13/11/2024

IA717 NLP, CHAI & fairness : linguistics of AI ethics charters & manifestos

Présentation des travaux

Superviseur : Maria Boritchev
Étudiants : Dan Hayoun ; Josephine Bernard ; Laury Magne ; Nicolas Allègre

I/ Introduction
II/ Initialisation (CORTEX & IRAMUTEQ)
III/ Corpus élargi (TXT)
IV/ Analyse approfondie

I/ Introduction

II/ Initialisation (CORTEX & IRAMUTEQ)

III/ Corpus élargi (TXT)

IV/ Analyse approfondie

I/Intro - Problématiques

“Fairness” semble essentiel dans l’usage éthique des IA

BUT : Dans quelle mesure il est employé dans les documents (légaux, techniques, business, …) officiels

- Est-ce qu’il est mis en avant ou peu présent dans les documents ?
- Peut-on lui donner une définition ? Générale ? Spécifique au document ?
- À quels mots l’associe-t-on ?
- Est-il est porteur de sens ?

En résumant, quels sont les différents usages du mot “fairness” ?

I/Intro - Initialisation

Récupération du projets et pré-processing :

- Récupération du projet depuis GitLab :
  - Problèmes d'encodage pour avoir des documents lisibles sur toutes machines
  - Rangement des données dans dossiers spécifiques
  - Création classe Python d'abstraction d'extraction des data par algo
- Script de pré-processing pour créer les datas :
  - doc / html => 624 documents (sur 730 URL)
  - transformation en txt => 620 documents (plus facile à parser)
  - algo cortext
  - algo iramuteq

I/ Introduction

II/ Initialisation (CORTEX & IRAMUTEQ)

III/ Corpus élargi (TXT)

IV/ Analyse approfondie

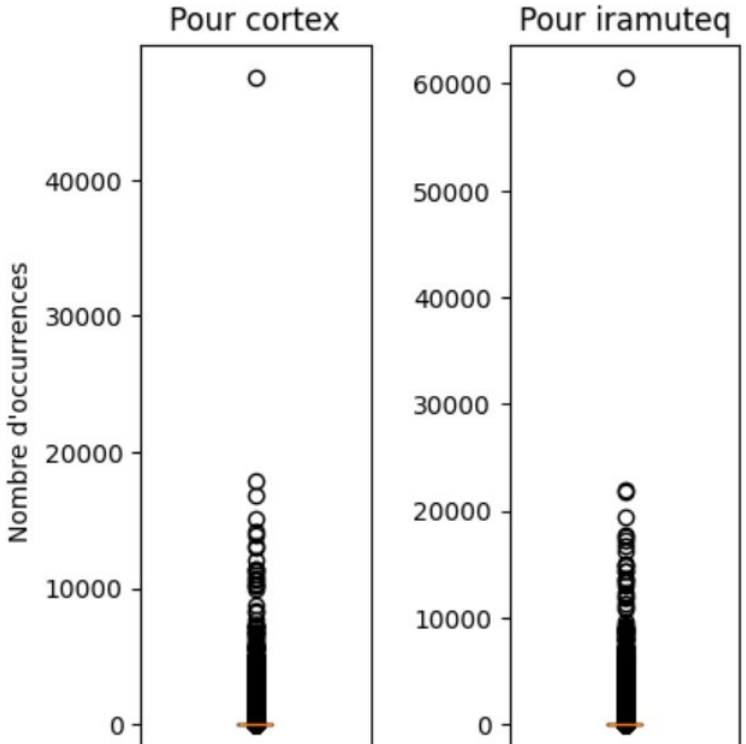
II/ Init - Corpus générés par iramuteq et cortex (1/3)

