Bases de données et data frame :

- Corpus_Short : base de données originales
- Corpus.Short.Small : base de données réduites avec 60 000 observations (10% du corpus original)
- Corpus.Short.Small.Filt : base de données réduites avec 60 000 observations et avec seulement les articles avec 15 auteurs et moins
- Corpus.CleanedNames : Corpus.Short.Small.Filt après nettoyage des prénoms et la création de la variable qui rassemble prénom et nom (full_name_i).
- **Corpus.CleanedNames.2** : corpus après création de l'algorithme et de la variable "gender", "proportion_male" et "proportion_female". Il s'agit de la base de données d'intérêt à utiliser à l'échelle de l'article pour les statistiques descriptives et les graphiques.
- gender_proba_ssa_df_2 : data frame associant des probabilités de genre à des prénoms uniques en utilisant la méthode ssa.
- gender_proba_napp_df_3 : data frame associant des probabilités de genre à des prénoms en utilisant la méthode napp et en gardant uniquement les prénoms qui n'ont pas été assignés à une probabilité de genre avec la méthode ssa.
- **names_proba_df** : data frame final associant des probabilités de genre à des prénoms en utilisant la méthode ssa et napp
- top10_female_economists : nom et prénom des 10 principales économistes
- common_chinese_names : prénoms chinois courants (auquel le prénom Juan est enlevé car il a de fortes chances d'être plutôt d'origine latine).
- ratip_by_nbauthors : cf suite fiche
- authors_df : permet une analyse de l'assignation du genre **par auteur unique**. Elle contient :
    - une colonne firstname ("all_firstnames_incorpus") qui regroupe l'ensemble des prénoms présents dans "Corpus.CleanedNames.2" et créée à partir de la variable f_name
    - une colonne lastname ("all_lastnames_incorpus") qui regroupe l'ensemble des noms présents dans "Corpus.CleanedNames.2" et créée à partir de la variable l_name
    - une colonne fullname ("all_fullnames_incorpus") qui regroupe l'ensemble des noms et prénoms présents dans "Corpus.CleanedNames.2" et créée à partir de la variable "full_name",
    - une colonne fullname2 ("all_fullnames_incorpus2)" qui regroupe l'ensemble des noms et prénoms présents dans "Corpus.CleanedNames.2" et créée à partir de la variable originale et non nettoyée "author_",
    - une colonne gender ("all_gender_incorpus") qui regroupe l'ensemble des genres présents dans "Corpus.CleanedNames.2" et créée à partir de la variable gender_
    - une colonne proportion_male ("all_proportionmale_incorpus") qui regroupe l'ensemble des probabilités de genre masculins présents dans "Corpus.CleanedNames.2" et créée à partir de la proportion_male
    - une colonne proportion_female ("all_proportionfemale_incorpus") qui regroupe l'ensemble des probabilités de genre féminins dans "Corpus.CleanedNames.2" et créée à partir de la variable proportion_female
    - n_articles qui compte le nombre d'articles par auteur existant

- Gender_of_Articles : analyse, **par rapport au nombre total d'articles**, des auteurs dont le genre a été correctement identifié. Les variables sont définies à partir de la variable ratio_identified_gender:
    - la variable Nb Articles correspondant au nombre d'articles,
    - une variable NoGender qui compte la proportion d'auteurs non identifiés (ratio_identified_gender = NA ou 0) par rapport au nombre total d'articles,
    - AllGender qui correspond au ratio d'article où tous les auteurs ont été identifiés,
    - Gender_Larger0.5 qui correspond au ratio d'articles où au moins 50% des auteurs sont identifiés,
    - Gender_Larger0.9 qui correspond au ratio d'article où au moins 90% des auteurs sont identifiés,
    - Gender_Larger0.25 qui correspond au ratio d'articles où au moins 25% des auteurs ont été identifiés
    - Mean_identifiedGender et Median_identifiedGender qui correspondent à la moyenne et à la médiane du ratio d'auteurs dont le genre a été identifié.

