Présentation challenge AED 2018

Dusart, Nicolaieff, Njamo, Zouitine

Université Paul Sabatier

November 27, 2018

Sommaire

Introduction

- Base de données : 308 extraits issus de 15 films
- Outils utilisés : Langage Python, librairie : Pandas, numpy, opencv, pyFeel, Sklearn, plotly, Matplotlib
- Temps imparti : 2.5 semaines (c'est dire la performance !)
- Encadré par : Estelle Randria, Jim Petiot, Isabelle Ferrané, Jérôme Farinas, Julien Pinquier, Lynda Tamine-Lechani, José Moreno.

Objectifs

Approche

Globale

Analyser le contenu des données

Classifier les extraits en fonction de critères de similarité

Locale

Sous objectif par modalité :

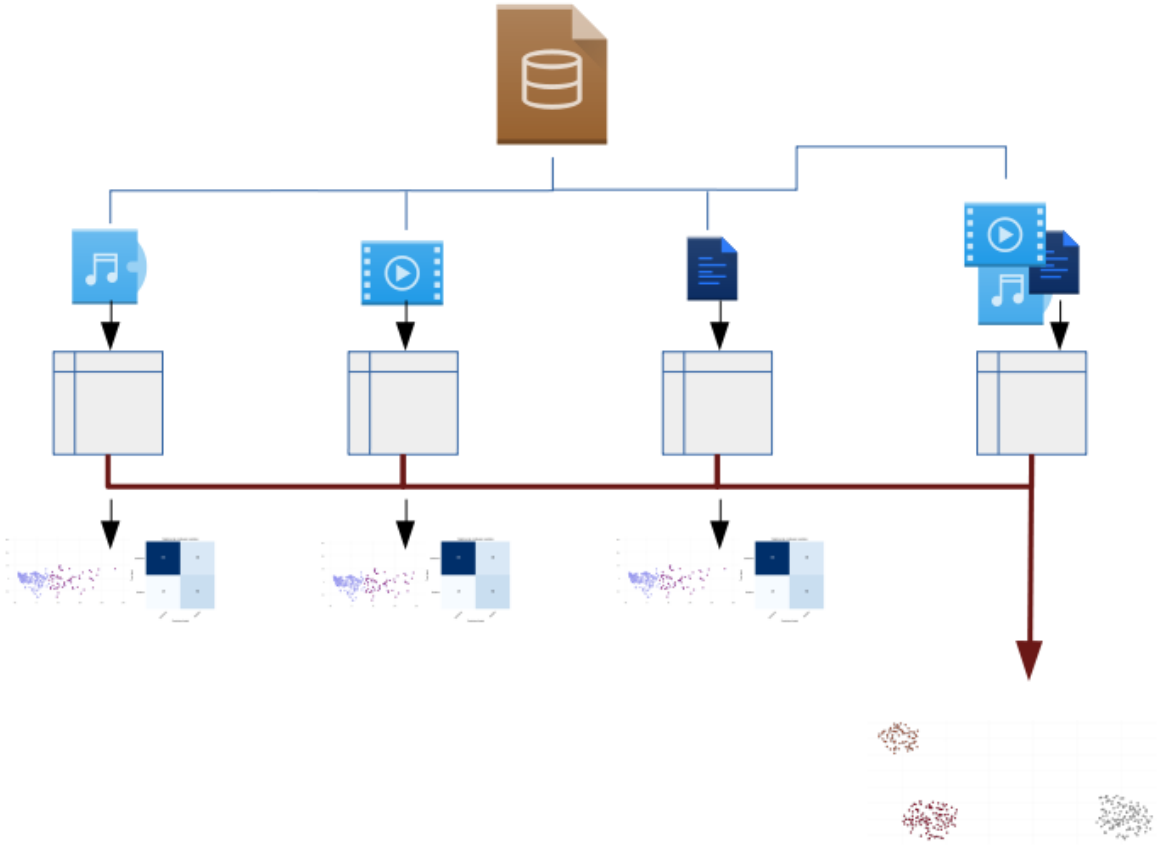
Texte: Extraire des thèmes/sentiments

Audio: Détection d'activité de parole, reconnaitre les bruits

Vidéo: Séparer les formes, reconnaitre des objets/lieux, couleurs

Objectifs

Structure



Présentation des données

- Base de données : 308 séquences de 15 films différents Pour chaque séquence on dispose de 3 modalités:
- Texte Format utilisé : `XML`. Taille des fichiers : 1.5 Mo
- Audio Format utilisé : `WAV`. Taille des fichiers : 459.5 Mo
- Vidéo Format utilisé : `MP4`. Taille des fichiers : 2.5 Go

Texte

TF-IDF (Term Frequency-Inverse Document Frequency)

Le TF-IDF est une méthode de pondération dans l'analyse de données textuelles.
Cela permet d'évaluer l'importance d'un terme contenu dans un document appartenant à un corpus.

