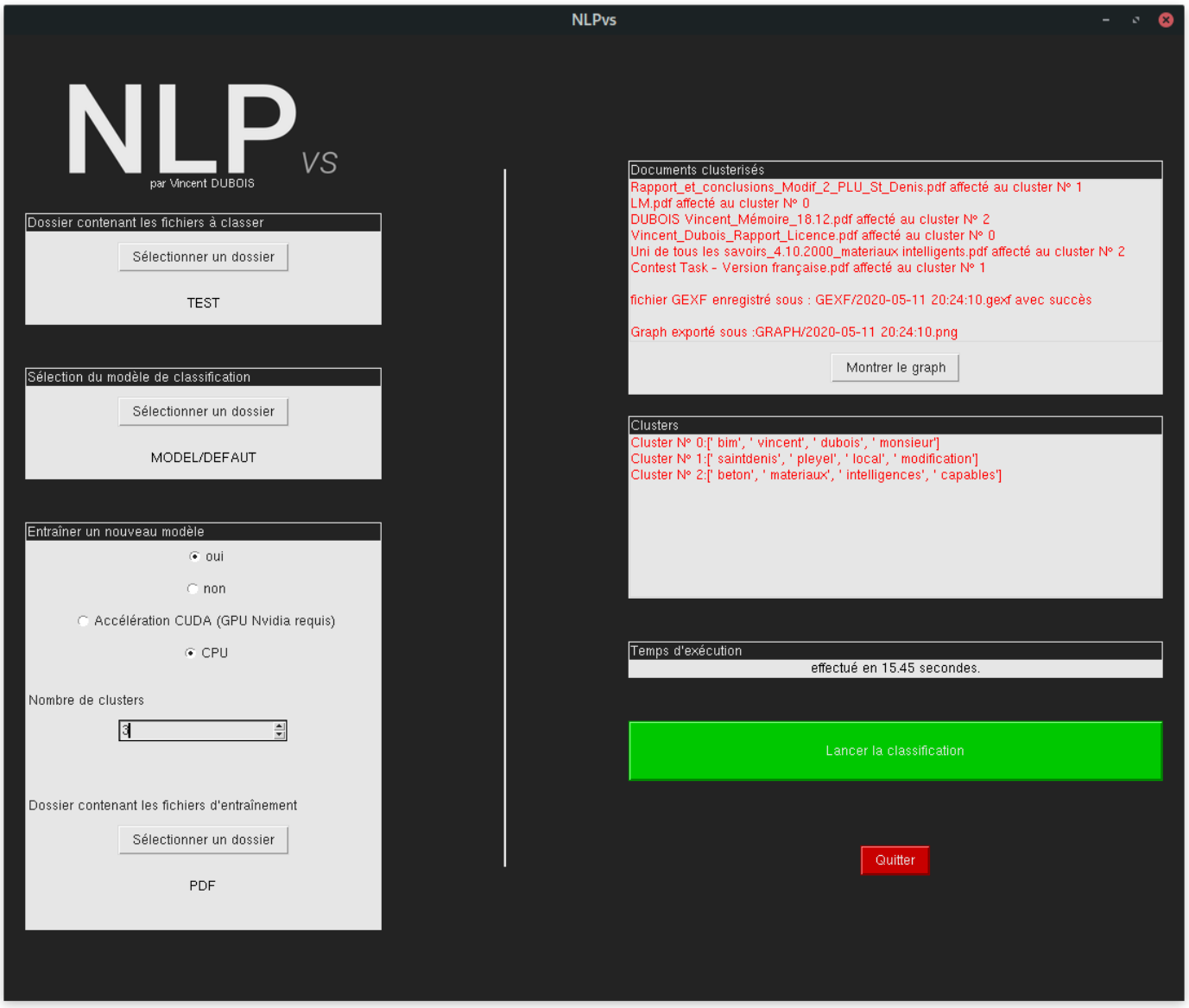
NLPvs

Outil de classification de documents au format PDF



Développé par DUBOIS Vincent dans le cadre d'une mission pour la chaire A.I. : Architecture Intelligence

ABSTRACT

NLPvs est un outil de classification non-supervisé de documents au format PDF.

