IA 717: CHAI & fairness: linguistics of AI ethics charters & manifestos

Sommaire

I) Introduction

Dans ce projet, nous nous donnons pour but d'analyser l'utilisation du mot "fairness" dans un corpus de papiers de recherches, industriels ou législatifs, qui abordent les liens entre éthique et IA. Nous nous posons plusieurs questions sur l'utilisation de ce mot :

1) Le mot "fairness" est-il mis en avant ou peu présent dans les documents ?
2) Peut-on donner une définition au mot "fairness" ? Une définition générale ou spécifique aux documents ? Indirectement, y-a-t-il un concept linguistique qui ressort de l'emploi du mot ?
3) Quels mots sont les plus présents au voisinage du mot "fairness" ?
4) Le mot "fairness" est-il utilisé de manière plus significative ? Ou plus décorative ?
5) Quels sont les différents usages du mot "fairness" ?

Ce projet se base sur le corpus mapaie. Nous récupérons ce projet à partir de Gitlab. Le corpus est composé de 624 documents, répartis entre documents au format PDF et HTML, qui sont ensuite convertis en texte brut (TXT). Le code de ce dépôt applique différents algorithmes de pré-process au corpus pour pouvoir l'analyser, la méthode CORTEX et la méthode IRAMUTEQ. Ensuite, nous travaillons sur 3 corpus, les corpus générés respectivement par les méthodes CORTEX et IRAMUTEQ et le corpus issu directement des fichiers texte brut TXT.

Nos livrables sont sur le dépôt GIT Gitlab dans le dossier notebook/, les Notebook Jupiter seront également convertis en HTML/PDF. Notamment :

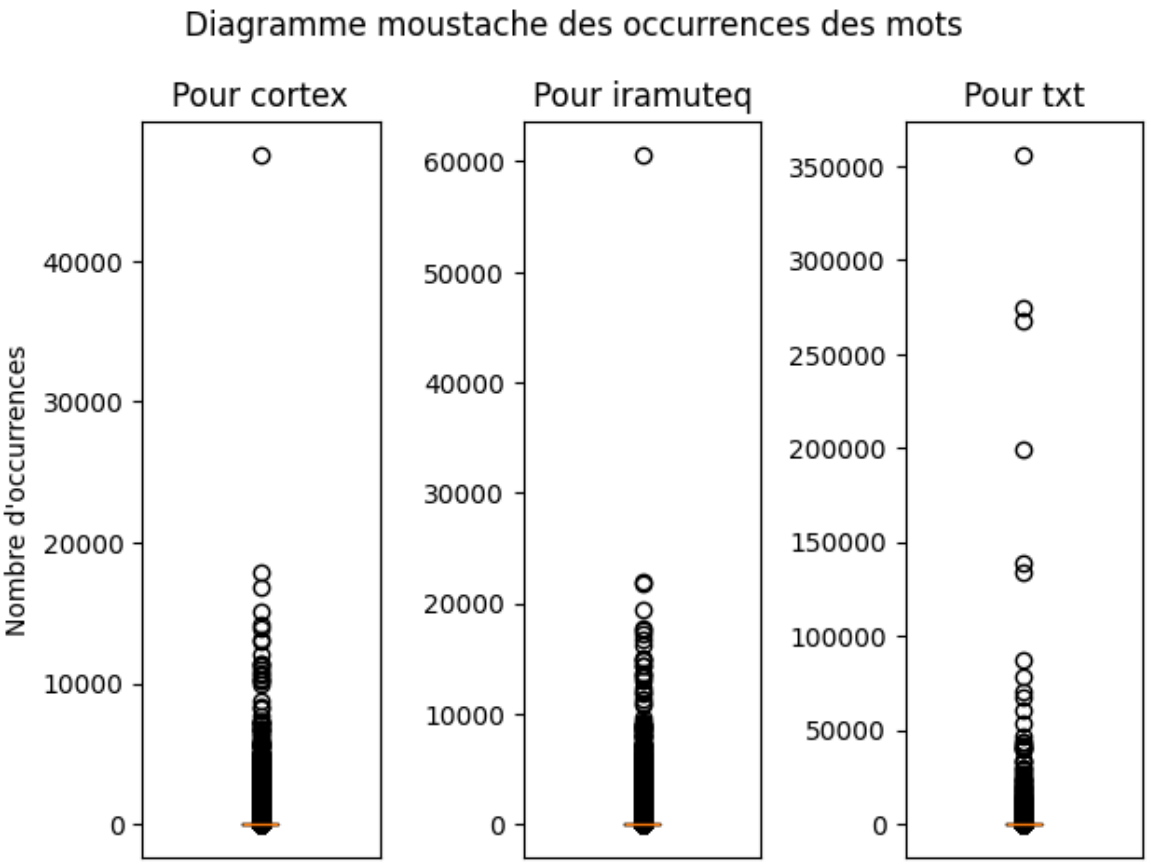
- *IA_717_ProjectCHAI-fairness_Student-FINAL.ipynb* : Notebook comprennant la première partie
    - *Corpus.py* : classe spécifiquement créée pour charge uniformément les 3 corpus
- *IA717_ProjectCHAI-fairness_AMR_Student.ipynb* : Notebook comprenant la partie seconde partie, AMR
- *metamorphosed/* (*amrdoc.py* et *amreditor.py*) : version modifiée des modules de l'application Metamorphosed pour une utilisation dans les Notebook