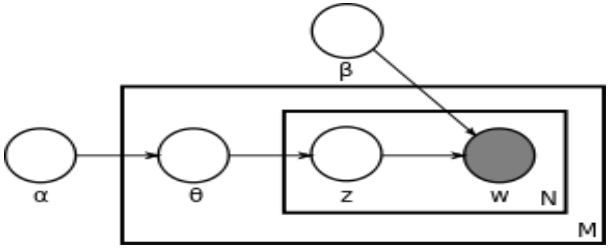
$$w_{i,j} = tf_{i,j} * log(\frac{N}{df_i})$$

$\text{tf}_{i,j}$: nombre d'occurences du mot i dans le texte j
df_i : nombre de documents contenant i
N : nombre total de documents

Méthodes

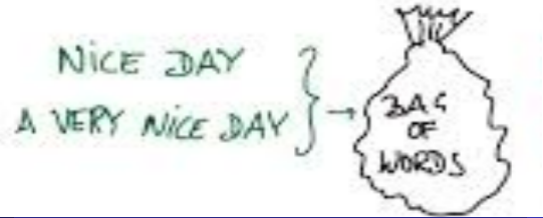
LDA(LatentDirichletAllocation)

Méthode générative probabiliste qui va attribuer à chaque mot du document un thème.



Méthode du Lexicon

FEEL: French Expanded Emotion Lexicon.
Calcul des sentiments



Audio

L'énergie du signal

Révèle l'alternance forte/faible puissance sonore du signal. Le pourcentage des trames à faible énergie permet de distinguer les blancs de la parole.

$$Energie = log(\sum_{i=0}^{n} x^2)$$

n : la taille de la fenêtre choisie pour découper le signal
x : un point du signal

Zcr (Zero Crossing Rate)

Le taux de passage à zéro de la forme d'onde temporelle révèle, par les brusques variations de son profil temporel, l'alternance voisée/non-voisée.

$$zcr = \frac{1}{T}\sum_{t=1}^{T} |s(t) - s(t-1)|$$

s : le signal
T : la longueur du signal
$s(t) = 1$ si le signal a une amplitude positive au temps t et 0 sinon.

AMDF(Average Magnitude Difference Function)

Fonction de distance utilisée ici pour estimer la fréquence fondamentale de l'audio.

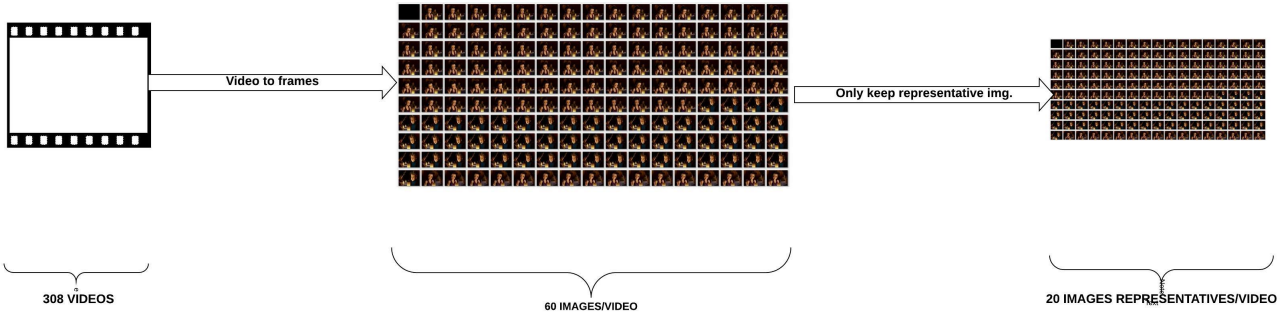
$$AMDF(\tau) = \sum_{n=0}^{N-1} |x(n) - x(n+\tau)|$$

- Utilisation d'une librairie **`'Snack'`** *qui calcule et renvoie automatiquement un fichier contenant la fréquence fondamentale pour chaque audio*

