Analyse des *claims* dans les articles de Traitement Automatique des Langues à l'aide d'une méthode par apprentissage non supervisé

Fanny Ducel, M1 Langue et Informatique
Encadrée par Karën Fort et Maxime Amblard

12 septembre 2022

LETTRES SORBONNE UNIVERSITÉ

Loria
Laboratoire lorrain de recherche en informatique et ses applications

Motivations

Computer AI passes Turing test in 'world first' - BBC News [1]

TECHNOLOGIE. Des ordinateurs qui parlent notre langue [2]

- **Problème :** amplification injustifiée
- **Enjeux :** crédibilité et progrès scientifiques, méfiance du public
- **Contexte :** *publier ou périr*
- **Domaines :** éthique, TAL pour le TAL, identification de la subjectivité
- **Objectif :** Mesurer la force des *claims* dans les articles de TAL

[1] https://www.bbc.com/news/technology-27762088
[2] https://www.courrierinternational.com/article/2003/06/05/des-ordinateurs-qui-parlent-notre-langue

Notions linguistiques

Claim

Affirmation qui rend compte de quelque chose qui entraîne un effet ou un résultat [Blake, 2010]

> « *L'expérience montre que RCNN est très efficace pour améliorer l'état de l'art [...]* » [a]

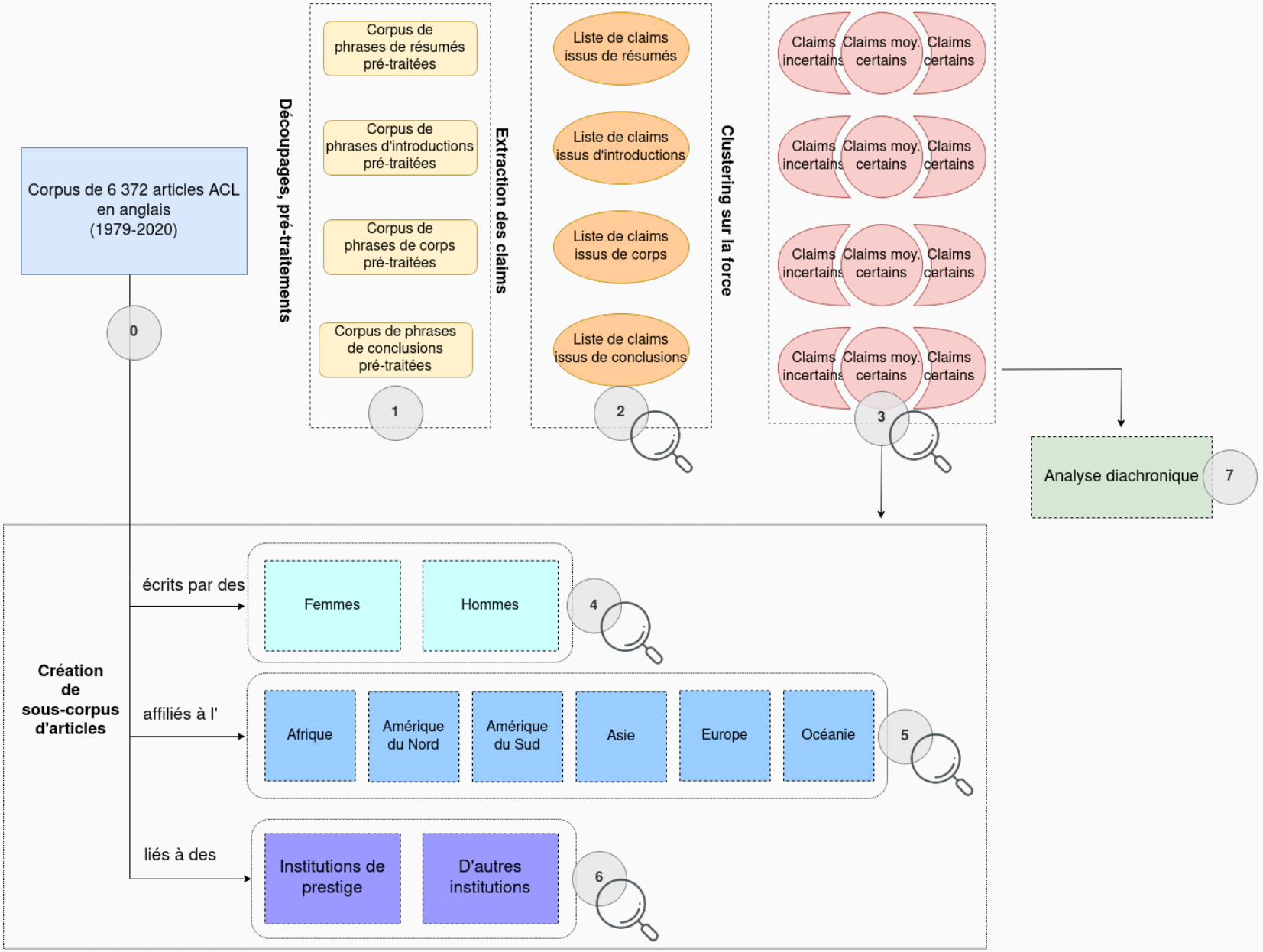
[a] Traduction d'une phrase de [Zhu et al., 2015]