II) Étude des corpus

II.1) Analyse des 3 corpus CORTEX, IRAMUTEQ et TEXTE brut

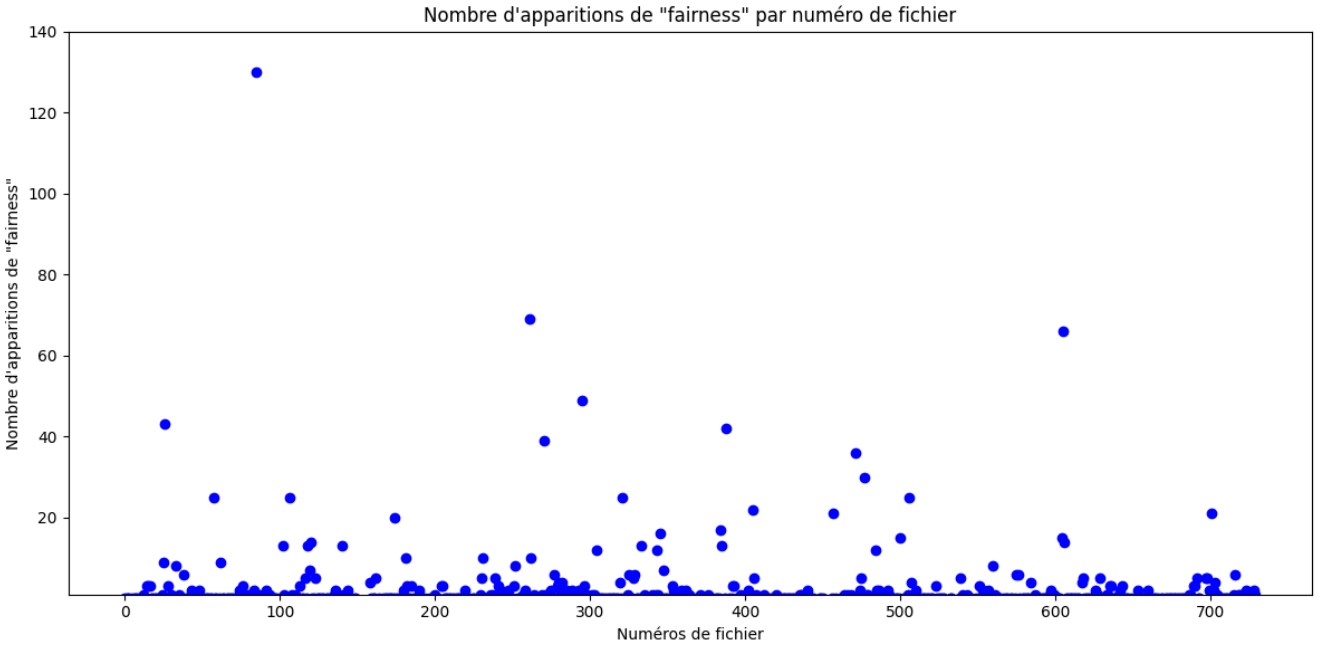
Nous étudions d'abord les mots directement des corpus en retirant les occurrences uniques :

```
*** MODEL : cortex
      cortex : Taille corpus = 30 012 540
      cortex : Mot le plus utilisé : data = 47 499
      cortex : Nb initial mot : 116180, Nb mot >1 occurence : 62 737
*** MODEL : iramuteq
      iramuteq : Taille corpus = 40 105 251
      iramuteq : Mot le plus utilisé : data = 60 534
      iramuteq : Nb initial mot : 154432, Nb mot >1 occurence : 81 494
*** MODEL : txt
      txt : Taille corpus = 65 113 631
      txt : Mot le plus utilisé : the = 355 298
      txt : Nb initial mot : 473509, Nb mot >1 occurence : 199 792
```



cortex	cortex	iramuteq	iramuteq	txt	txt
data	47499	data	60534	the	355298
systems	17878	systems	21949	and	274848
use	16875	use	21791	of	267744
intelligence	15079	intelligence	19373	to	198824
also	14170	also	17698	in	138304
public	13926	human	17608	a	133383
system	13097	public	17303	for	87271
human	12998	system	16635	AI	78249
artificial	12003	artificial	16222	is	70804
research	11421	des	15060	that	67586

Nous constatons que les corpus CORTEX et IRAMUTEQ sont sur des données ayant déjà été nettoyées donc les stopword ont été retirés. Mais grâce aux corpus brut TXT, nous voyons qu'il y a un mot "AI" qui a été complètement retiré or dans notre cas c'est un mot important. De plus, les deux corpus préprocessés sont similaires en termes de fréquence de mot et représentent le thème des documents.



Le mot “fairness” est très présent globalement dans les documents. Selon la définition qu’il lui ait donné, nous pouvons également lui associer d’autres mots, comme des antonymes :

- discrimination = 3 011
- bias = 4 679

Cependant, sa fréquence ne permet pas à elle seule de déterminer son impact ou son importance réelle, ni de connaître l’usage de ce mot (significatif, décoratif) ni son contexte (forme positive, forme négative).

