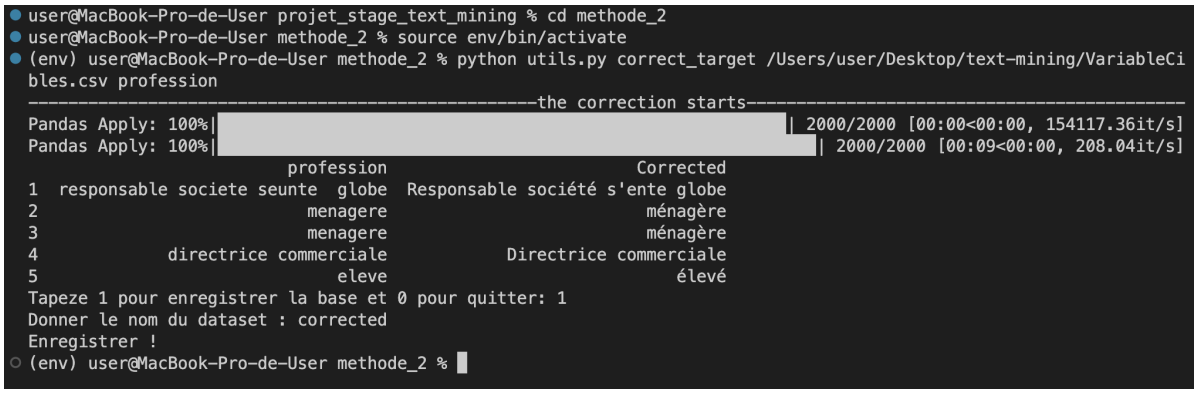
TEXT MINING PROJECT

- Pour lancer le projet il faut cloner le projet

```
https://github.com/Aida73/text-mining.git
```

- Activer l’environnement virtuel:
  - Ouvrir le projet cloné dans un terminal et se déplacer dans le dossier **methode_2**
  - Tapez **source env/bin/activate**
- Installer les packages nécessaires
  - **pip install -r requirements.txt**
- Corriger un dataset
  - Tapez dans le terminal ***python utils.py correct_target /path/of/csv/dataset target***
  - exemple : ***python utils.py correct_target /Users/user/Desktop/text-mining/VariableCibles.csv profession***

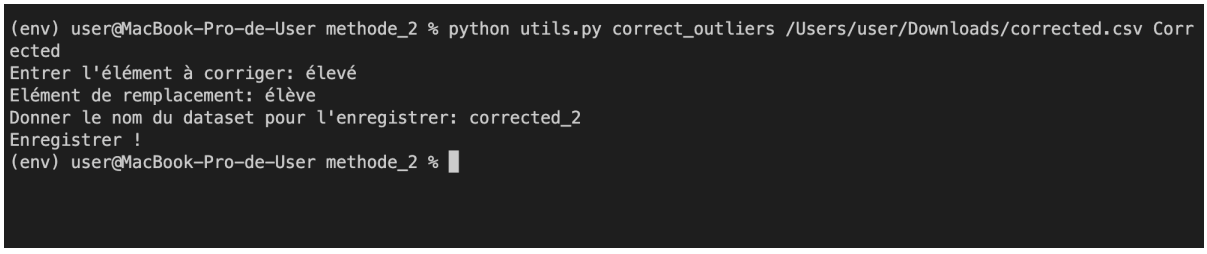


- On peut des fois rencontrer quelques corrections mal effectuées. Cependant on pourra toujours les corriger.

corrected.csv

	profession	Corrected
	responsable societe seunte globe	Responsable société s'ente globe
	menagere	ménagère
	menagere	ménagère
	directrice commerciale	Directrice commerciale
	eleve	élevé
	eleve	élevé
	chauffeur	chauffeur
	ouvrier agricole	Ouvrier agricole
)	agent du cadastre impots et domaine	Agent du cadastre impôts et domaine
l	autres	autres
2	pilote engin	Pilote engin
1	commercant	commercant