Modalité épistémique

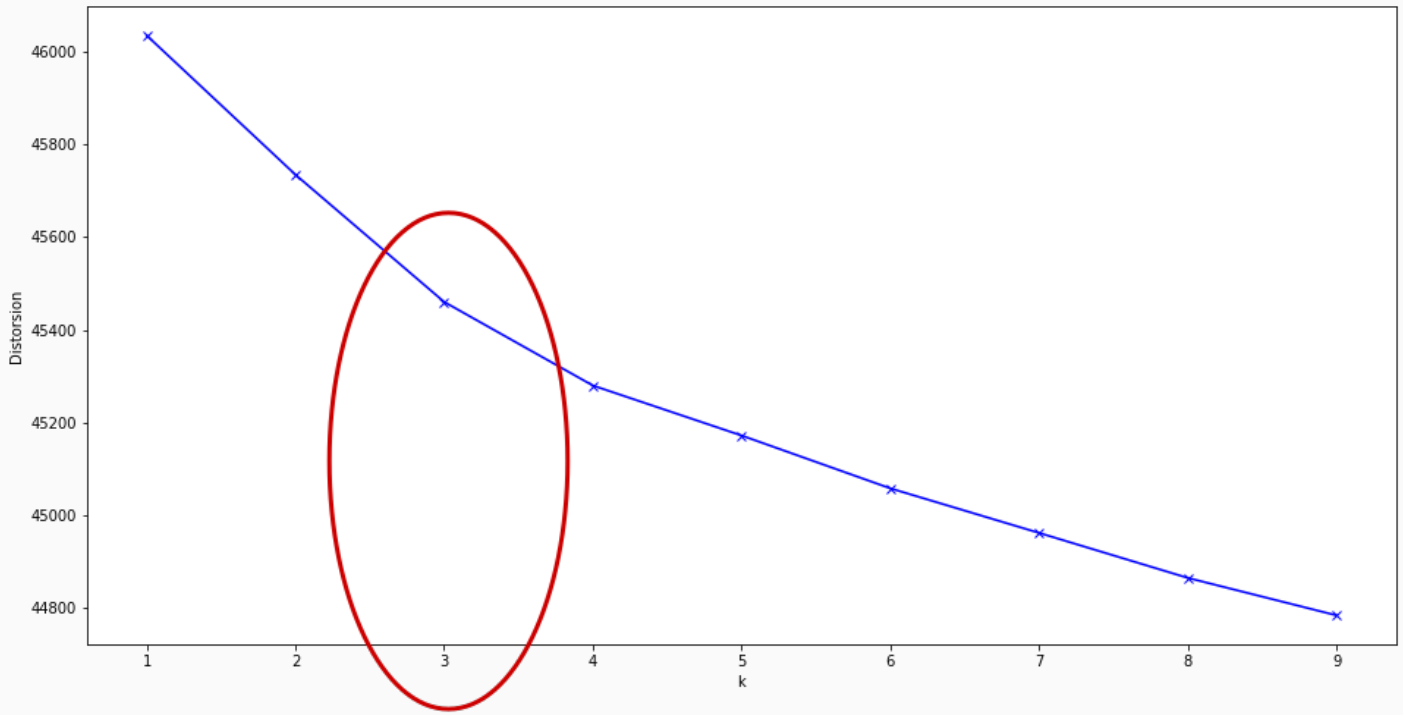
Expression linguistique de la probabilité, de la certitude [Rubin, 2006]

> *probable, évident, sembler, prouver, ...*

Étapes du travail



Déterminer le nombre de *clusters* par modalité épistémique



Méthode du coude [Syakur et al., 2018] sur les conclusions

Résultats des métriques d'évaluation du *clustering*

	Score de CH	Score de DB
Lem., stopwords	**142,26**	**8,83**
Lem., pas de stopwords	127,70	8,71
Pas de lem., stopwords	119,48	13,89
Pas de lem., pas de stopwords	100,22	10,37
Matrice aléatoire	*7,3*	*10,1*

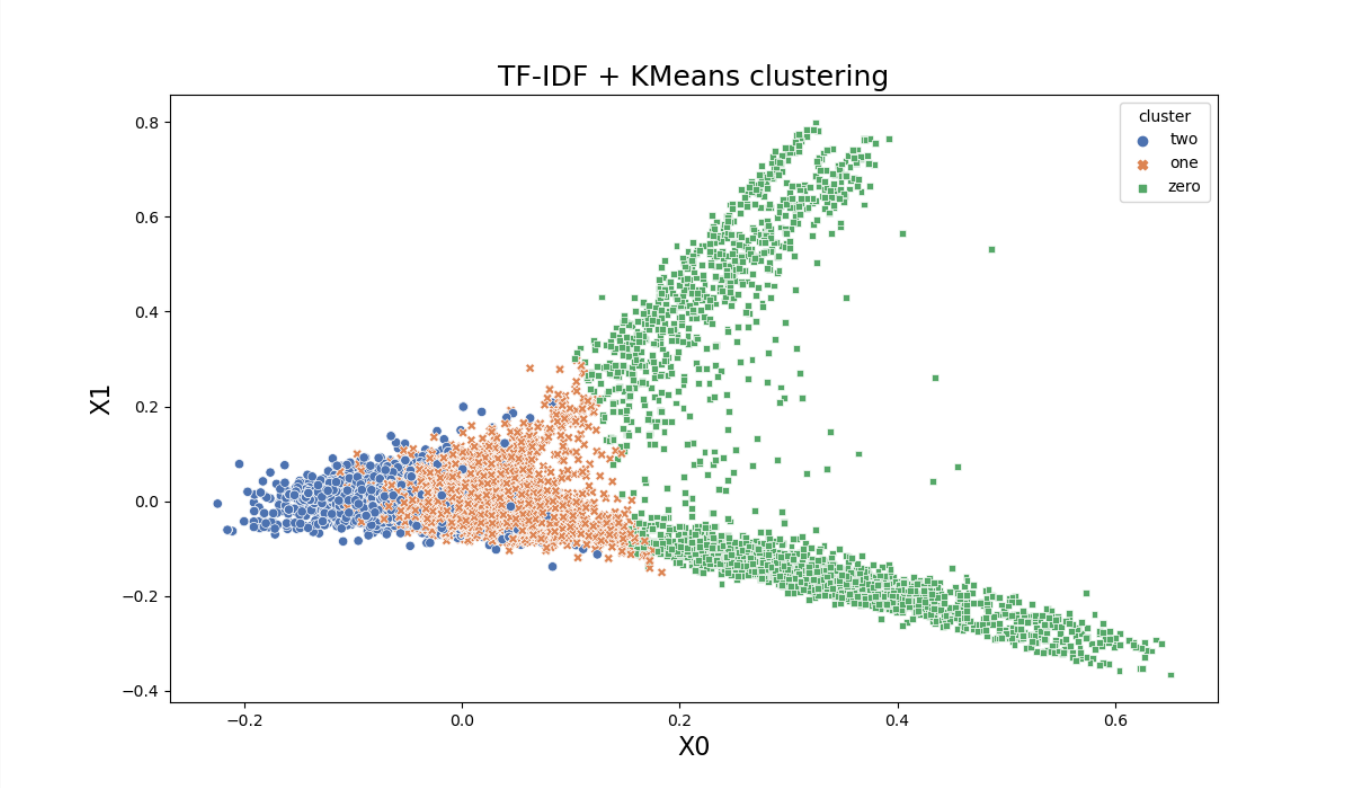
Résultats pour les *clusterings* sur les résumés

(CH : Calinski-Harabasz, le plus grand possible

DB : Davies-Bouldin, le plus petit possible)

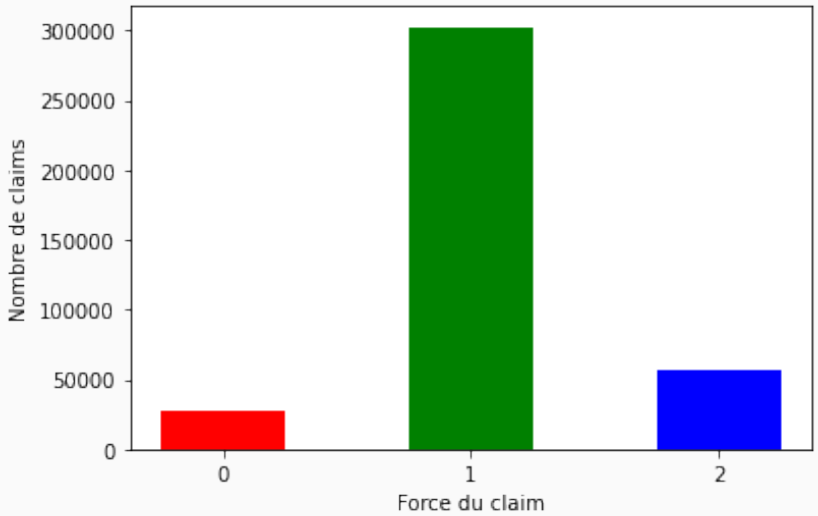
Exemple : *clustering* sur les conclusions