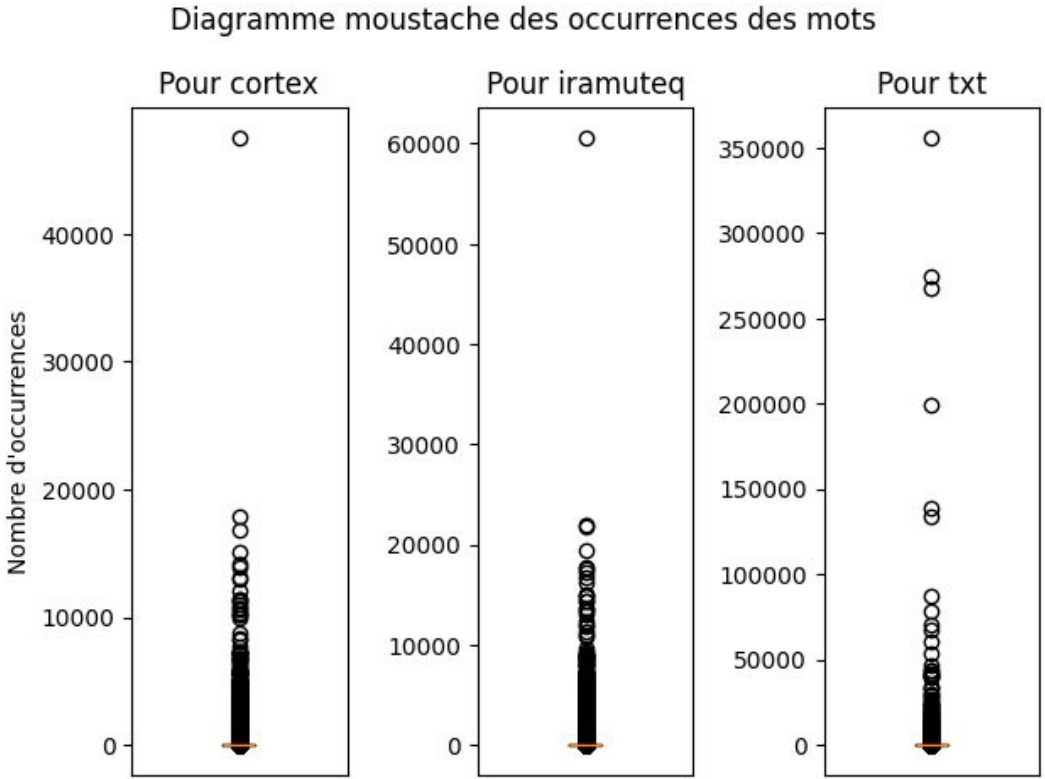
Étudier le mots directement dans les corpus générés iramuteq et cortex

	cortex	iramuteq
data	47499	60534
systems	17878	21949
use	16875	21791
intelligence	15079	19373

Pas de surprise, peu d'information, résultat similaire

=> Mais beaucoup d'extrême lié au thème des documents





Résultat sans retirer les stopword.

Uniquement en retireant les mots à 1 caractère

II/ Init - Corpus générés par iramuteq et cortex (2/3)

Étudier le mot directement dans les corpus générés iramuteq et cortex

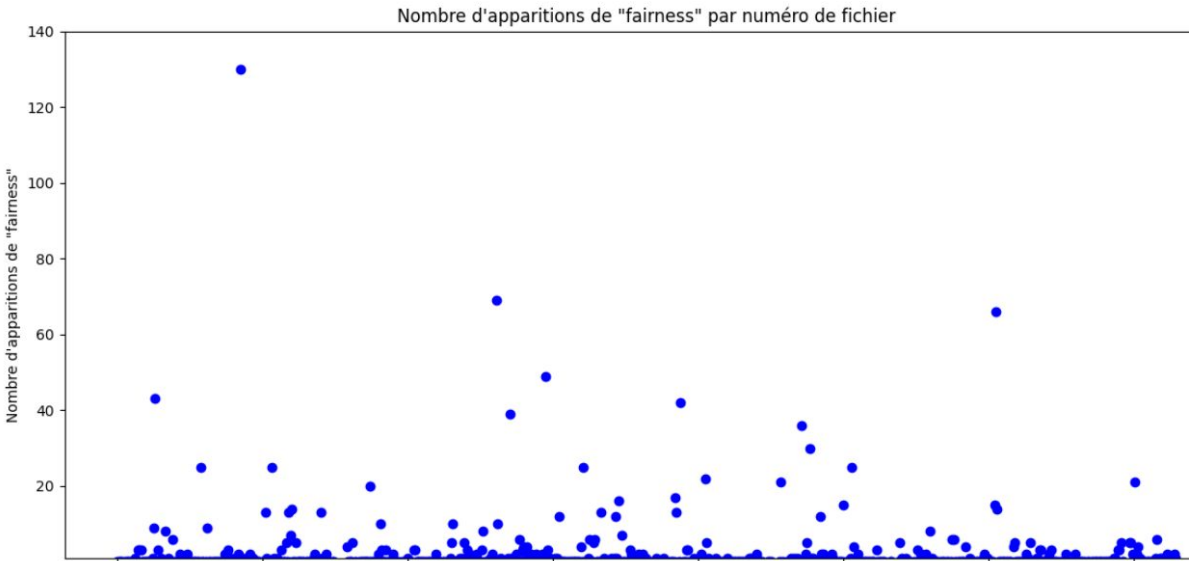
1. compter les mots (au sens classique)
    - …, Artificial, Intelligence,... et le “mot” [Artificial Intelligence] ?
    - d’ailleurs “qu’en est-il de with or without fairness ?
2. on récupère les 3-gram

II/ Init - Corpus générés par iramuteq et cortex (3/3)

Focus sur des mots qui portent sur l'éthique

	cortex	iramuteq
fairness	**2275**	**2807**
algorithmic fairness	68	80
bias	4035	4679
gdpr	2095	2619
discrimination	2364	3011
biometric	1658	1803
regulation	3659	5050
artificial general intelligence	87	106

Le nombre d'occurrence du mot fairness reste significatif (mot le plus fréquent à 45 - 60k)



II/ Init - Cortex et iramuteq (conclusion)

Le mot “fairness” est très présent globalement dans les documents.