Méthodes

Vidéo

- **Features extraction : Video to Frames**

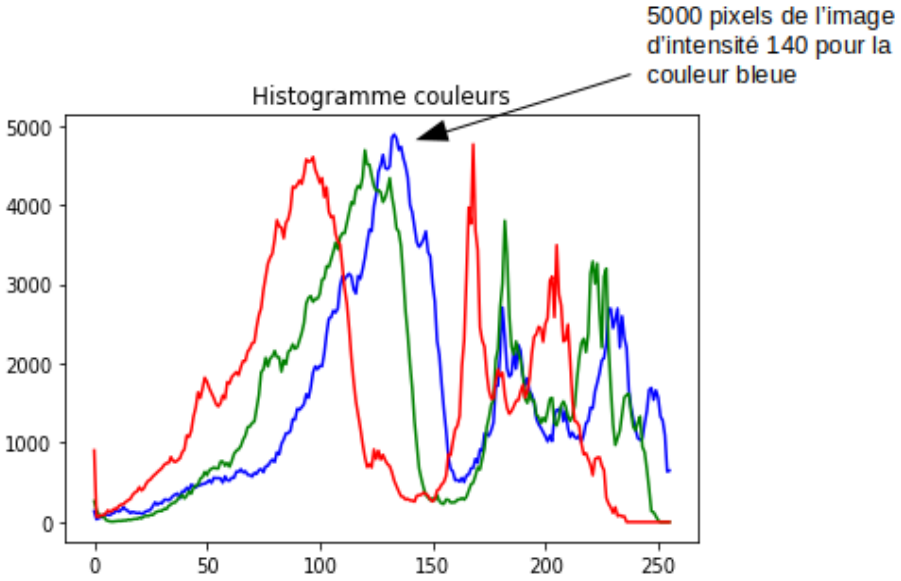


Description du processus

- Découper la vidéo en frames à raison d'une image par seconde.
- Regrouper toutes les images similaires en lot et garder une image pour chaque lot.
- Les images retenues sont stockées dans un fichier portant le No de la vidéo.
- Passer à la vidéo suivante et répéter les 3 points ci-dessus.

Méthodes

- Histogramme de couleur



Méthodes

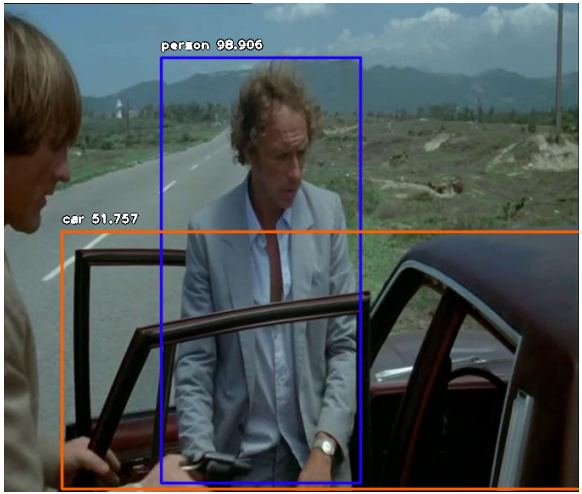
- Quantité de mouvement

Calcul

Déplacement d'un point entre deux images successives
Somme des "déplacements" de l'image

Méthodes

- Détection d'objets

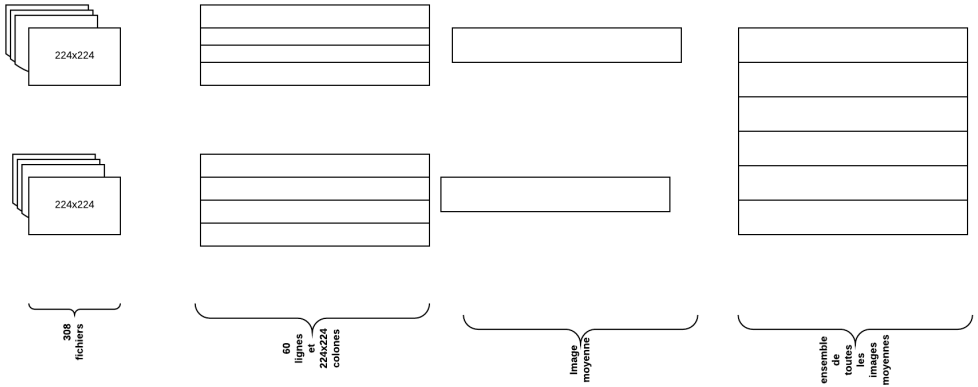


Approche

Détection d'objets à l'aide de réseaux pré-entrainés
Comparaison entre une liste des objets détectés dans l'image et une liste de ce que l'on recherche

Méthodes

- **Features extraction : Frames to Matrix**

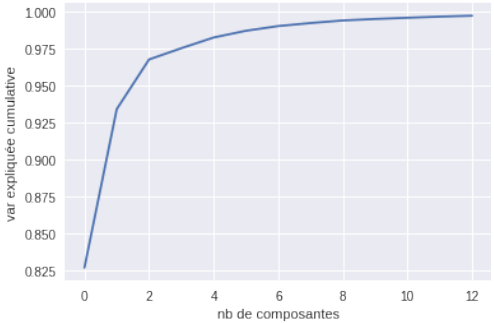


Description du processus

- Transformer toutes les images en niveaux de gris et convertir en vecteur ligne.
- Calculer l'image moyenne et transformer en vecteur ligne.
- Ajouter ce vecteur comme nouvelle ligne à la matrice **M**.
- Passer au fichier suivant et répéter les 3 points ci-dessus.

Méthodes

- **Features extraction : Reduction de dimension**

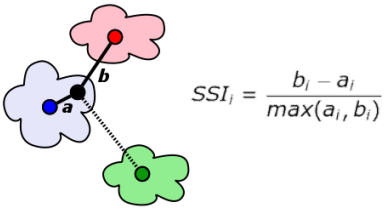


Analyse en Composantes Principales ACP

- La matrice **M** de dimension *308*x*50176* constitue notre base d'apprentissage.
- On applique une ACP sur **M** et avec **12** composantes, on retrouve la totalité des informations initiales.
- La base d'apprentissage est donc reduite à *308*x*12*.