Degré 0	*papier, université, technologie, trouver, partiellement*
Degré 1	*soutenir, plus, prédire, peut, trouver, pourrait, doit, devrait*
Degré 2	*significatif, meilleur, améliorer, démontrer, surpasser, réaliser*

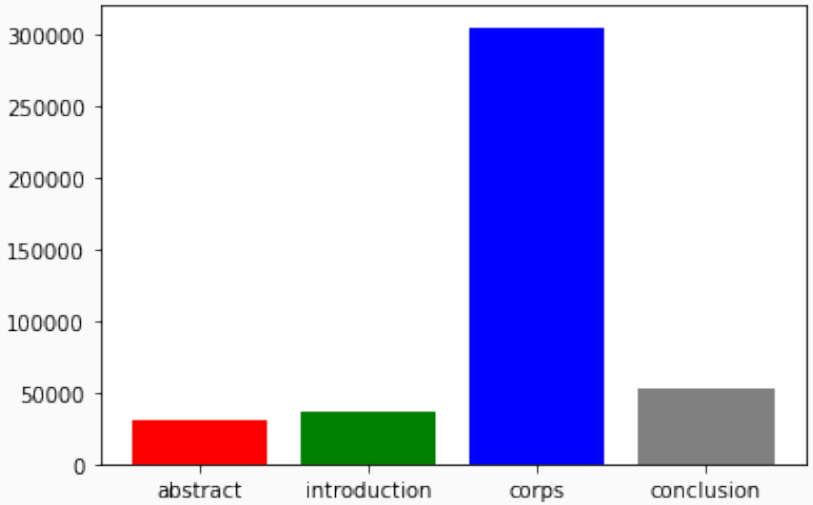


Claims selon le degré et la partie

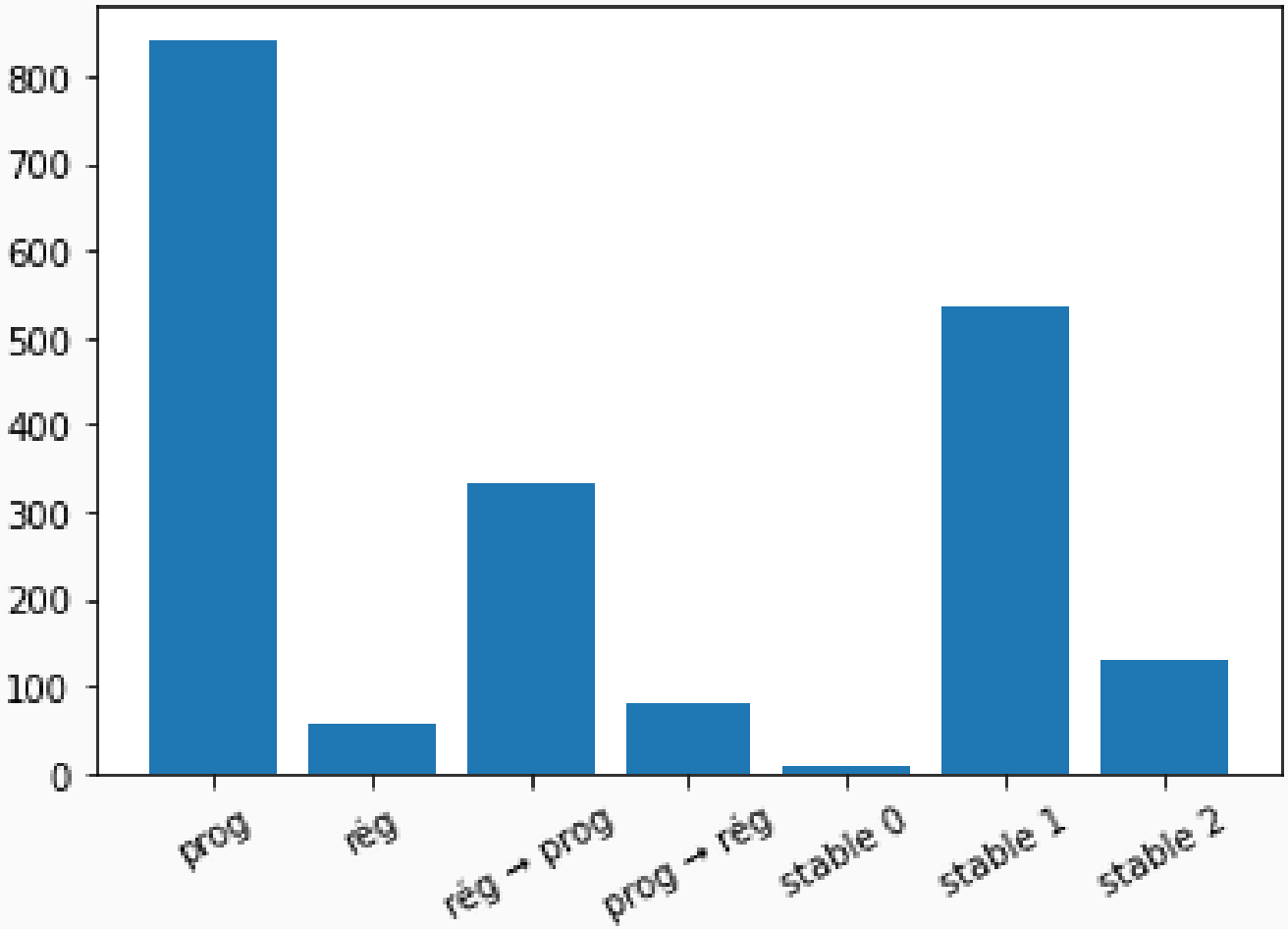
Nombre **absolu** de *claims* selon la modalité épistémique