II.2) Analyse du corpus élargi

Maintenant, notre objectif est désormais d'analyser plus en profondeur la représentation de ce terme dans l'ensemble des documents pour comprendre les usages du mot Fairness :

- Fairness est-il utilisé de manière significative ou seulement décorative ?
- À quoi est-il associé syntaxiquement et lexicalement ?

Mais dans un premier temps, il faut construire notre propre corpus afin de garder les éléments qui nous semblent essentiels. Notamment, il faudrait construire un corpus plus permissif avec des stopword prédéfinis, notamment pour garder le contexte, comme les mots de liaison (‘or’, ‘and’, ‘with’).

II.2.1) Définition du corpus

Notre but est ainsi de nettoyer le corpus pour garantir une bonne analyse syntaxique. Cependant, nous avons pris parti de garder un maximum d’information syntaxique afin d’aider le modèle à associer une catégorie grammaticale à chaque terme, faisant que l’on ne peut ici utiliser les tokens définis précédemment.

Nettoyages effectués :

- Suppression des caractères spéciaux non significatifs et uniformisation mais en conservant les ponctuations utiles afin de délimiter les phrases et structurer le texte :
    - Conversion en ASCII et minuscule
    - Retrait caractère non-alphabétique sauf les `'-.` (=>apostrophe, mot-lié, acronyme ou url)
        - mais retrait des ponctuations finales pour la séparation des stopwords
- Retrait mot à une seule lettre
- Puis retrait des stopwords spécifiques

Choix des stopwords :

Nous avons listé les mots les plus fréquents puis nous avons filtré la liste afin de garder les mots permettant de garder le contexte notamment (énumération, négation, relation). Après analyse du top100, nous avons décidés de ne garder que les lettres alphanumérique (cf. nettoyages effectués) et nous avons ainsi choisi de retirer les stopword suivants :

```
stop_words = ['the','of','to','in','a','for','is','that','on','be','are','de','by','an',
'this','it','will','which','la','their','at','et','also','in','such','des','this']
```

II.2.2) Analyse fréquentielle

L’analyse fréquentielle effectuée précédemment sur les 3 corpus donne les mêmes résultats globaux constatés. Il faut donc, et nous pouvons maintenant grâce à un corpus nettoyé et ayant gardé le sens sémantique, étudier des méthodes plus avancées de NLP.

II.2.3) Analyse utilisant le Part-of-Speech (POS) Tagging

Dans un premier temps, nous allons appliquer le POS tagging pour identifier les rôles syntaxiques du terme fairness dans le corpus. Cela permet de comprendre si fairness joue un rôle central dans les phrases (par exemple, en tant que sujet) ou s'il est davantage utilisé de manière secondaire (par exemple, en tant qu'objet ou dans des prépositions).

Il faut noter que le modèle SpaCy “en_core_web_sm” est utilisé et les documents du corpus sont traités en parallèle avec la méthode nlp.pipe, en ajustant le batch_size pour réduire le temps de calcul. Le corpus nettoyé, utilisé pour les tokens et le n-gram, ne sera pas employé ici. En effet, Le POS-tagging a une nécessité de garder plus de sémantique grammaticale. Comme le fait de garder les majuscules (ex. identifier un nom propre) ou certains mots à une seule lettre qui ont un sens important afin de comprendre la structure linguistique (ex. “a” : article indéfini, “I” : pronom personnel sujet).

Résultat :

Rôle grammatical	Nb occurence
pobj	648
compound	363
conj	249
dobj	218
nmod	93
nsubj	80
nsubjpass	18
appos	14
npadvmod	13
ROOT	9
attr	7
xcomp	5
dep	2
relcl	2
pcomp	2
advcl	2
amod	2
ccomp	2
punct	1
acl	1

Donc l’analyse montre que fairness est principalement utilisé dans des relations secondaires avec d’autres termes, notamment comme :

- Objet de prépositions : Associé à des idées ou des valeurs à travers des phrases structurées par des prépositions.
- Composition de mots (compound).

De plus, Fairness est très peu utilisé (facteur 10) comme sujet de phrase que comme objet de préposition. Cela suggère que fairness est rarement le sujet principal d’une phrase. Cela est surprenant dans un corpus centré sur l’éthique de l’intelligence artificielle où nous pourrions nous attendre à ce qu’il soit au cœur des discussions.

Enfin, son rôle syntaxique d’objet souligne son importance indirecte mais renforce aussi l’idée qu’il est souvent mentionné en association avec d'autres concepts.

II.2.4) Analyse des n-grams

Pour approfondir, nous avons étudié les relations lexicales de fairness à travers une analyse des n-grams (ici bigrams et trigrams associés au terme Fairness) en utilisant CountVectorizer. Cela permettra d'identifier les termes qui nuancent fairness.

