Analysis of French theses since 1985

Bioinfo-fr

2023-07-23

Table of contents

Preface

1 Analysis of French PhD theses

1.1 Load data and libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import datetime
```

```
# Set the plotting style
plt.style.use("../bioinfo-fr.mplstyle")
```

```
# Load the dataframe
df = pd.read_csv(
    "../tmp/theses-soutenues-filtered.csv", quotechar='"', sep=",", header=0
)
df["date_soutenance"] = pd.to_datetime(df["date_soutenance"])
df.head()
```

```
/tmp/ipykernel_67155/3583511864.py:2: DtypeWarning: Columns (5,6,7) have mixed types. Specify
  df = pd.read_csv(
```

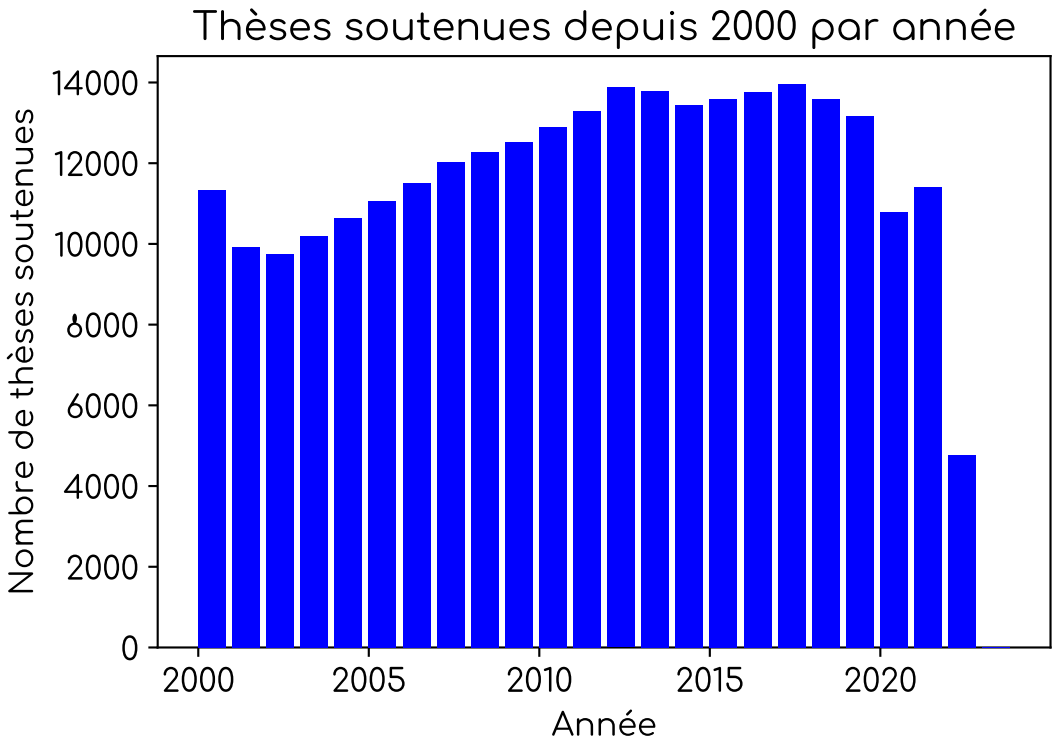
```
/home/sortion/.local/share/mambaforge/envs/theses-bioinfo-fr/lib/python3.11/site-
packages/IPython/core/formatters.py:344: FutureWarning: In future versions `DataFrame.to_late
  return method()
```

	auteurs.0.idref	auteurs.0.nom	auteurs.0.prenom	date_soutenance	directeurs_th
0	076645665	Wu	Tao	2003-01-01	
1	102611777	Simonin	Clémence	2011-11-09	
2	196371645	Poupon	Lenaic	2017-02-15	
3	251153770	Snider-Giovannone	Marie-Noëlle	2015-12-15	
4	156674697	Teixeira	Cédric	2011-11-21	

1.2 Distribution of Theses Defense Dates