Nombre **absolu** de *claims* selon la partie



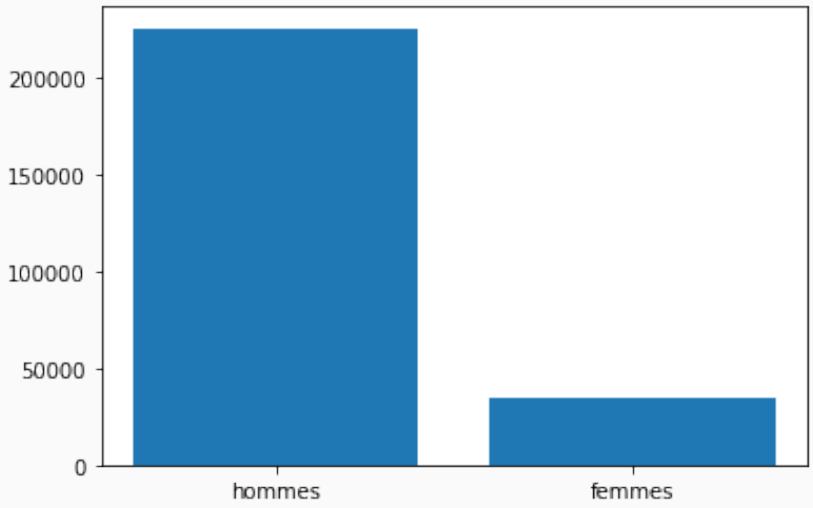
Progression intra-article



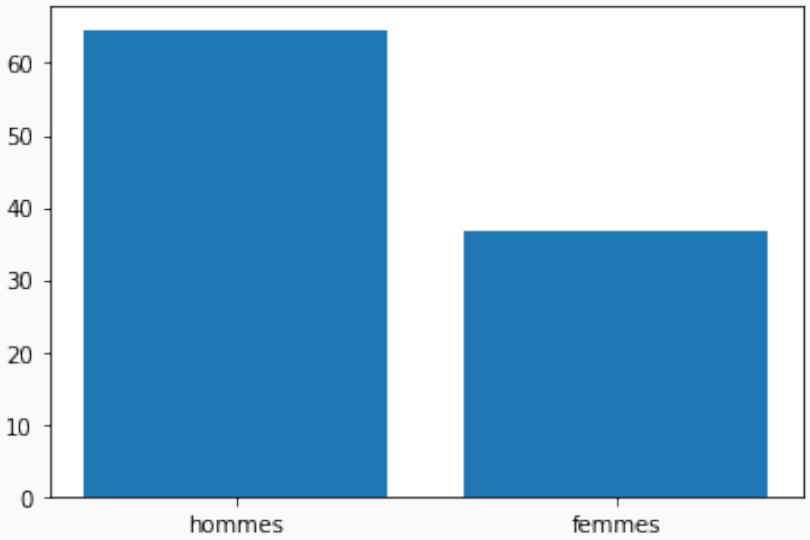
prog : progression, rég : régression, -> : puis

Corrélation avec le genre des scientifiques

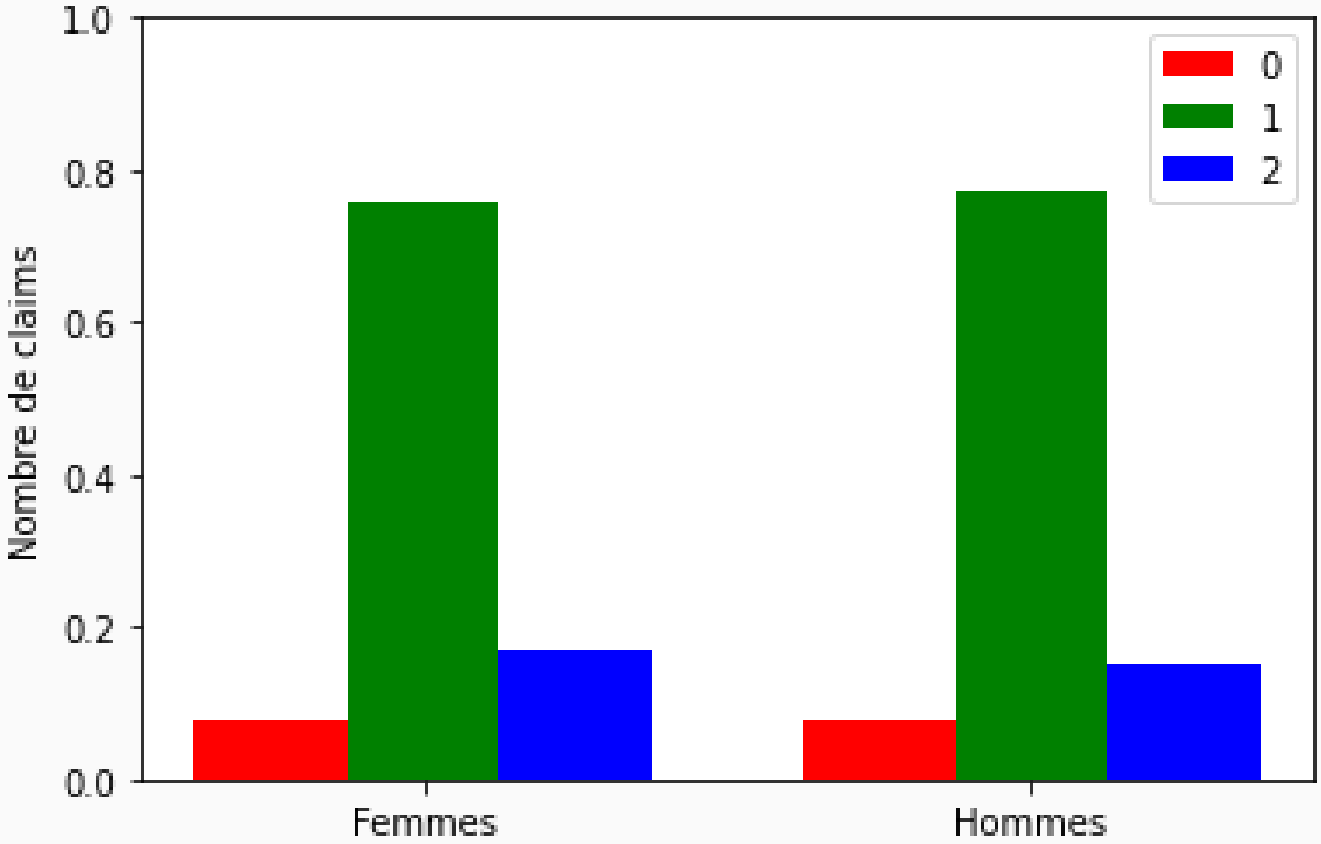
Nombre **absolu** de *claims* selon le genre



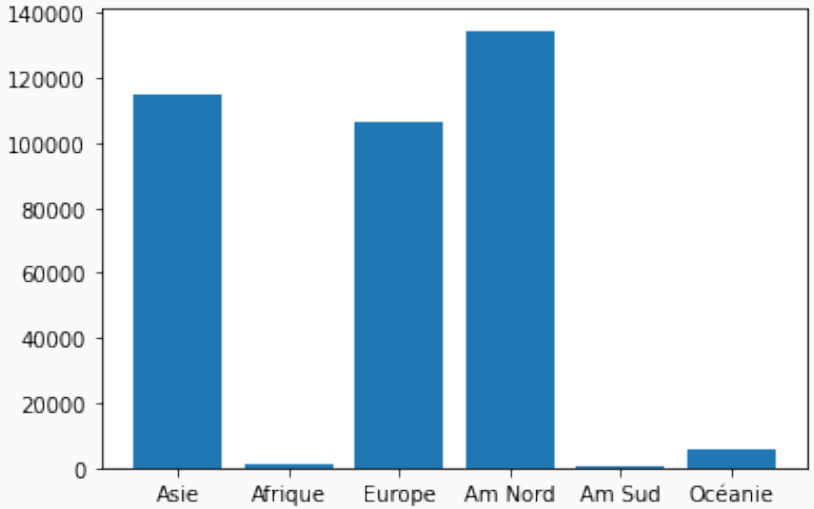
Nombre **moyen** de *claims* par article selon le genre