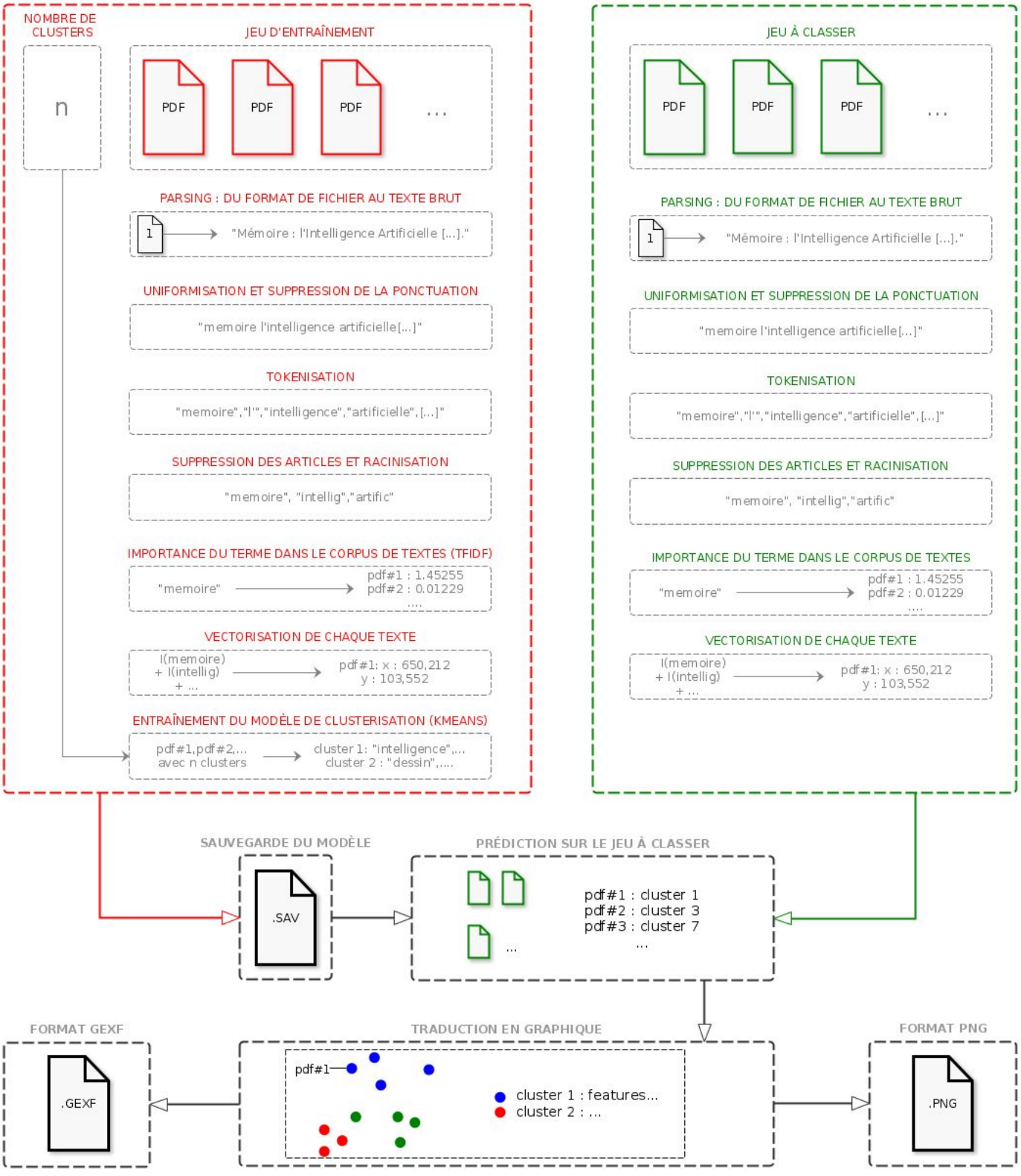
Pour un corpus de documents à classer défini par l'utilisateur, l'outil se charge de les répartir au sein de clusters selon le modèle de Machine Learning pré-entraîné choisi, et de restituer cette classification sous forme graphique montrant la proximité des documents entre eux ainsi que leurs répartition au sein des clusters.
Il est également possible d'entraîner son propre modèle, en spécifiant un corpus de documents d'entraînement ainsi qu'un nombre de clusters (avec une possibilité de l'accélérer grâce aux coeurs CUDA sur un GPU Nvidia). Après l'entraînement, ce modèle est automatiquement sauvegardé (pour permettre une réutilisation pour un autre jeu de documents) et procède à la classification du corpus de documents à classer.