Nous allons donc:

- Utiliser les tokens et stopword définis précédemment. En effet, les ensembles de mots contenant des stopwords ne seront jamais pris en compte par Countvectorizer.
- Lemmatizer : Ramener les mots à leur forme canonique pour éviter les ambiguïtés puisque CountVectorizer sépare binairement les mots (donc sensible à la case, etc.)
- Garder l'ordre des termes dans les ensembles de mots car lors d'une énumération par exemple le contexte peut différer selon la position des termes.

Nous savons d'après le nuage de mots que les bigrams et trigrams les plus fréquents sont d'abord "fairness and" et "and fairness", puis "fairness accountability" et "fairness accountability and". Nous pouvons aussi distinguer facilement "conference fairness", "conference fairness accountability" et "algorithmic fairness".

<u>Interprétation:</u>

- Les bigrams les plus fréquents indiquent que fairness est très souvent utilisé dans des propositions associatives (énumérations ou conjonction) qui placent fairness dans une relation d'interdépendance. Le terme est perçu comme un composant parmi d'autres, plutôt qu'un thème clé unique.
  Il faut cependant noter que dans une énumération, l'ordre des mots peut montrer l'importance des mots. Fairness apparaissant principalement en première position est présenté comme la notion initiale de l'énumération. Ce ne serait donc plus seulement un élément d'une liste exhaustive mais une notion centrale.
- Ensuite dans "fairness accountability" et "algorithmic fairness", Fairness caractérise des notions spécifiques (comme "algorithmic"). Il a donc un rôle plus ciblé car ajoute une dimension éthique à un concept clé. Cependant il n'est pas ici le sujet principal.

En fin de compte, l'analyse syntaxique et lexicale du terme fairness montre une dualité dans son usage dans le corpus. Malgré le fait que fairness peut être mis en avant, notamment comme objet ou pour enrichir un concept, il est souvent employé dans des énumérations et des conjonctions, jouant ainsi un rôle décoratif ou répondant peut être à une intégration conventionnelle dans les discussions sur l'éthique en IA.

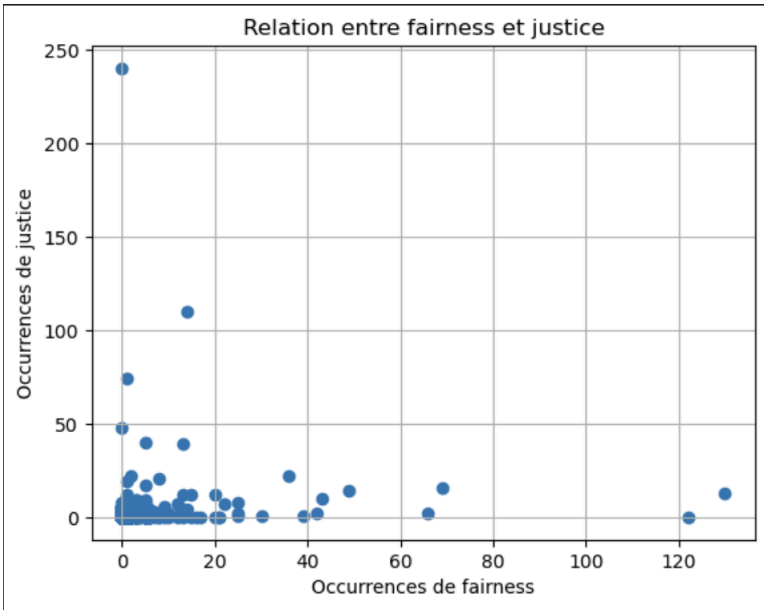
II.2.5) Élargissement du vocabulaire

Nous étudions maintenant les synonymes et antonymes du mot fairness, pour voir s'ils peuvent nous apporter des informations supplémentaires sur l'utilisation de fairness. Particulièrement, nous regardons notamment la corrélation entre fairness et justice, et, fairness et bias.

II.2.5.1) Analyse du mot 'justice'

Corrélation de 'justice' avec 'fairness' r=0,1 :

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



Nous remarquons que fairness et justice sont très peu corrélés. Cela signifie que les deux mots apparaissent dans des textes différents. Nous pouvons donc supposer dans des contextes différents.

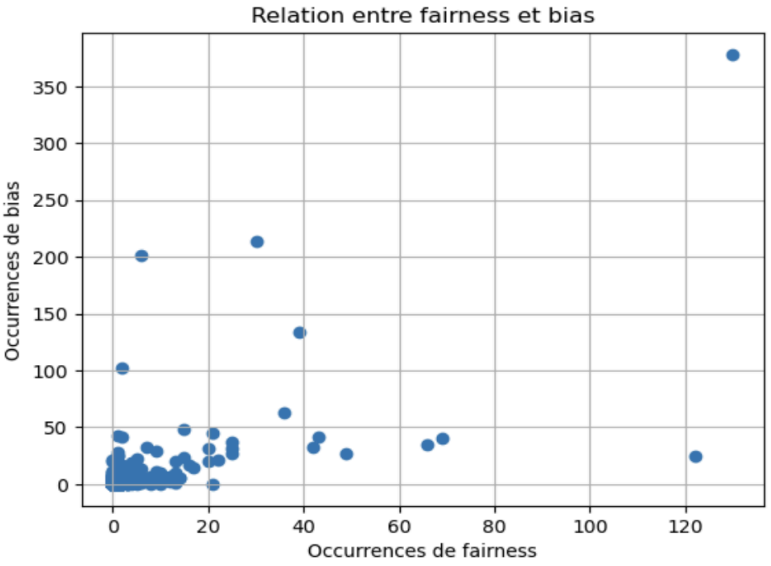
Nuage de mots pour 'justice' :