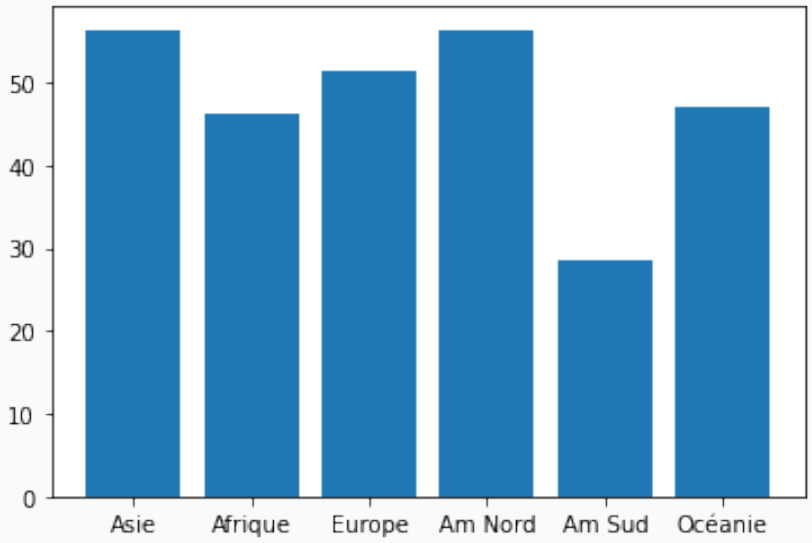
Une différence seulement quantitative



Corrélation avec le continent d'origine

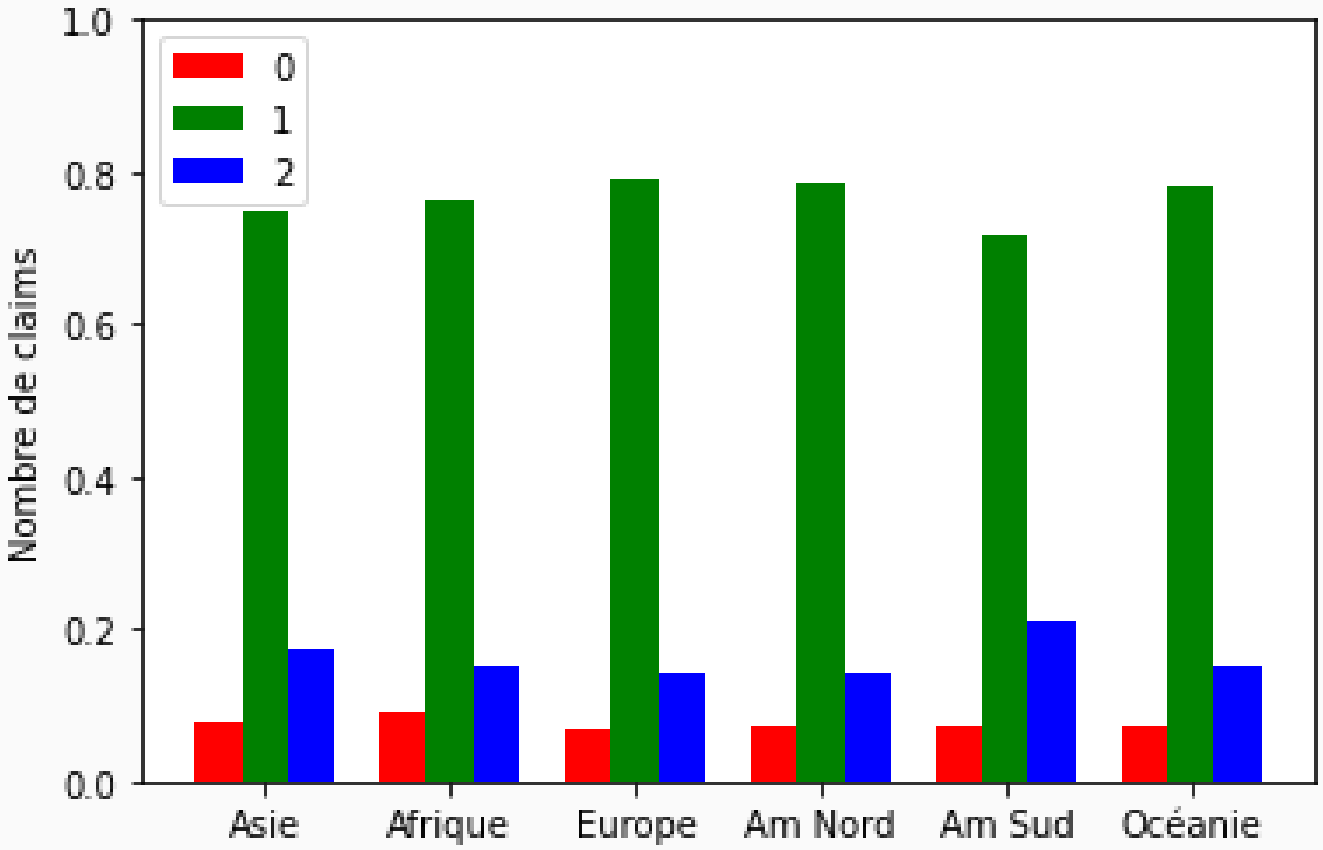


Nombre **absolu** de *claims* selon le continent majoritairement affilié



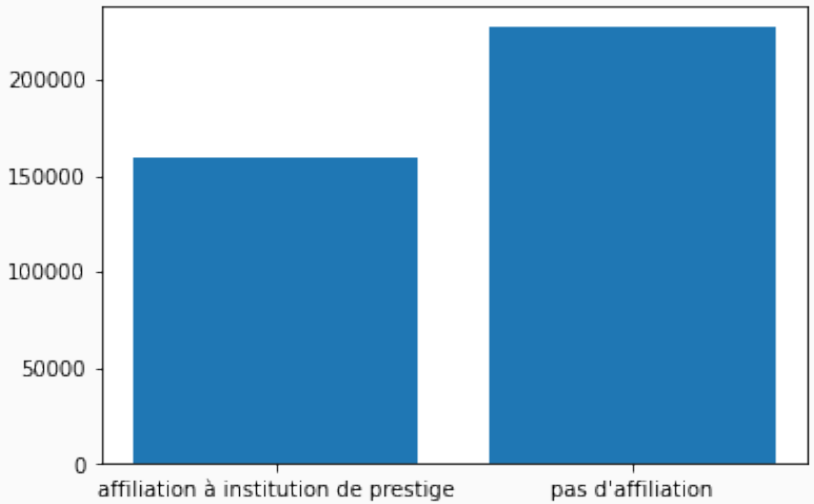
Nombre **moyen** de *claims* par article selon le continent majoritairement affilié