Pour chaque corpus (entraînement ou classification) de PDF, un traitement initial est requis:

- "Parsing" des fichiers PDF : il s'agit ici d'extraire le contenu textuel de chaque fichier (sous forme de chaîne de caractères) et d'en supprimer toute mise en page.
- Uniformisation et suppression de la ponctuation de chaque texte.
- "Tokenisation" : découpage d'un texte en termes (mots) afin de créer une entrée de vocabulaire (un "jeton")
- Suppression des articles et "stop words" (verbes être et avoir, etc…) : élimination des termes les plus communs entre chaque texte afin de les alléger pour les étapes suivantes.
- "Racinisation" : réduction de chaque terme à sa "racine" en enlevant conjugaison, suffixe ou préfixe.
- Mesure du "Text Frequency - Inverted Frequency Index" (TF-IDF) : Estimation de la fréquence d'un terme (ou groupe de termes) à travers un corpus, sur la base de sa rareté au sein du corpus (moins un terme est fréquent, plus il a de chances d'être porteur de sens pour pouvoir différencier les textes entre eux).
- Vectorisation de chaque texte : Associe un vecteur à chaque texte sur la base des indices de fréquence.

Un tel corpus peut ensuite servir de base pour entraîner un modèle (basé sur l'algorithme de clusterisation Kmeans) afin d'en tirer des clusters propres à ce modèle puisque lié aux vocabulaire de termes du corpus d'entraînement (les features de ces clusters seront les termes les plus significatifs qui ont servi au modèle pour différencier les textes entre eux).

Le modèle est ensuite chargé de "prédire" auquel de ses clusters appartient chaque texte du corpus à classer. Enfin, la distance cosinus entre chaque texte vectorisé sera calculée, dans le but de générer un graphique légendé qui traduira la clusterisation et cette distance en un seul document au format PNG. Un document au format GEXF sera également produit de manière systématique.


NOMBRE DE CLUSTERS
n
JEU D'ENTRAÎNEMENT
PDF
PDF
PDF
...
PARSING : DU FORMAT DE FICHIER AU TEXTE BRUT
1
"Mémoire : l'Intelligence Artificielle [...]."
UNIFORMISATION ET SUPPRESSION DE LA PONCTUATION
"memoire l'intelligence artificielle[...]"
TOKENISATION
"memoire","l'","intelligence","artificielle",[...]"
SUPPRESSION DES ARTICLES ET RACINISATION
"memoire", "intellig","artific"
IMPORTANCE DU TERME DANS LE CORPUS DE TEXTES (TFIDF)
"memoire"
pdf#1 : 1.45255
pdf#2 : 0.01229
....
VECTORISATION DE CHAQUE TEXTE
l(memoire)
+ l(intellig)
+ ...
pdf#1: x : 650,212
y : 103,552
ENTRAÎNEMENT DU MODÈLE DE CLUSTERISATION (KMEANS)
pdf#1,pdf#2,...
avec n clusters
cluster 1: "intelligence",...
cluster 2 : "dessin",....
JEU À CLASSER
PDF
PDF
PDF
...
PARSING : DU FORMAT DE FICHIER AU TEXTE BRUT
1
"Mémoire : l'Intelligence Artificielle [...]."
UNIFORMISATION ET SUPPRESSION DE LA PONCTUATION
"memoire l'intelligence artificielle[...]"
TOKENISATION
"memoire","l'","intelligence","artificielle",[...]"
SUPPRESSION DES ARTICLES ET RACINISATION
"memoire", "intellig","artific"
IMPORTANCE DU TERME DANS LE CORPUS DE TEXTES
"memoire"
pdf#1 : 1.45255
pdf#2 : 0.01229
....
VECTORISATION DE CHAQUE TEXTE
l(memoire)
+ l(intellig)
+ ...
pdf#1: x : 650,212
y : 103,552
SAUVEGARDE DU MODÈLE
.SAV
PRÉDICTION SUR LE JEU À CLASSER
pdf#1 : cluster 1
pdf#2 : cluster 3
pdf#3 : cluster 7
...
...
FORMAT GEXF
.GEXF
TRADUCTION EN GRAPHIQUE
pdf#1
cluster 1 : features...
cluster 2 : ...
FORMAT PNG
.PNG