categories with ***"&"*** inside : a second column "Category\_other" has been created to take the 2nd term after the "&"
* Then the category words between ***"(…)"*** has been put inside "Category\_other" column, and the first term only remains on the "Category" column
* Then the category words with ***","*** has been splitted in two : the first term before the "," on "Category" column, and all the remaining termns in "Category\_other" column

## After cleaning & both columns

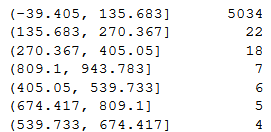
### Lots of Fiction : 35.28 % of all books



[…]

* Quite the same as previously

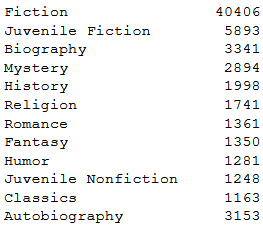
### Repartition of the number of elements insed unified categories : lots of categories with less than 140 books inside



[…]

### To keep 95% of the most important unified categories, we could keep the 4143 first categories (on 5113)

### Unified categories representing at least 1% of all the books



# Author genres

## Cleaning

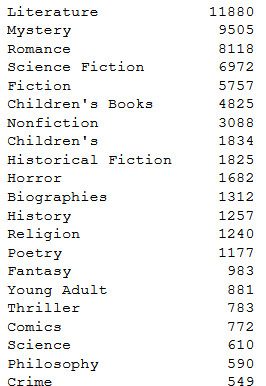
Cleaning (on Colab ) as multiple author genres in one cells, genres separated by ***&***

* 2 columns : "author\_genres" and "author\_genres\_other"

## author\_genres

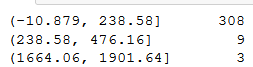
* author genres with only one element inside saved in ***"author\_gender\_one\_book.csv"***

### Lots of Literature as author gender



[…]

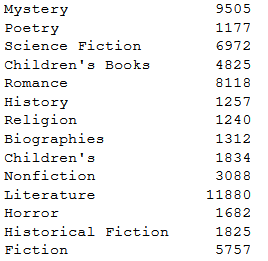
### Repartition of the number of elements insed author genres : lots of author genres with less than 240 books inside



[…]

### To keep 95% of the most important author genres, we could keep the 303 first author genres (on 338)

### author genres representing at least 1% of all the books

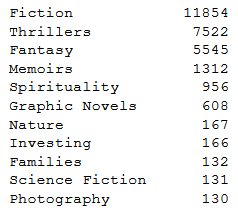


* Mystery, Romance, History, Religion, biography and fiction in common with categories.
* The same after cleaning categories.

## author\_genres\_other

* author genres with only one element inside saved in ***"author\_other\_gender\_one\_book.csv"***

### Lots of Fiction as author other gender



[…]

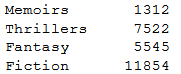
### Repartition of the number of elements insed author other genres : lots of author other genres with less than 120 books inside



[…]

### To keep 95% of the most important author other genres, we could keep the 34 first author other genres (on 39)

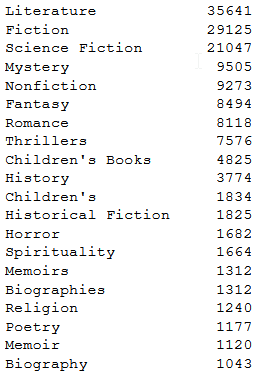
### author other genres representing at least 1% of all the books



## Both columns

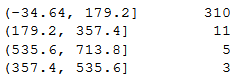
* author unified genres with only one element inside saved in ***"author\_unified\_gender\_one\_book.csv"***

### Lots of Literature as unified author genres



[…]

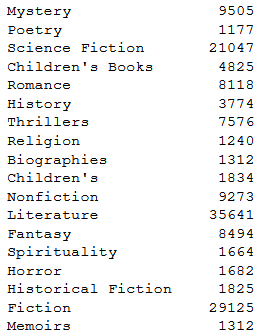
### Repartition of the number of elements insed unified author genres : lots of author genres with less than 180 books inside



[…]

### To keep 95% of the most important unified author genres, we could keep the 324 first author other genres (on 352)

### unified author genres representing at least 1% of all the books



* Thriller, Fantasy, Spirituality and Memoirs is appearing ***from the author\_genres column***
* Mystery, Romance, History, Religion, Biography, Fantasy, Fiction in common ***with categories***.
* The same after cleaning categories

# Duplicates

## Cleaning

***Duplicates have been removed***. Duplicates have been serached according to :

* ***ISBN*** comparison

*Remark : ISBN 13 values seem to be very often 9780000000000*

* *Column corrupted, no used then !*

Then ***title + author*** (exact comparison) have been compared. Duplicates have been :