Selon la définition qu’on lui donne, on peut également lui associer d’autres mots, comme des antonymes :

- discrimination = 3 011
- bias = 4 679

Par contre :

- Impossible de savoir s’il est décoratif ou significatif ?
- Ou s’il est utilisé sous une forme négative ?

II/ Nouvelles idées

- **Reconstruire un corpus plus permissif au stopwords (pour capter le context par exemple)**
  - Utiliser les données brutes textuelles
  - Exemple : And et Or à supprimer des stopwords pour détecter les simples énumérations
- **Élargir le vocabulaire à étudier avec les synonymes et les antonymes de “fairness”**
- **And et Or à supprimer des stopwords pour détecter les simples énumérations**
- **Analyse des phrases pour voir la portée du mot Fairness**

I/ Introduction

II/ Initialisation (CORTEX & IRAMUTEQ)

III/ Corpus élargi (TXT)

IV/ Analyse approfondie

III/ Corpus élargi (TXT)

Nous utilisons maintenant les fichiers TXT générés à partir des PDF & HTML

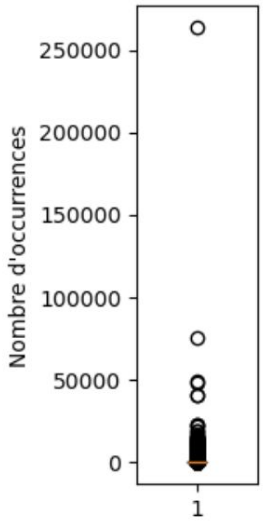
- **Nettoyages :**
  - Conversion en ASCII et minuscule
  - Retrait caractère non-alphabétique sauf ‘-. (=>apostrophe, mot-lié, acronyme ou url)
    - mais retrait des ponctuation finale
  - Retrait mot à une seule lettre
  - Puis retrait des stopwords spécifiques

III/ Corpus élargi (TXT)

Travaux sur les stopwords :

- **1- Lister les mots les plus fréquents**
  - beaucoup sont jugés peu significatifs
  - le plus fréquent est “the” = 355 298
- **2- Filter en analysant manuellement pour affiner la liste :**
  - the : X
  - AI : on garde
  - is : X
  - are : X
  - or : on garde
  - data : on garde
  - …
- **But : garder le context d’usage comme :**
  - **énumérations, négations et relation**

Pour corpus élargi



III/ Corpus élargi (TXT)

Les usages de Fairness: **Décoratif ou significatif** ? À quoi est-il associé?

1. **Le POS tagging** (associer une catégorie grammatical comme nom, verbe, adjectif etc.)

Les rôles de "fairness" les plus représentatifs:

- Objets de prépositions
- Composition de mots

Donc fairness dans l'ensemble du corpus est souvent mentionné avec d'autres idées ou valeurs.

Rôles	Fréquence
pobj	731
compound	411
conj	246
dobj	294
nsubj	63

III/ Corpus élargi (TXT)

Les usages de Fairness: Décoratif ou significatif ? À quoi est-il associé?

2. Focalisation sur les bi-gram/tri-grams associé à Fairness.

Qu'est-ce que CountVectorizer

```
vectorizer = CountVectorizer(vocabulary=vocab, ngram_range=(1, 3), stop_words=stop_words)
X = vectorizer.fit_transform(list_corpusbytxtclean)
```

Points d'attention :

- Définition des stopwords
- Pour préciser le vocabulaire donné il faut prétraiter le corpus (ramener les mots à leur forme racine ou canonique).

III/ Corpus élargi (TXT)

Les usages de Fairness: Décoratif ou significatif ? À quoi est-il associé?