Méthodes

- **Métrique : silhouette score**



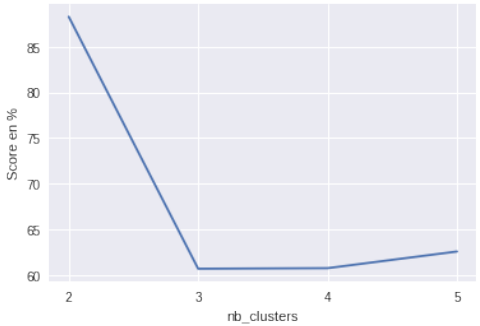
Définition et calcul

C'est une mesure d'interpretation et de validation de la cohérence au sein des clusters. *a(i)* est la distance moyenne entre i et tous les autres points du cluster et *b(i)* est la plus petite distance moyenne de i à tous les points de tout autre groupe.

$$-1 \leq \mathbf{S} = \frac{\sum_{i=1}^{308} SSI_i}{308} \leq 1$$

Les données sont bien séparées lorsque **S** est proche de 1 et mal séparées lorsqu'il est proche de -1.

KMeans : Evaluation de la dispersion des clusters



En appliquant le modèle de KMeans sur la matrice **M** tout en faisant varier le nombre de cluster entre 2 et 154, On s'est rendu compte qu'avec deux clusters, on a un score de sihouette maximal (S = 0.9). Donc une bonne dispersion entre les clusters.

$$\underset{\mathbf{k}}{\text{argmax}} \sum_{k=2}^{154} \mathbf{S}(kmeans(n\_cluster{=}k))$$

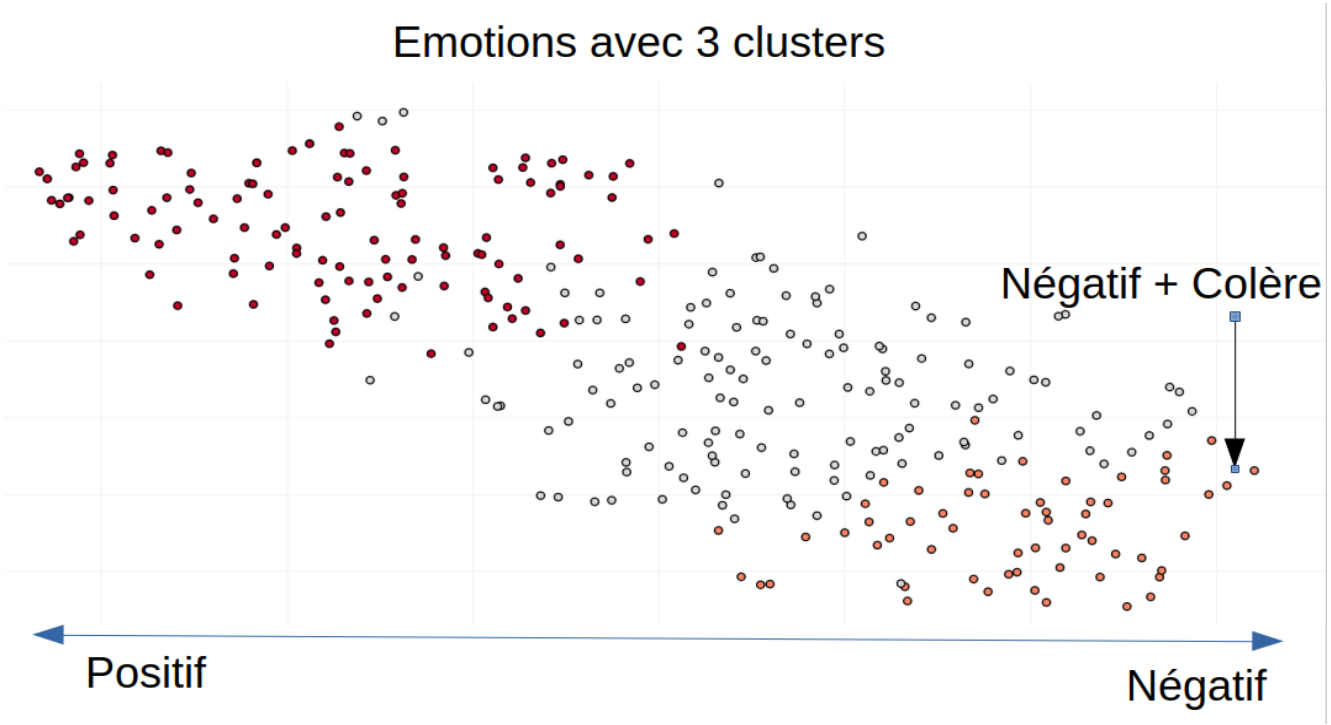
Méthodes

Méthode utilisée pour la fusion

- PCA : reduire la dimension des descripteurs en minimisant la perte d'information.
- TSNE : pour visualiser en 2 dimensions et respecter la distance euclidienne entre les points.