Variables dans Corpus.CleanedNames.2:

- nb_authors : nombre d'auteurs par articles
- author_i : nom et prénom non nettoyés des auteurs, avec i allant de 1 à 15 et un auteur par article. On se focalise seulement sur les articles avec 15 auteurs ou moins car ils représentent 99,93% des articles totaux.
- l_name_i : avec i allant de 1 à 15 qui répertorie le nom de famille de chaque auteur pour chaque article
- **f_name_i** : avec i allant de 1 à 15 qui répertorie le prénom de chaque auteur pour chaque article
- full_name_i : avec i allant de 1 à 15 qui répertorie le prénom et le nom de chaque auteur pour chaque article. Si il y des valeurs manquantes dans les deux colonnes l_name_i et f_name_i alors cette crible renvoie une valeur manquante, s'il y des valeurs manquantes dans une des deux colonnes l_name_i et f_name_i alors la variable renvoie une valeur manquante et sinon elle renvoie le prénom et le nom de chaque auteur
- **gender** : variable qui, grâce à notre algorithme, associe un genre à chaque auteur à partir de la variable "f_name_i". Cette variable peut prendre les valeurs "male", "female", "NA" ou "unknown".
- proportion_male : variable qui donne la probabilité qu'un auteur soit un homme, seulement si le genre est assigné
- proportion_female : variable qui donne la probabilité qu'un auteur soit une femme, seulement si le genre est assigné
- nb_authors_gendered : variable qui compte le nombre d'auteurs qui ont été assignés à un genre (en comptant le nombre de valeurs non manquantes dans la colonne proportion_male)
- ratio_identified_gender : variable qui compte le pourcentage d'auteurs qui ont été assignés à un genre parmi le nombre d'auteurs total (ratio entre nb_authors_gendered et nb_authors). Cette variable est égale à 1 si tous les auteurs de l'article ont été

correctement identifiés, est supérieure à 0.5 si au moins 50% des auteurs ont été identifiés.

- sum_gender_male : variable qui fait la somme des probabilités de genre masculins pour chaque article
- sum_gender_female : variable qui fait la somme des probabilités de genre féminins pour chaque article
- proportion_gender_male_all : variable qui correspond au ratio de la somme des probabilités de genre masculin sur le nombre d'auteur total, pour chaque article
- **proportion_gender_male_id** : variable d'intérêt principale qui correspond au ratio de la somme des probabilités de genre masculin sur le nombre d'auteur correctement assignés, pour chaque article ∆ NaN > Na
- proportion_gender_female_all : variable qui correspond au ratio de la somme des probabilités de genre féminin sur le nombre d'auteur total, pour chaque article
- **proportion_gender_female_id** : variable d'intérêt principale qui correspond au ratio de la somme des probabilités de genre féminin sur le nombre d'auteur correctement assignés, pour chaque article ∆ NaN > Na

Etapes :