```
# Select only theses defended after 1985
start_year = 2000
current_year = 2023  # For the dataset we have.
# Load precomputed dataframe
df_after = pd.read_csv("../tmp/year_distribution.csv")
```

```
plt.figure()
plt.bar(df_after["year"], df_after["count"], color="C0", zorder=3, align="edge")
plt.xlabel("Année")
plt.ylabel("Nombre de thèses soutenues")
plt.title("Thèses soutenues depuis 2000 par année")
plt.show()
```



1.3 Distribution of Theses by Discipline

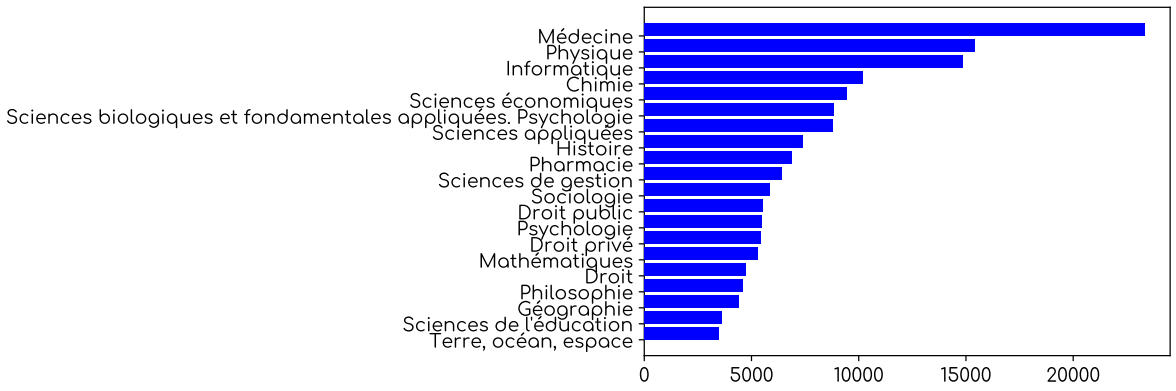
Code inspiré de https://github.com/richarddelome/theses_fr/

```
df_discipline = (
    df["discipline.fr"].explode().value_counts()[:20].sort_values(ascending=True)
)
df_discipline.head()
```

```
/home/sortion/.local/share/mambaforge/envs/theses-bioinfo-fr/lib/python3.11/site-
packages/IPython/core/formatters.py:344: FutureWarning: In future versions `DataFrame.to_late
  return method()
```

	discipline.fr
Terre, océan, espace	3461
Sciences de l'éducation	3594
Géographie	4391
Philosophie	4616
Droit	4732

```
plt.figure()
plt.barh(
    df_discipline.index,
    df_discipline.values,
    color="C0",
    zorder=3,
    align="edge",
)
plt.show()
```


Médecine
Physique
Informatique
Chimie
Sciences économiques
Sciences biologiques et fondamentales appliquées. Psychologie
Sciences appliquées
Histoire
Pharmacie
Sciences de gestion
Sociologie
Droit public
Psychologie
Droit privé
Mathématiques
Droit
Philosophie
Géographie
Sciences de l'éducation
Terre, océan, espace
0
5000
10000
15000
20000

2 Analysis of French PhD theses in Julia

2.1 Load packages

```
using CSV
using DataFrames
using Dates
using Plots
using StatsPlots
```

2.2 English prevalence in French PhD theses since 1985

```
df = CSV.read("../tmp/english_prevalence.csv", DataFrame)
```

	year	english	french	missing
	Int64	Int64	Int64	Int64
1	1985	1	1138	0
2	1986	5	3408	0
3	1987	4	4573	0
4	1988	4	6297	0
5	1989	2	6262	0
6	1990	7	6326	0
7	1991	1	6645	0
8	1992	7	7431	1
9	1993	10	7892	1
10	1994	15	8097	0
11	1995	14	6349	0
12	1996	16	6883	0
13	1997	20	7126	2
14	1998	15	6546	0
15	1999	30	6544	1
16	2000	31	6203	8
17	2001	81	5129	9
18	2002	93	5109	11
19	2003	134	5388	5
20	2004	151	5904	1
21	2005	256	5852	0
22	2006	327	6117	5
23	2007	410	7010	4
24	2008	690	7241	8
25	2009	887	7406	3
26	2010	1133	7837	5
27	2011	1418	9042	0
28	2012	1890	9592	0
29	2013	2356	9560	0
30	2014	1877	7360	0
…	…	…	…	…