Résultats

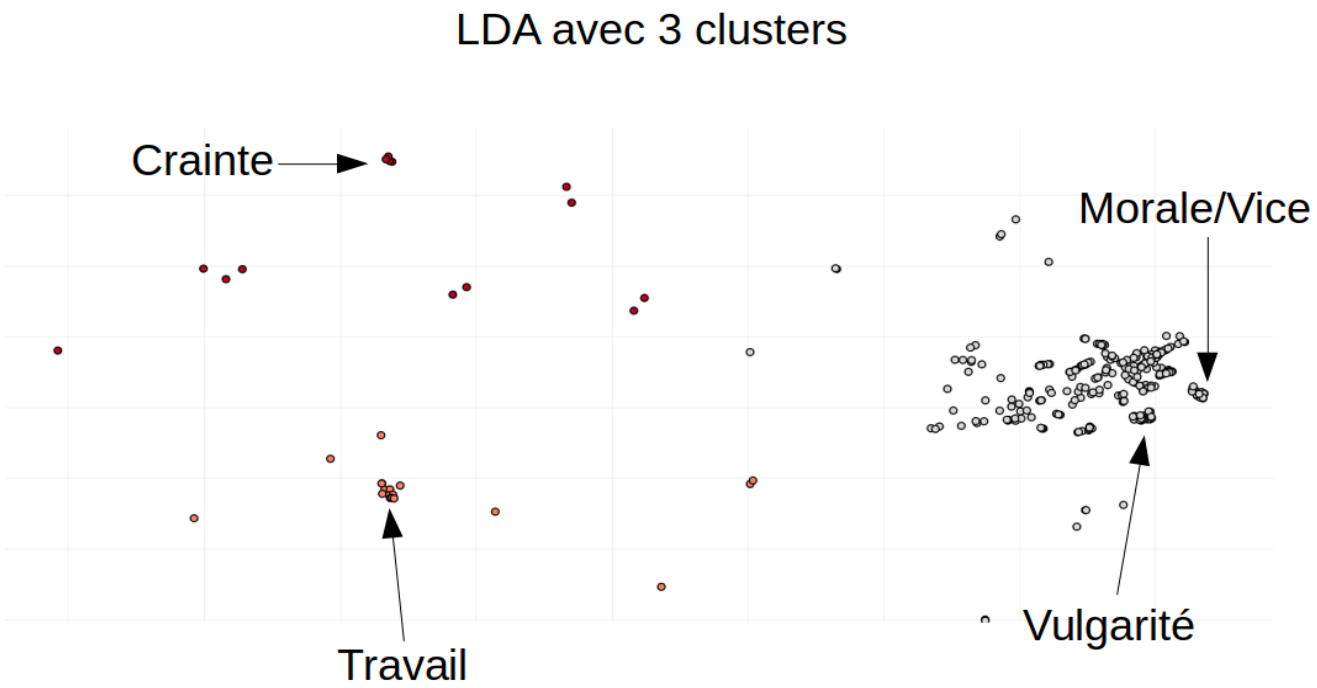
Texte



Résultats

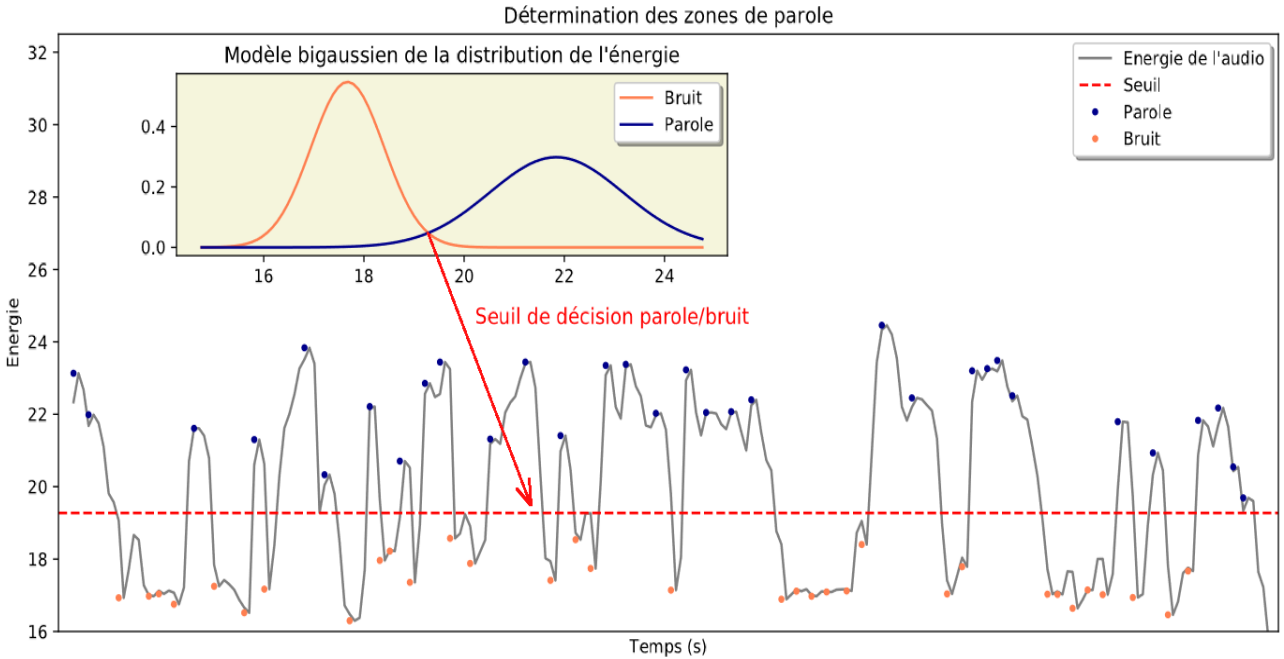
Nuage de mots du thème 'Sport' trouvé dans la LDA