Nous remarquons que justice est aussi utilisée dans des énumérations, son sens n'est pas plus fort dans le corpus que celui de fairness. De plus, ce mot a l'air plus significatif. En effet, nous remarquons qu'il est clairement utilisé pour parler de décisions pénales, juridiques, et un peu plus rarement pour parler des décisions qu'est amené à prendre une IA. Ainsi, il semble déjà que le mot fairness ne soit pas trop utilisé pour parler des thèmes juridiques/pénales de l'éthique de l'IA, car il est préféré par le mot 'justice'. Donc fairness n'est pas utilisé dans un sens juridique, pénal, dans ce corpus, étant donné qu'il n'est pas corrélé à justice. Nous pouvons donc exclure ce sens pour le mot fairness.

II.2.5.2) Analyse du mot ‘bias’

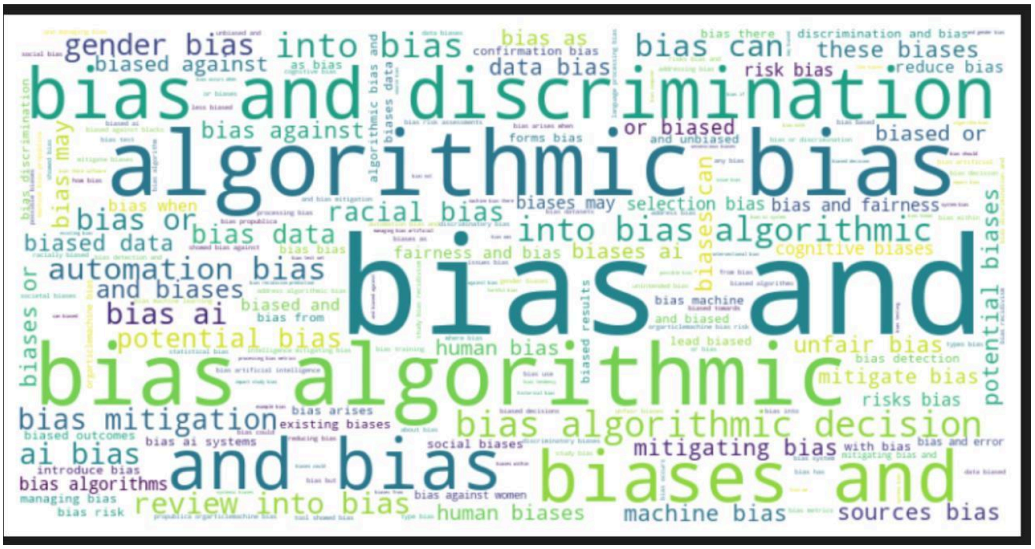
Corrélation ‘bias’ avec ‘fairness’ r=0,64 :

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



Fairness et bias possèdent quant à eux une corrélation assez élevée. Ainsi, ils sont utilisés dans les mêmes textes et probablement dans les mêmes contextes.

Nuage de mots pour ‘bias’ :



Le terme bias est fréquemment employé dans des énumérations, comme en témoigne la forte occurrence des expressions « bias and » et « and bias ». Il n'a donc pas une importance sémantique supérieure à celle de fairness dans l’ensemble du corpus. Cependant, bias est souvent associé à des notions de discrimination, notamment « gender », « racial » ou « discrimination ». Cela suggère que fairness est également largement utilisé dans ce contexte, en raison de la corrélation entre ces deux termes. Par ailleurs, bias est aussi utilisé en algorithmique, en particulier dans les réseaux de neurones le ‘bias est sur les poids. Cette utilisation est sans intérêt ici pour notre étude, car fairness ne fait pas référence à des coefficients.

Les contextes d’utilisation de bias et justice peuvent donc nous en apprendre davantage sur le contexte d’utilisation de fairness. Ainsi, l’étude des mots proches nous permet de préciser les sens d’utilisation du mot fairness : il n’est pas ou peu utilisé dans le thème juridique, et utilisé pour parler des discriminations.

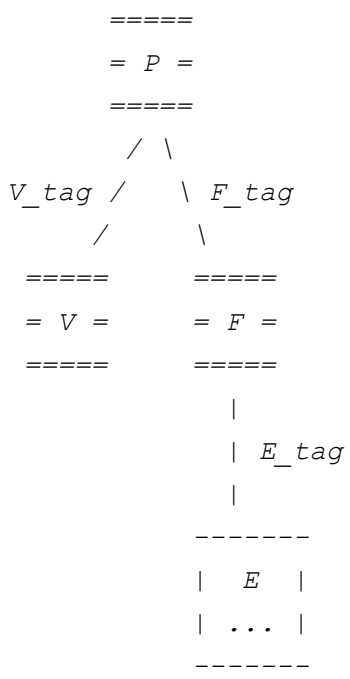
III) Étude des AMR

III.1) Objet

L'objectif est toujours de comprendre comment le mot fairness est utilisé, ses relations sémantiques et son rôle dans différents contextes. Pour pousser cette étude, nous allons nous pencher sur les concepts proches dans les phrases en utilisant les représentations AMR (Abstract Meaning Representation).