Une différence (encore) seulement quantitative

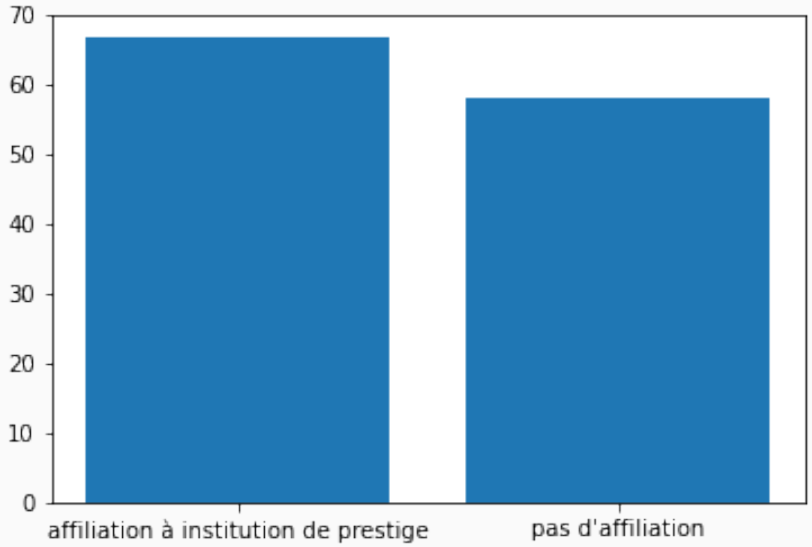


Corrélations avec des institutions de prestige

Nombre **absolu** de *claims* selon les affiliations à des institutions de prestige



Nombre **moyen** de *claims* par article selon les affiliations à des institutions de prestige



Conclusions

Contributions

- Le plus grand corpus utilisé pour étude sur les *claims* (6 372 articles, +216 millions de caractères, +30 millions de tokens)
- 1er sur le TAL

A retenir

- Majorité de degré 1, dans corps et conclusions, progression ⇧
- Différences quantitatives selon genre, continent, affiliations
- /!/ Corrélation ≠ causalité : variables cachées

Perspectives

- Affiner découpage de la structure (corps)
- Annotations manuelles
- Croiser les corrélations
- *Spin* (décalage *claims* et valeur des résultats)
- Système pour attirer l'attention sur les *claims*
- Autres corpus

Merci de votre attention !

https://github.com/FannyDucel/MemoireM1Claims

Annexes

Annexes

- Liste de *hedges*
- Institutions de prestige
- Analyse par genres
- Détails sur la corrélation avec les institutions de prestige
- Répartition selon les parties
- Diachronie
- Moyenne de *claims* par partie

Liste de *hedges*

find, can, could, may, might, must, should, claim, admit, discover, suggest, predict, prove, show, explain, infer, conclude, demonstrate, lead, succeed, intend, indicate, assume, appear, seem, favor, think, believe, analyze, examine, report, dare, point out, note, assert, declare, remark, comment, observe, reveal, disclose, confirm, convince

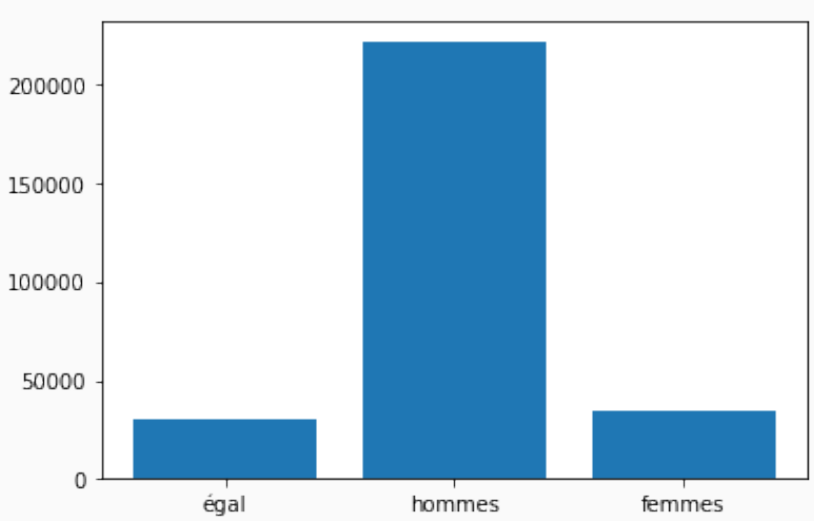
Algorithme de Boyer-Moore-Horspool

Liste d'institutions de prestige

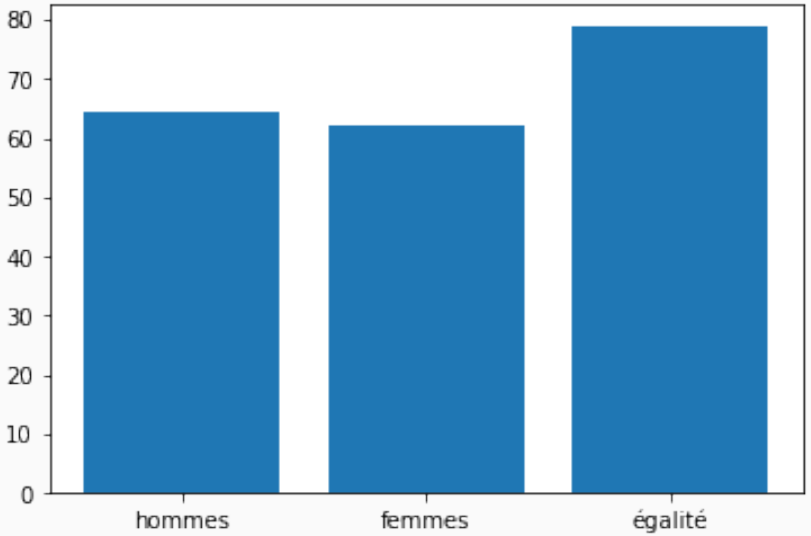
Google, Amazon, Facebook, Apple, Microsoft, Harvard University, Stanford University, University of Cambridge, Massachusetts Institute of Technology, MIT, University of California, Princeton University, University of Oxford, Columbia University, California Institute of Technology, Caltech, University of Chicago, Yale University, Cornell University, Université Paris-Saclay, University of California, University of Pennsylvania, Johns Hopkins University, University College London, University of Washington, Swiss Federal Institute of Technology Zurich, ETHZ, University of Toronto, Washington University, University of Tokyo, Imperial College London, University of Michigan, New York University, Tsinghua University, University of North Carolina, University of Copenhagen, University of Wisconsin-Madison, Duke University, University of Melbourne, Northwestern University, University of Manchester, Sorbonne University, Sorbonne Université, Kyoto University, University of Edinburgh, PSL Research University Paris,Université PSL,Université Paris Sciences Lettres, University of Minnesota, University of Texas at Austin, University of British Columbia, Karolinska Institute, Rockefeller University, Peking University, University of Colorado at Boulder, King's College London, University of München, University of Texas Southwestern Medical Center at Dallas, Utrecht University.