- 1. Création des colonnes author_ qui listent les auteurs pour chaque article en prenant une base de données réduites avec 60 000 observations et avec seulement les articles avec 15 auteurs et moins
- 2. Création des colonnes l_name_i et f_name_i
- 3. Nettoyage des colonnes l_name_i et f_name_i : Corpus.CleanedNames
    - Etape 1 : création des colonnes full_name_i : suppression des textes entre parenthèses, des espaces supplémentaires et des espaces de début et de fin, enlever les accents
    - Etape 2 : nettoyer la colonne f_name_i pour améliorer l'identification des prénoms avec la commande gender : supprimer les lettres suivies d'un point, remplacer les ponctuations par des espaces (prénoms composés), supprimer les doubles espaces et les espaces de début et de fin, garder seulement le premier prénom des prénoms composés
- 4. Création de data frames pour assigner des probabilités de genre à chaque prénom
    - Nous démarrons par assigner le genre au prénom des auteurs avec la **méthode ssa** qui semble la plus adaptée pour notre corpus (car issue de données américaines relativement récentes "US Census List", et les Etats-Unis sont composés d'un nombre important de cultures/civilisations, augmentant ainsi l'ensemble de prénoms possibles) : gender_proba_ssa_df et gender_proba_ssa_df_2
    - Deuxième data frame utilisant la méthode "napp" répertoriant les prénoms qui n'ont pas été assignés à une probabilité de genre lors de l'utilisation de la méthode ssa : gender_proba_napp_df et gender_proba_napp_df_3
    - Data frame final regroupant les deux data frame précédentes : names_proba_df.
    - Test pour savoir le nombre de prénoms uniques par rapport au nombre de prénoms correctement assignés : Nous avons un peu plus d'1/3 (10 400) des prénoms dont nous pouvons déterminer le genre à partir des méthodes SSA et NAPP, et 2/3 (17 000) qui ne seront pas reconnus par cette méthode. Les

prénoms qui ne sont pas appariés semblent être issus de cultures non occidentales. Cependant, ces prénoms sont assez rares et reviennent peu régulièrement dans les articles, ce qui affecte donc assez peu l'appariement.

- 5. Création d'un algorithme pour assigner des prénoms à un genre : utilisation de names_proba_df, common_chinese_names et top10_female_economists. Cet algorithme permet de créer les variables "gender", "prop_male" et "prop_female".
- 5.1. Création de la variable "gender"
    - Si la colonne f_name_i contient une valeur manquante ("NA"), alors gender renvoie une valeur manquante ("NA"). Sinon :
    - Si la colonne f_name_i est dans la liste des prénoms chinois courants (common_chinese_names), alors gender renvoie une valeur "unknown". Sinon :
    - Si la colonne f_name_i est dans la liste des femmes économistes (top10_female_economists), alors gender renvoie une valeur "female". Sinon :
    - Si la colonne f_name_i n'est pas dans la data frame names_proba_df, alors gender renvoie une valeur valeur manquante ("NA"). Sinon :
    - Si la colonne f_name_i est dans la data frame names_proba_df et que le genre associé à ce prénom à une probabilité supérieure à 90% d'être celui d'un homme (proportion_male > 0.9) , alors gender renvoie une valeur valeur "male". Sinon :
    - Si la colonne f_name_i est dans la data frame names_proba_df et que le genre associé à ce prénom à une probabilité inférieure à 10% d'être celui d'un homme (proportion_male < 0.1) , alors gender renvoie une valeur valeur "female". Sinon :
    - Si la colonne f_name_i est dans la data frame names_proba_df et que le genre associé à ce prénom à une probabilité comprise entre 90% et 10% d'être celui d'un homme (0.1 < proportion_male < 0.9) , alors gender renvoie une valeur valeur "unknown". Sinon :
- 5.2. Création de la variable "proportion_male" et "proportion_female" dans la data frame "Corpus.CleanedNames.2"
    - Si la variable "proportion_male" (resp. "proportion_female") dans la dataframe "names_proba_df" renvoie une valeur manquante ("NA") alors la variable "proportion_male" (resp. "proportion_female") dans la data frame "Corpus.CleanedNames2" renvoie une valeur manquante ("NA")
    - Si la variable "f_name_i" donc si le prénom est présent dans la liste des prénoms chinois ("common_chinese_names") alors la variable "proportion_male" (resp. "proportion_female") dans la data frame "Corpus.CleanedNames2" renvoie une valeur manquante ("NA")
    - Si la variable "f_name_i" donc si le prénom est présent dans la liste des femmes économistes ("top10_female_economists") alors la variable "proportion_male" dans la data frame "Corpus.CleanedNames2" renvoie 0 (la probabilité que l'auteur soit une femme est de 0) et la variable "proprotion_female" renvoie 1 (la probabilité que l'auteur soit une femme est de 1).
    - Si la variable "f_name_i" donc si le prénom n'est pas présent dans la data frame "names_proba_df" alors la variable "proportion_male" (resp. "proportion_female") dans la data frame "Corpus.CleanedNames2" renvoie une valeur manquante "NA". Sinon (si le prénom est présent) :