Résultats



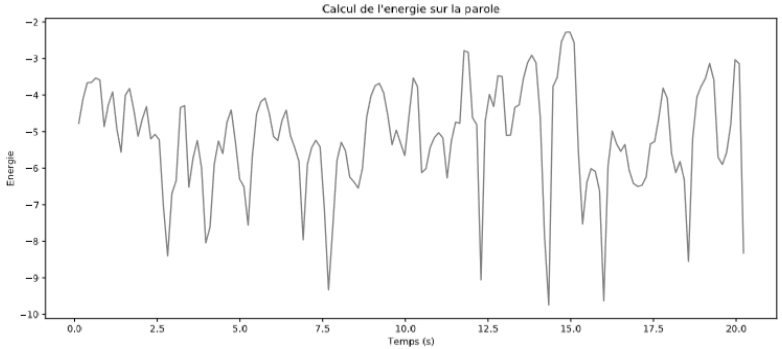
Audio

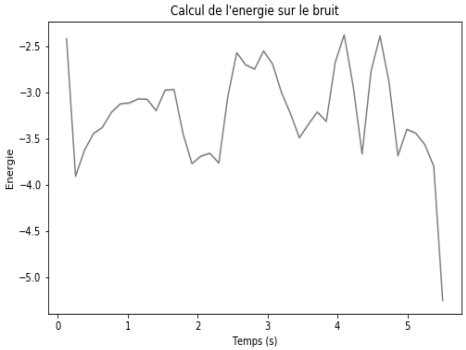
- Détecteur d'activité vocale

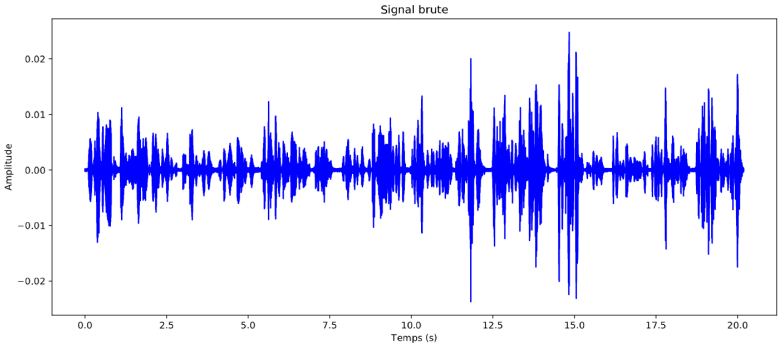


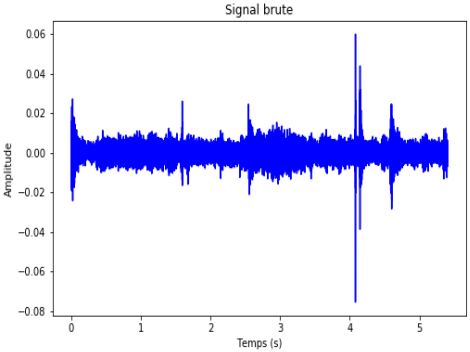
Résultats

- L'énergie du signal de l'audio 104 coupée du bruit









Résultats

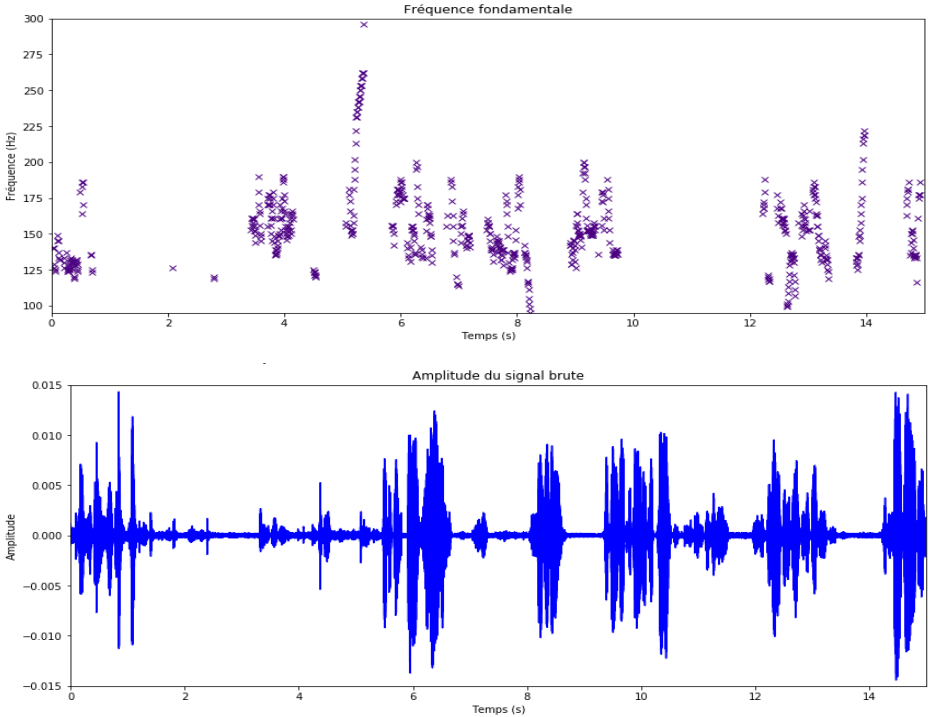
- Taux de parole :

$$Taux = \frac{longueur\ du\ signal\ de\ parole}{longueur\ du\ signal\ initial}$$

  Renseigne sur la quantité de parole présente dans chaque audio. En moyenne il y a 62% de parole dans une séquence. Le taux de parole pour chaque audio varie entre 26% et 94%. Les audios sont répartis uniformément.