En particulier, analyser :

- les parents (concepts supérieurs) du mot "fairness" ;
- les voisins (concepts ayant le même parent) ;
- les relations définies dans les graphes.



III.2) Implémentation

Le fichier `key_penmans.amr` nous a été fourni en amont. Il a été généré par un modèle à base réseau de neurones qui transforme une phrase en langage naturel (anglais ici) en arbre AMR, via le module ARMlib en prenant toutes les phrases contenant le mot fairness.

Ainsi, nous nous retrouvons avec 1201 instances d'objet AMR prenant la structure suivante :

 # ::snt [phrase en langage naturel]
 # File [numéro du fichier = document]

[arbre]

Par exemple, voici le premier item de notre corpus :

```
# ::snt First of all, how can existing regulations or new ones facilitate algorithmic accountability  or fairness?
# File 167

(p / possible-01
    :ARG1 (f / facilitate-01
        :ARG0 (o / or
            :op1 (r / regulate-01
                :ARG1-of (e / exist-01))
            :op2 (r2 / regulate-01
                :ARG1-of (n / new-01)))
        :ARG1 (o2 / or
            :op1 (a / accountable-02
```

```
            :mod (a2 / algorithm))
        :op2 (f2 / fair-01))
    :manner (a3 / amr-unknown))
:mod (f3 / first-of-all))
```

Nous utilisons des versions modifiées des modules AMRdoc et AMReditor extraits d'une application "metamorphosed" pour d'une part charger les données dans des structures facilement lisibles (AMRdoc) et d'autre part pour manipuler plus simplement la syntaxe penman et donc de lire les arbres ARM (AMReditor).

Le module AMRdoc permet de charger l'intégralité des données dans la structure *amr_mapaie* et donne l'accès à chaque phrase en parcourant *amr_mapaie.sentences.*

Initialisation : `amr_mapaie = amrdoc.AMRdoc(PATH_DATA_FILE_AMR)`

Par exemple :

- `amr_mapaie.sentences[0].text` : la phrase de l'AMR d'indice 1
- `amr_mapaie.sentences[0].comments[0]` : le numéro du fichier (Format : 'File xxx')
- `amr_mapaie.sentences[0].amr` : accès direct en format penman de l'AMR

Ensuite à l'aide de AMReditor, nous manipulons la syntaxe penman et l'arbre AMR *amr_sentence* :

Initialisation :

```
ap = amreditor.AMRProcessor()
ap.lastpm = amr_sentence.amr
if not ap.isparsed:
   ap.readpenman(ap.lastpm)
```

Deux propriétés sont particulièrement utiles ensuite :

- `ap.triples` qui renvoie la liste des triplets de chaque noeud de l'arbre
  [..., ('a2', ':instance', 'algorithm'), ('f2', ':instance', 'fair-01'), …]
- `ap.vars` qui renvoie l'ensemble des objets dans un dictionnaire dont la clé est l'identifiant du concept et la valeur, le concept.
  {..., 'o2': 'or', 'a': 'accountable-02', 'a2': 'algorithm', 'f2': 'fair-01', …}

Enfin, pour récupérer les relations qui nous intéresse, nous avons implémentés les fonctions suivantes :

- `get_parent_infos(targets, amr_sentence)` qui prend en paramètres une liste de concept et une phrase AMReditor et, pour chaque triplet contenant l'un des concepts de *targets,* renvoie les tuples :
  - *`parent_var`* : l'identifiant du parents
  - *`parent_type`* : le type du parents
  - *`relation`* : le type de relation entre le parent et le target
- `get_node_children(node_id, amr_sentence, list_concept_tosee)` qui prend en paramètre un identifiant d'un noeud (ou concept), la phrase AMReditor et une liste de concept et renvoie, pour chaque triple contenant l'un des concepts de list_concept_tosee les tuples :
  - *`child_id`* : l'identifiant du concept enfant
  - *`child_type`* : le type du concept enfant
  - *`relation`* : la relation entre le parent et l'enfant trouvé
- Et les fonctions pour récupérer tous les parents et tous les enfants

III.3) Analyse

III.3.1) Le concept Fairness

Dans un premier temps, nous avons étudié les concepts pouvant se rattacher au mot fairness par une recherche sur les concepts commençant par “f”. Nous avons sélectionné les concepts 'fair-01', 'fairness', 'fairwash-01'. Malgré la naïveté du filtrage, les résultats sont bons. Nous travaillerons avec ces trois concepts dans la suite de l’étude. Nous les appellerons concept “Fair” pour simplifier l’écriture. À noter, le concept “fairness” n’existe pas dans la banque de données PropBank, et si pour une analyse plus approfondie la liste des concepts des utilisés une normalisation de fearness->fair-01 serait à entreprendre.

```
concepts = set(x[0] for x in [x.split('') for x in amrdoc.relations_between_concepts([amr_mapaie], depth=1)])
```