- Leur correction a été prise en compte. Pour ce faire:
  - Tapez sur le terminal ***python utils.py correct_outliers /path/to/corrected/dataset target***
  - Exemple: *python utils.py correct_outliers /Users/user/Downloads/corrected.csv Corrected*



```
(env) user@MacBook-Pro-de-User methode_2 % python utils.py correct_outliers /Users/user/Downloads/corrected.csv Corr
ected
Entrer l'élément à corriger: élevé
Elément de remplacement: élève
Donner le nom du dataset pour l'enregistrer: corrected_2
Enregistrer !
(env) user@MacBook-Pro-de-User methode_2 %
```

corrected_2.csv

	Unnamed: 0	profession	Corrected
0	1	responsable societe seunte globe	Responsable société s'ente globe
1	2	menagere	ménagère
2	3	menagere	ménagère
3	4	directrice commerciale	Directrice commerciale
4	5	eleve	élève
5	6	eleve	élève
6	7	chauffeur	chauffeur
7	9	ouvrier agricole	Ouvrier agricole
8	10	agent du cadastre impots et domaine	Agent du cadastre impôts et domaine
9	11	autres	autres
10	12	pilote engin	Pilote engin
11	14	commercant	commerçant
12	15	commercante femme au foyer	commerçante femme au foyer
13	16	etudiant droits des affaires	étudiant droits des affaires

- Trouver les categories: on va essayer de catégoriser la colonne corrigé. Ceci, en utilisant les catégories données par l'utilisateur. Il pourra donner autant de catégories qu'il voudra. Pour ce faire:
  - Tapez **python utils.py find_categories /path/of/corrected/dataset target**
  - Exemple : ***python utils.py find_categories /Users/user/Downloads/corrected_2.csv profession***

```
(env) user@MacBook-Pro-de-User methode_2 % python utils.py find_categories /Users/user/Downloads/corrected_2.csv pro
fession
Tapeze 1 ajouter une categorie: 1
categorie : ménagère
['ménagère']
Tapeze 1 ajouter une autre catégorie categorie ou 0 pour la classification: 1
categorie : enseignement
['ménagère', 'enseignement']
Tapeze 1 ajouter une autre catégorie categorie ou 0 pour la classification: médical
Vous ne pouvez choisir qu'entre 0 et 1 !
Tapeze 1 ajouter une autre catégorie categorie ou 0 pour la classification: 1
categorie : transport
['ménagère', 'enseignement', 'transport']
Tapeze 1 ajouter une autre catégorie categorie ou 0 pour la classification: 1
categorie : administration
['ménagère', 'enseignement', 'transport', 'administration']
Tapeze 1 ajouter une autre catégorie categorie ou 0 pour la classification: 0
['ménagère', 'enseignement', 'transport', 'administration']
```

```
----------------------------------------finding categories----------------------------------------
Pandas Apply: 100%|| 2000/2000 [00:00<00:00, 10619.96it/s]
Pandas Apply: 100%|| 2000/2000 [00:00<00:00, 74802.78it/s]
Pandas Apply: 100%|| 2000/2000 [00:00<00:00, 7507.40it/s]
Pandas Apply: 100%|| 2000/2000 [00:00<00:00, 24723.42it/s]
Pandas Apply: 100%|| 2000/2000 [00:00<00:00, 10071.20it/s]
Pandas Apply: 100%|| 2000/2000 [00:00<00:00, 21324.66it/s]
Pandas Apply: 100%|| 2000/2000 [00:00<00:00, 7035.96it/s]
Pandas Apply: 100%|| 2000/2000 [00:00<00:00, 24168.81it/s]
Pandas Apply: 100%|| 2000/2000 [00:00<00:00, 580727.45it/s]
Tapeze 1 pour enregistrer la base et 0 pour quitter: 1
Donner le nom du dataset : categorized
Enregistrer !
None                          1099
enseignement                   372
administration                 246
ménagère                       189
transport                       61
transport,administration        13
enseignement,administration     12
transport,enseignement           8
Name: Categorie, dtype: int64
(env) user@MacBook-Pro-de-User methode_2 %
```

	profession	Corrected	Categorie
	responsable societe seunte globe	Responsable société s'ente globe	administration
	menagere	ménagère	ménagère
	menagere	ménagère	ménagère
	directrice commerciale	Directrice commerciale	None
	eleve	élève	enseignement
	eleve	élève	enseignement
	chauffeur	chauffeur	transport
	ouvrier agricole	Ouvrier agricole	enseignement
	agent du cadastre impots et domaine	Agent du cadastre impôts et domaine	enseignement,administration
	autres	autres	None
	pilote engin	Pilote engin	None
	commercant	commerçant	None
	commercante femme au foyer	commerçante femme au foyer	ménagère
	etudiant droits des affaires	étudiant droits des affaires	enseignement
	auxilliaire de vie	auxiliaire de vie	None
	commercante	commerçante	None
	aide soignante	aide-soignante	None