    - Si le prénom est trouvé mais que le genre n'est pas "male" ou "female", “proportion_male” (resp. “proportion_female”) prend la valeur NA.
    - Si le prénom est trouvé et que le genre est "male" ou "female", l “proportion_male” (resp. “proportion_female”) prend la valeur de proportion_male (resp. “proportion_female”) associée à ce prénom dans names_proba_df.
- 5.3. Création des variables nb_authors_gendered et ratio_identified_gender
- 6. Vérification de l’algorithme
    - ratio_by_nbauthors : regarder, **par nombre d’auteurs**, la moyenne et la médiane du ratio d’auteurs correctement assignés par rapport au nombre d’auteurs total ainsi que le nombre d’articles
    - **Analyse par auteur unique** (jusqu'à présent l'analyse était effectuée par article) : création d’un dataframe authors_df. Ce data frame montrera qu’il existe environ 111 116 auteurs différents pour 60 000 articles et représente seulement les auteurs existants (dont le nom et prénom existent).
    - ⚠ Regarder, **par nombre d’auteurs pour chaque article ou par nombre d’articles écrits pour chaque auteur**, le nombre d’auteurs dont le genre a été correctement assigné. Ex : sur 64 160 + 25 5113 articles qui ne contiennent qu’un auteur 64 160 ont été correctement assignés (75% correctement assignés). Ce ratio diminue jusqu'à 6 auteurs par article (ou auteurs avec 6 articles). Nous constatons que plus le nombre d’auteurs par articles est important, plus le nombre d’auteurs correctement assignés est important. On constate également de façon générale que 25% des auteurs ont un genre non identifié.
- Etape 7 : Agrégation en travaillant à l’échelle de l’article et vérification de l’algorithme :
    - Agrégation:Créationsum_gender_male,sum_gender_female,proportion_gender_male_all,proportion_gender_male_id,proportion_gender_female_all,proportion_gender_female_id dans Corpus.CleanedNames.2
    - Vérification de l’algorithme :
        - Comparaison du ratio de la somme des probabilités de genre féminin sur le nombre d’auteur total, pour chaque article (proportion_gender_female_all), au ratio de la somme des probabilités de genre féminin sur le nombre d’auteur correctement assignés, pour chaque article. Nous constatons qu’il y a un nombre important d’articles pour lesquels tous les auteurs ont été assignés mais qu’il y a également beaucoup d’articles pour lesquels certains auteurs n’ont pas été assignés à un genre.
        - Analyse, **par rapport au nombre total d’articles**, des auteurs dont le genre a été correctement identifié à partir de la variable ratio_identified_gender : création de la data frame Gender_of_Articles. Nous voyons que pour environ 15% des articles, il nous est impossible d'assigner le genre à l'un des auteurs. Pour plus de 50% des articles nous sommes capable d'identifier le genre de tous les auteurs. Pour 75% des articles nous sommes capables d'assigner la moitié des auteurs. Pour la moitié des articles nous sommes capable d'assigner le genre de 90% des auteurs. Pour environ 80% des articles, nous pouvons assigner le genre de 25% des auteurs.

- Étape 8 : Analyse des valeurs manquantes - analyse de l’assignation du genre à l’échelle des auteurs. 30 191 auteurs qui n'ont pas été assignés sur 111 116 auteurs différents (27% d'auteurs non assignés). Parmi eux il s’agit de nombreux non occidentaux.
- Etape 9 : Création de variables catégorielles

Limites :

- Omission des auteurs qui n’ont pas de prénoms. Ceci est dû aux données brutes et à son codage.
- Biais de sélection : non appariement des auteurs avec des prénoms non occidentaux.