III.3.2) Primo-analyse

Analyse sommaire initiale en commençant par le top20 des concepts du corpus amr_mapaie :

concept	Nb occurence
and	5658
name	2995
multi-sentence	1020
or	572
possible-01	569
publication-91	395
person	379
publication	362
intelligent-01	354
use-01	354
have-degree-91	334
mean-01	305
date-entity	300
principle	278
recommend-01	257
fair-01	256
develop-02	249
system	246
cause-01	234
ensure-01	233

Puis primo-vision sur les relations sur le concept de fairness sur la base de sortie de la fonction (`amrdoc.relations_between_concepts([amr_mapaie], depth=2)`):

	fair-01	fairness	fairwash-01
:mod	13	105	0
:polarity	58	42	0
:topic	0	10	0
:li	0	2	0
:ARG0	5	0	1
:ARG1	156	0	1
:degree	1	0	0
:source	0	1	0
:prep-on	0	1	0
:beneficiary	0	2	0
:domain	1	5	0
:ARG3	1	1	0
:quant	0	1	0
:manner	5	6	1
:location	3	5	0
:poss	0	15	0
:ARG2	6	0	0
:ARG4	1	0	0
:prep-in	0	9	0
:example	0	1	0
:condition	6	0	0
:time	0	1	0

III.3.3) “Fair” en tant que racine

Nous trouvons que le concept “Fair” n’est utilisé que 13 fois en tant que racine de l’arbre (donc comme concept central) et une seule fois, si nous excluons les titres ou les occurrences seules. Cette phrase est celle avec le concept Fairwash-01 où il apparaît qu’une seule fois.

III.3.4) Petit aparté sur Fairwash-01

La phrase originel :

“They fairwash a model by taking a closed-box model and produce an ensemble of interpretable models that approximate the original model but are much fairer, which then hide the unfairness of the original model.”

Ce concept n'existe pas dans PropBank et a donc été inventé par le NN. Cependant, le réseau de neurones a considéré ce concept particulier et donc de distance dimensionnelle plus éloignée que les concepts de fair/fairness. D'où, l'intérêt pour nous d'effectuer une analyse plus poussée sur ce résultat de la sortie du NN.

L’idée est d’expliquer que certain font du fairwashing, c’est-à-dire qu’ils présentent un modèle explicable qui approxime le vrai modèle utilisé. C’est un néologisme à l’image du greenwashing par exemple.

III.3.5) Les parents de “fair”

Les trois type de parents apparaissant le plus sont :

- `and` (476 fois) : représente des jonctions où "fairness" est combiné avec d'autres concepts. Cela indique que l'équité est souvent vue comme un élément parmi d'autres dans des énoncés.
- `define-01` (57 fois) : évoque l'action de "définir l'équité", suggérant une préoccupation pour clarifier ou conceptualiser ce terme.
- `principle` (51 fois) : associe le concept de "fairness" à des principes ou idéaux, soulignant son rôle comme une valeur fondamentale.

“And” est largement surreprésenté (5 fois plus que le second terme le plus utilisé). Il concerne 36.25% des concepts parents de fairness. Cela confirme bien notre première analyse des postags. Nous pouvons donc en conclure que Fairness est souvent associé à un autre concept par une conjonction de type union.

III.3.6) L'union de "fair"

Nous comptons cette fois-ci, les occurrences des voisins (i.e. même parent) de "fair" dont le parent est "and". Nous trouvons le top 5 suivant :

- 'transparency': 131,
- 'accountable-02': 111,
- 'safe-01': 50,
- 'explain-01': 37,
- 'privacy': 36,

Plus de la moitié des concepts associés portent sur la possibilité de comprendre ('transparency', 'accountable', 'explain'). Nous pouvons en déduire que fairness est en effet un but mais qu'il n'est pas possible d'y parvenir sans comprendre le fonctionnement (explicabilité).

En continuant dans cette voie et en regardant cette fois les relation entre le concept "fair" et son parent, nous voyons que les relations les plus présentes sont :

- :ARG1: 407
- :op1: 263
- :op2: 175
- :topic: 113
- :op3: 61

Les relations opX concernent globalement les opérateurs "and" (en grande majorité comme nous l'avons vu) ou "or". Nous nous penchons donc sur l'opérateur :ARG1 qui représente globalement un deuxième tier de relation avec le concept "fair" (le premier est opX qui rejoint le concept "and").

- Proportion de ':ARG1': 31.00%
- Proportion de ':opX': 38.00%

Nous pourrions aussi nous concentrer sur "topic" ultérieurement.