```
# Get the english / (french + english) ratio on a new column
df.ratio = df.english ./ (df.french + df.english)
```

```
38-element Vector{Float64}:
 0.000877963125548727
 0.001464986815118664
 0.0008739348918505571
 0.0006348198698619266
```

```
0.00031928480204342275
0.0011053213327017212
0.0001504664459825459
0.0009411132024737832
0.0012655024044545685
0.0018491124260355029
0.0022002200220022
0.0023191766922742428
0.00279876854184159

0.1355640535372849
0.16460546943041282
0.19771735481705272
0.20320450362671863
0.206891817682592
0.20976545213158282
0.23020459159768858
0.2577067669172932
0.3180814770218744
0.3430755395683453
0.3850658857979502
0.46273291925465837
```

Plot:

```
plot(df.year, df.ratio, label="ratio", legend=:topleft, xlabel="Year", ylabel="English", t
```

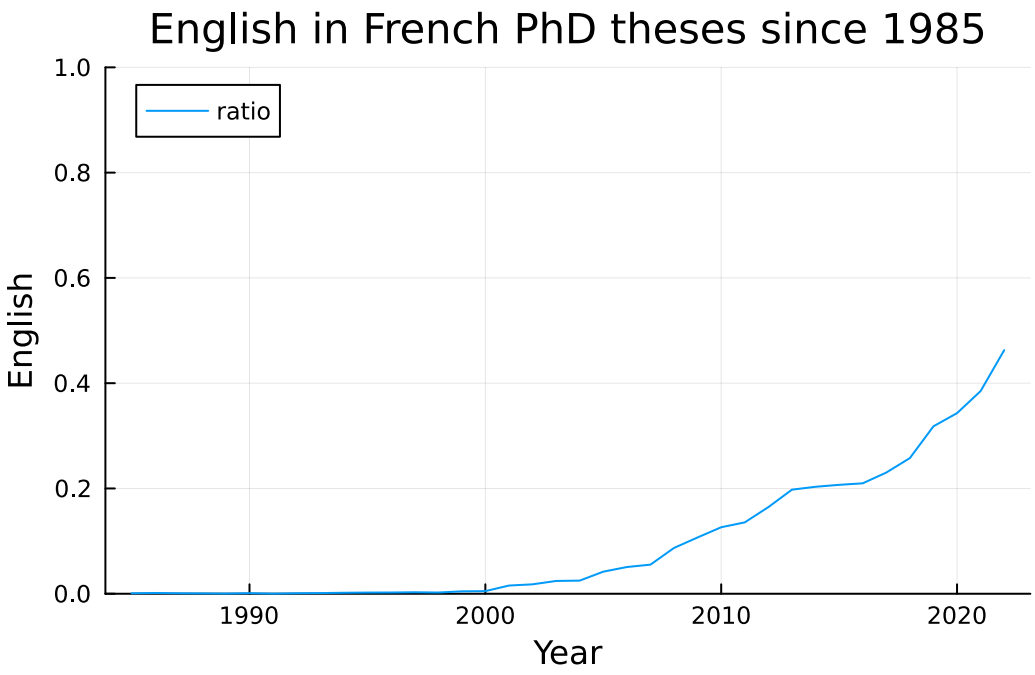


Figure 2.1: Rate of French PhD theses in English since 1985

Save:

```
savefig("../media/plots/english_prevalence.png")
```

```
"/home/sortion/Documents/Projects/bio/bioinfo-fr/bioinfo-theses-analyses/media/plots/english_
```

References