- Zcr : Méthode appliquée sur le bruit (la non parole), afin d'obtenir une distinction entre les scènes calmes et les scènes bruyantes.

Résultats

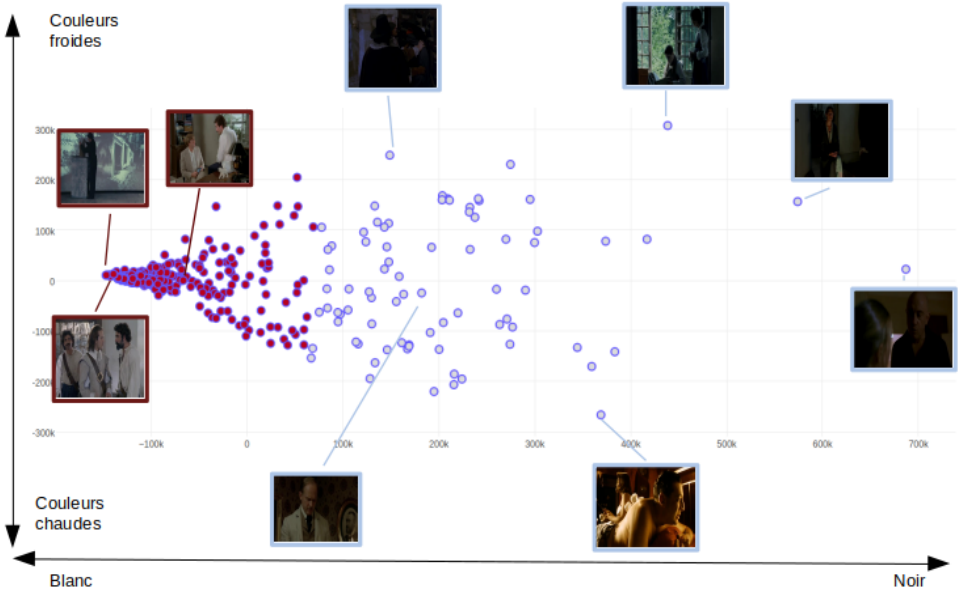
- Fréquence fondamentale



Résultats

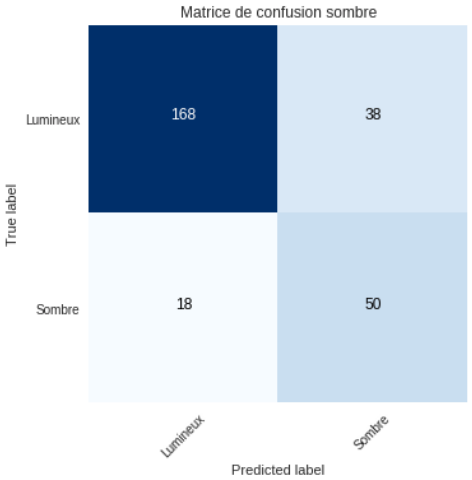
Vidéo

- Histogrammes de couleurs



Résultats

- KNN sombre/lumineux histogrammes de couleurs



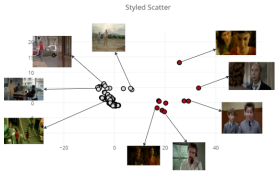
50/68 = 73% de séquences sombres bien détectées
168/206 = 81% de séquences lumineuses bien détectées