* duplicates saved in ***"Books\_Duplicates.csv"***
* ***removed from "bothWebSites\_InternetSearch\_AllBooks\_BookCrossing\_cleaned.csv"***

# Category and author genre

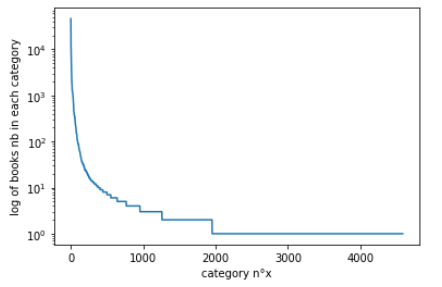
## Selection of the most famous categories

* Joined categories of
  + "Books\_Duplicates.csv"
  + + "bothWebSites\_InternetSearch\_AllBooks\_BookCrossing\_cleaned.csv" selected,
    - saved in ***"Sel\_cat\_autgenr.csv"***
      * Variable sel\_cat
* Joined categories of
  + "Books\_Duplicates.csv"
  + + "bothWebSites\_InternetSearch\_AllBooks\_BookCrossing\_cleaned.csv" NOT selected,
    - saved in ***"Other\_cat\_autgenr.csv"***
      * Variable other\_cat

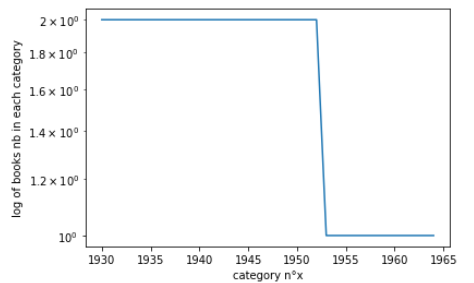
Joined categories of

* "Books\_Duplicates.csv"
* + "bothWebSites\_InternetSearch\_AllBooks\_BookCrossing\_cleaned.csv",

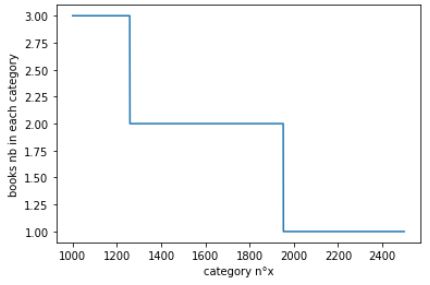
and the number of each categories are:



Zoomed view around the 2000th category :

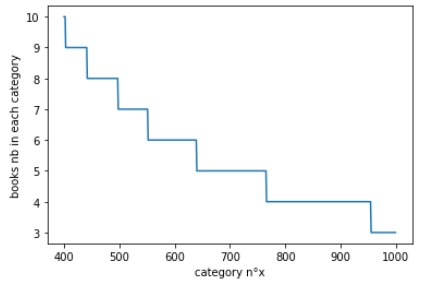


I could chose to keep only the first 1953ieth categories…



…but it corresponds to categories with only 2 books inside.

The first ***403*** categories have at least 10 books inside : it' my limit ;



## Cleaning an Creation of train data set

* Skip Gram model saved in ***"skipGram\_cat.model"***

FAILURE :

gensim.downloader.load('glove-wiki-gigaword-50')

and gensim.downloader.load('word2vec-google-news-300')

* NOK because our words categories are not inside

AND colab do not accept such huge data set loading !

I have joined content of columns (in "sent" variable) of "Books\_Duplicates.csv"and "bothWebSites\_InternetSearch\_AllBooks\_BookCrossing\_cleaned.csv" :

* "book\_description"
* "Category"
* "Category\_other"
* "author\_genres"
* "author\_genres\_other"

Cleaning of "book\_description" column by remowing

* ***stopwords***
* ***words of length < 2***
* ***'...'***

I used the last version of Word2Vec inside gensim.models :

* construction of vocabulary
* model training
  + Skip Gram
  + numerical vector of length 50

## Find the Category the most close to the selected categories

* Udpate of file ***"Other\_cat\_autgenr.csv"***
  + Variable other\_cat

I used the previous model to find for each categories inside other\_cat the closest category inside sel\_cat, with model.wv.most\_similar\_to\_given.

## Cleaning of Category and Author\_genre

I have ***updapted the Category and author\_genre columns of*** "bothWebSites\_InternetSearch\_AllBooks\_BookCrossing\_cleaned.csv" ONLY ***thanks to those previous w2v link***.

# Category\_other and Category\_other

There are 612 unique values inside Category\_other.

There are 38 unique values inside author\_genres\_other.

* No additional cleaning with word2vec performed, too few values.