Figures pour le genre majoritaire

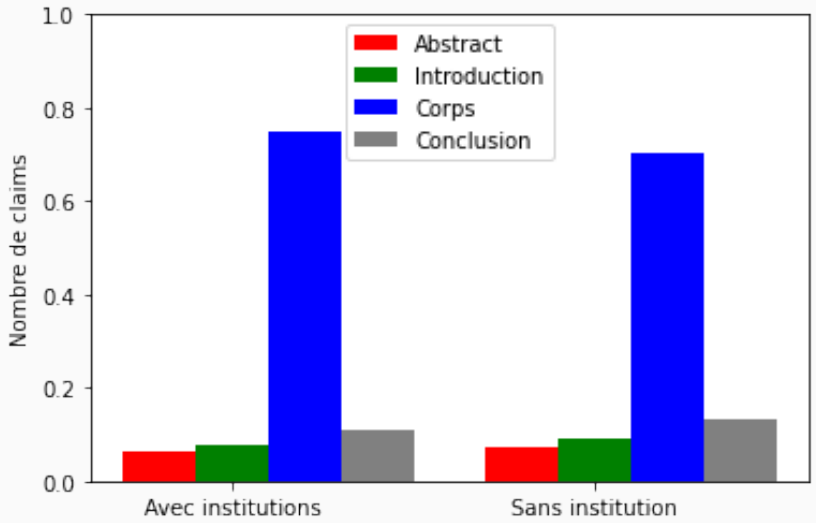
Nombre **absolu** de *claims* selon le genre

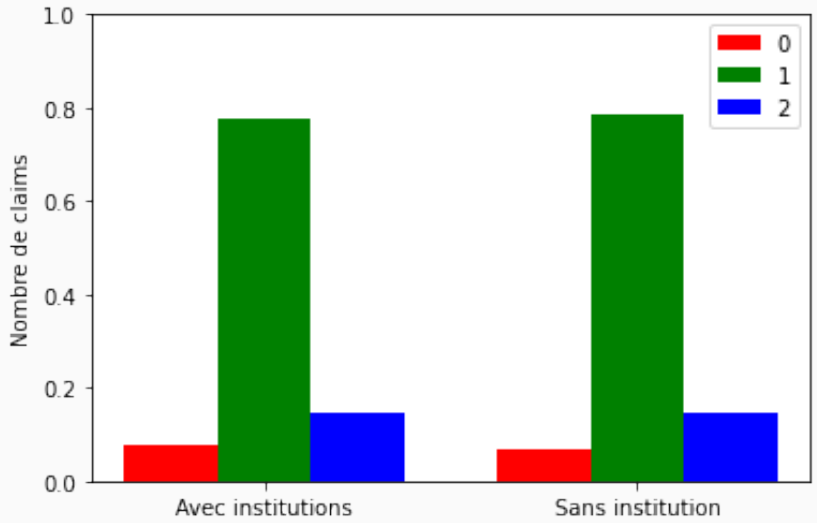


Nombre **moyen** de *claims* par article selon le genre

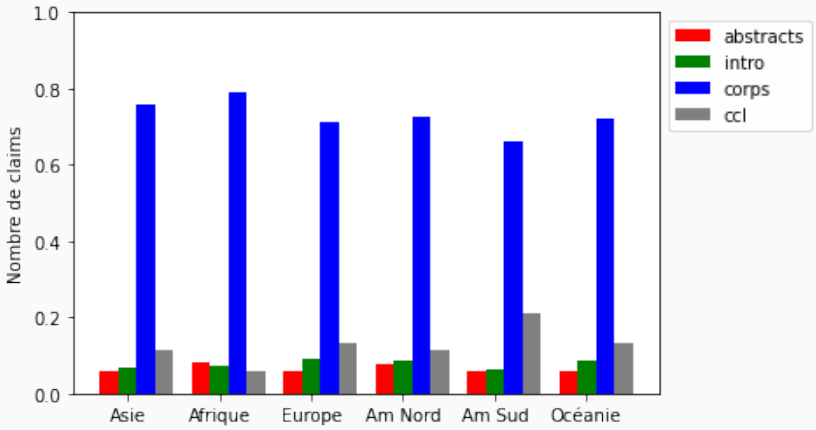


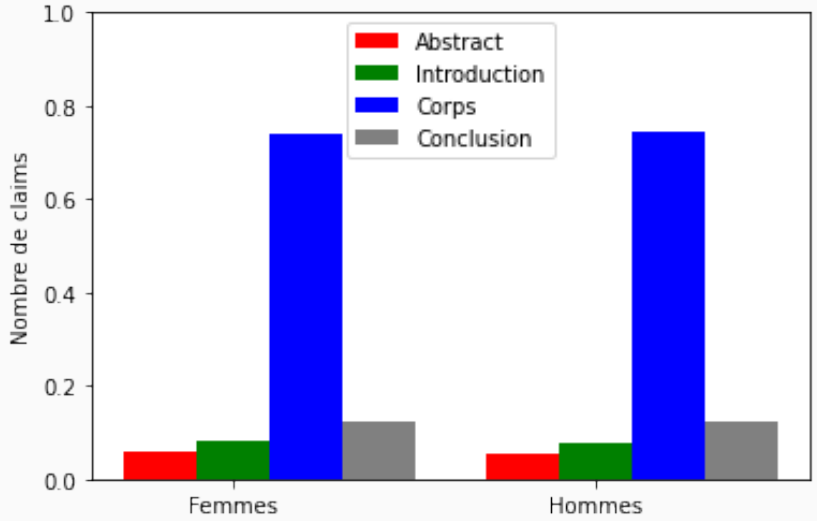
Figures pour la modalité épistémique et la partie des institutions