I/ Introduction

II/ Initialisation (CORTEX & IRAMUTEQ)

III/ Corpus élargi (TXT)

IV/ Analyse approfondie

IV/ Idée : élargissement du vocabulaire

Mots voisins de fairness :

bias, justice, unbiased, impartiality, objectivity, balance, honesty, neutrality, equity

Quelques observations :

- Justice, bias particulièrement présents dans le corpus
- Bias et Fairness sont très présents ensemble (txt.85)

IV/ Idée : élargissement du vocabulaire

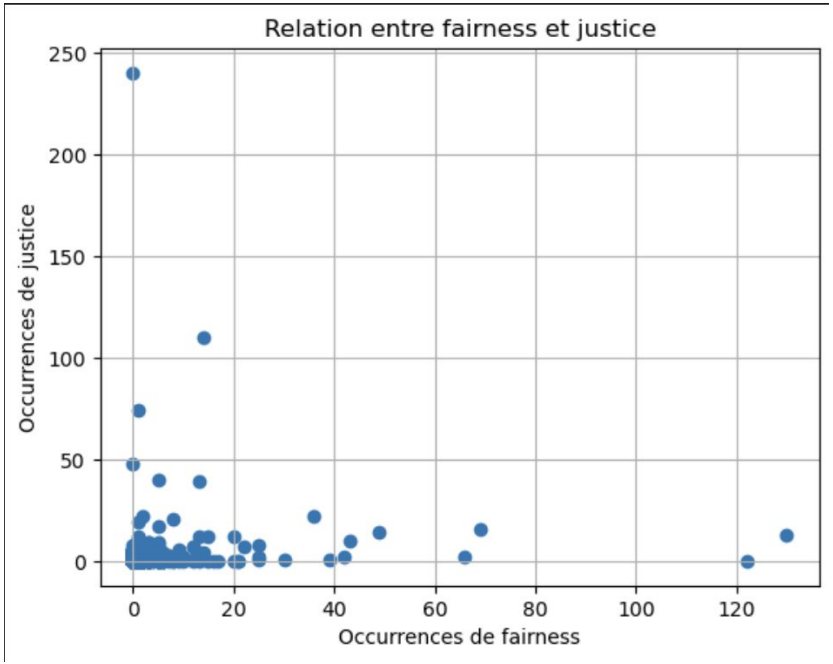
Terme	Frequence
justice	1777
unbiased	79
impartiality	83
objectivity	26
balance	607
honesty	10
integrity	582
equity	548
neutrality	82

Terme	Frequence
bias	3565
inequity	27
unfairness	86
partiality	1
prejudice	161
favoritism	2
injustice	20
discrimination	2365

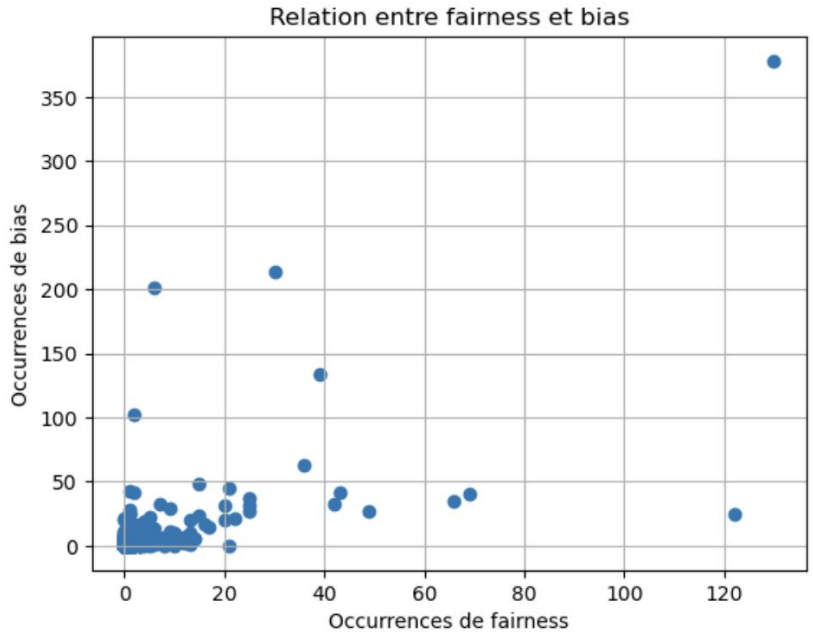
fairness : 2892 apparitions

IV/ Idée : élargissement du vocabulaire

r = 0.10



r = 0.64 : corrélation présente



Conclusion

- **L’étude des mots proches nous en apprend davantage sur fairness**
- **Il peut être plus intéressant d’étudier d’autres mots que fairness, pour étudier cette notion d’”équitabilité”**
- **(À venir) Étudier les bi et trigrammes des ‘mots proches’ (notamment des mots corrélés)**
- **Étudier la distance au verbe**
- **Globalement, fairness utilisé dans des énumérations et de manière peu significative. Pour étudier la question il faut continuer le travail sur les synonymes**


THANK
YOU
FOR
YOUR
ATTENTION