III.3.7) La relation ":arg1"

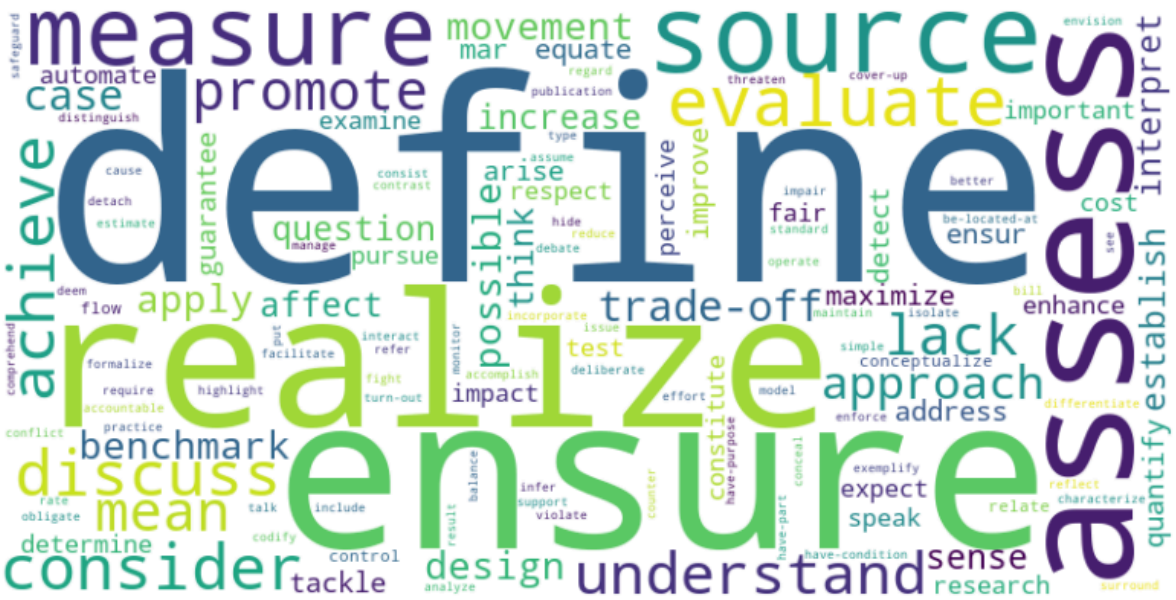
Si "fairness" est un argument (:ARG1), cela signifie que le concept est défini ou assuré par son parent.

Top 10 des concepts utilisés avec :ARG1 :

- define-01: 57
- ensure-01: 31
- realize-01: 23
- assess-01: 16
- source-02: 12
- evaluate-01: 10
- measure-01: 10
- discuss-01: 8
- consider-02: 8
- achieve-01: 8

On pourrait traduire par :

- define-01 : "Définir l'équité" (ou "définition de l'équité")
- ensure-01 : "garantir l'équité" (même remarque)
- etc.



Il serait intéressant ensuite de considérer les associations en regardant le parent du parent et les enfants de “fair” pour éventuellement approcher de l’usage. Trouver par exemple si la phrase pourrait résumer “définir fairness est …” ou “la définition de fairness est …”

IV) Conclusion

En définitive, cette étude nous permet de tirer plusieurs conclusions sur l'utilisation du mot 'fairness' dans notre corpus.

Premièrement, le mot "fairness" est largement présent dans le corpus mapaie. Principalement utilisé en tant qu'objet dans les phrases, il apparaît au cours d'énumérations où il jouerait donc en premier lieu un rôle décoratif. Cela peut-être lié à une intégration conventionnelle du mot dans les discussions sur l'éthique en IA.

Dans l'analyse des bi et tri-grams, le mot "fairness" est associé à d'autres concepts. Notamment, les deux concepts liés au mot fairness : "conference fairness" et "fairness accountability". La première notion suggère une utilisation en référence à des conférences en liens entre fairness et IA. Ainsi, "fairness" est donc beaucoup abordé pour représenter des conférences, des présentations. Cela peut suggérer une utilisation décorative, ou "fairwash". La deuxième notion aborde la responsabilité des acteurs développant des IAs. Cette utilisation est plus concrète et aborde la question de la responsabilité en cas de problèmes ou d'accident (qui serait responsable ? Le développeur ?). Ainsi, cette notion aborde surtout un sujet juridique.

L'étude des mots proches nous permet de préciser les sens d'utilisation du mot fairness. En effet, il n'est pas ou peu utilisé dans le thème juridique mais il est plus utilisé pour parler des discriminations. D'autres sens pourraient encore être trouvés si on continuait la recherche sur d'autres synonymes ou antonymes.

La transformation en AMR montre que le concept "fairness" est peu employé comme racine syntaxique. À contrario, il apparaît fréquemment dans les conjonctions ou en tant qu'argument d'un concept qui le définit. Cela indique qu'il est souvent subordonné à d'autres notions clés, comme la 'transparency', 'accountability' ou 'privacy'. Ou alors qu'il ne porte pas de notion intrinsèque mais le concept "fairness" est donc porté par une action dans la plupart des cas (définir fairness).

D'un point de vue plus sémantique, "fairness" est étroitement lié à l'explicabilité. Cela se remarque comme en témoignent ses associations fréquentes avec des termes tels que 'transparency', 'explain' ou 'accountable'. Cela souligne que l'équité est perçue comme atteignable uniquement si les systèmes sont compréhensibles. De plus, par l'association en tant qu'argument, nous pourrions en déduire qu'une des problématiques est de pouvoir comprendre ("define") le concept fairness avant de pouvoir le "garantir" ou l'appliquer ("realize").