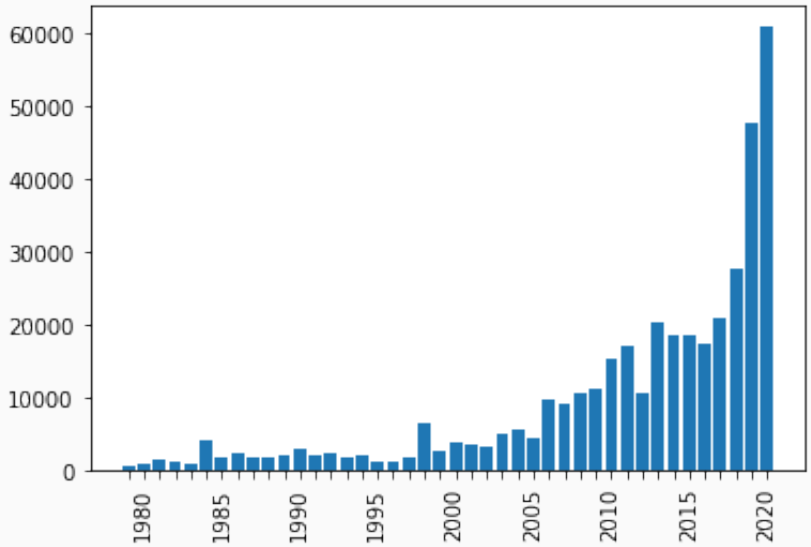


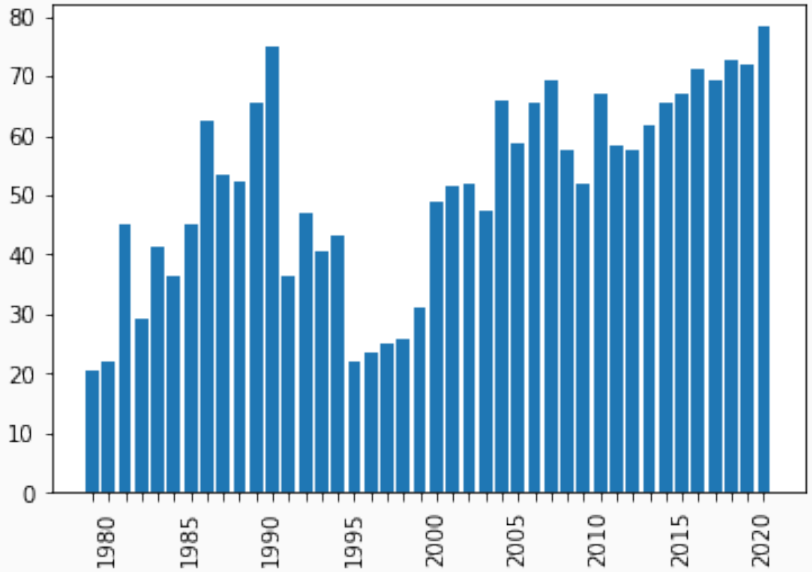
Figures pour la répartition selon les parties (genre et continent)



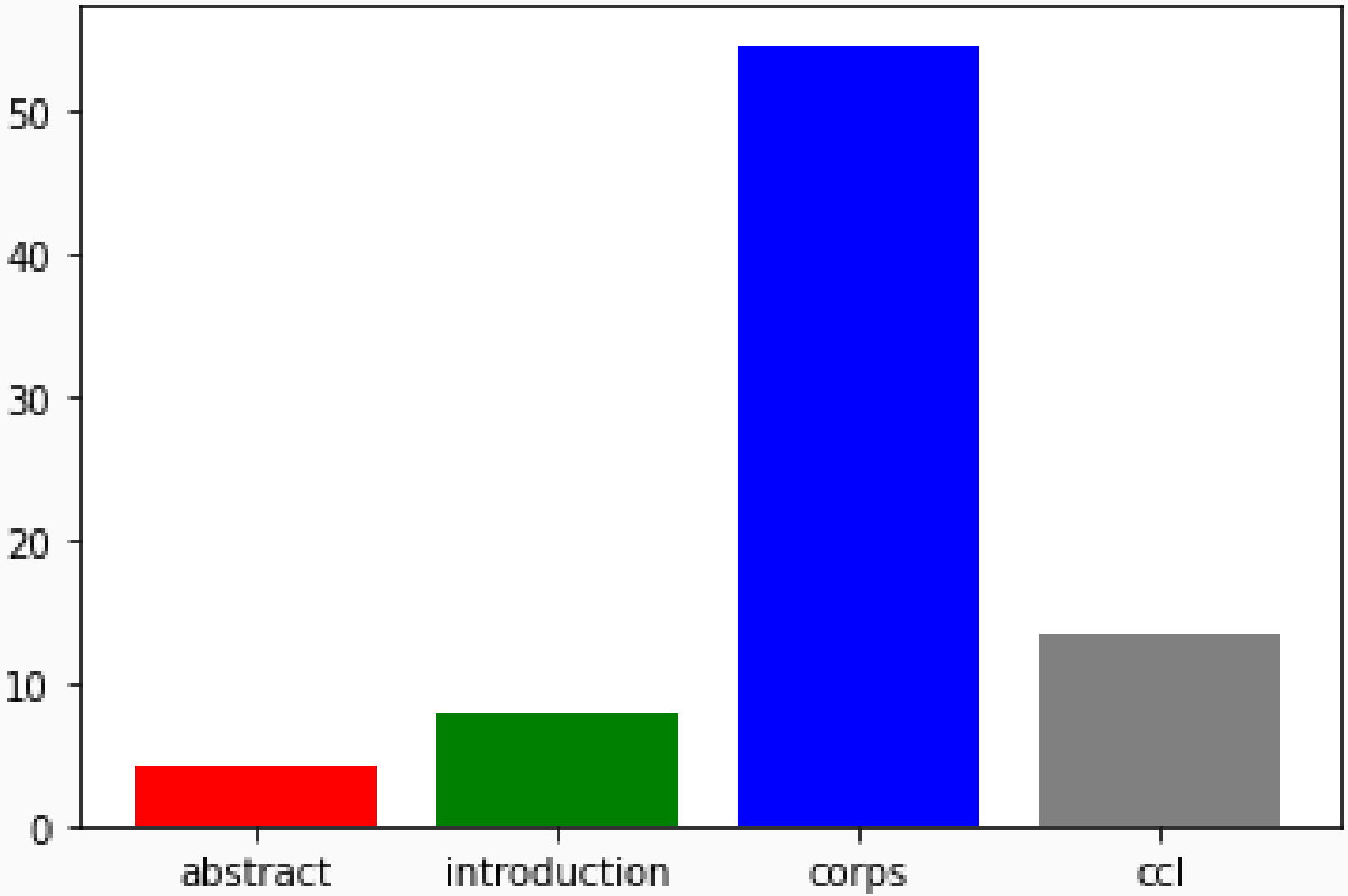


Diachronie


60000
50000
40000
30000
20000
10000
0
1980
1985
1990
1995
2000
2005
2010
2015
2020


80
70
60
50
40
30
20
10
0
1980
1985
1990
1995
2000
2005
2010
2015
2020

Moyenne de *claims* par partie



Références

Blake, C. (2010).
Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles.
Journal of Biomedical Informatics, 43(2):173–189.

Rubin, V. (2006).
Identifying Certainty in Texts.
PhD thesis, Syracuse University.

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., and Satoto, B. D. (2018).
Integration k-means clustering method and elbow method for identification of the best customer profile cluster.
IOP Conference Series: Materials Science and Engineering, 336:012017.

Zhu, C., Qiu, X., Chen, X., and Huang, X. (2015).
A re-ranking model for dependency parser with recursive convolutional neural network.