Résultats

KMeans : 2 clusters

Résultats

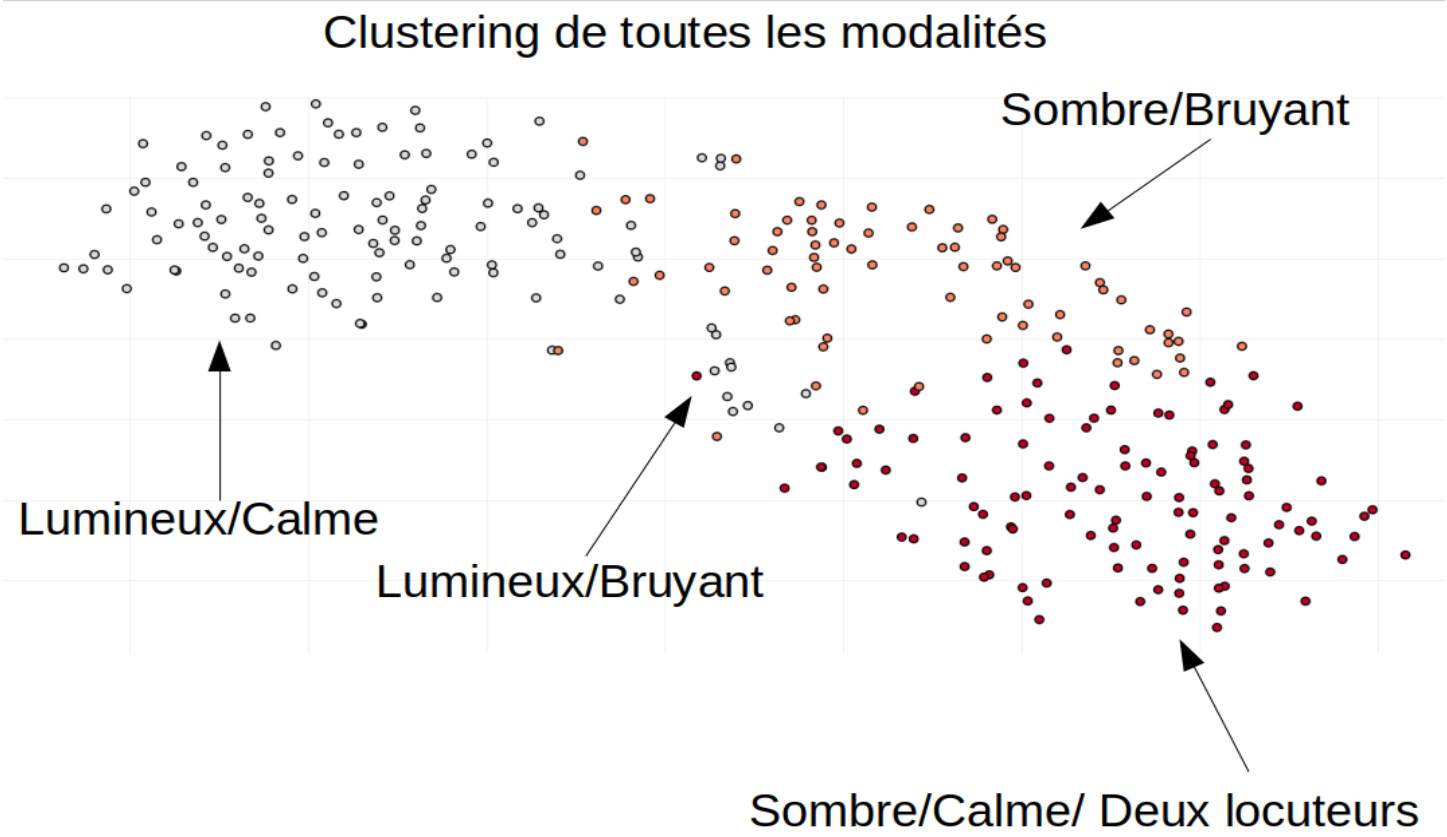
KMeans : 2 clusters



Interpretation

- Les 308 vidéos sont dispatchées en **deux groupes**(rouge et bleu ciel).
- En bleu, les vidéos avec un plan appelé **plan d'ensemble** et qui permet de situer les personnages(dans un jardin, dans la rue, dans un bureau,etc.)
- En rouge, les vidéos avec un plan appelé plan moyen qui permet de zoomer sur les personnages afin de lire l'émotion sur leurs visages.

Fusion


Clustering de toutes les modalités
Sombre/Bruyant
Lumineux/Calme
Lumineux/Bruyant
Sombre/Calme/ Deux locuteurs

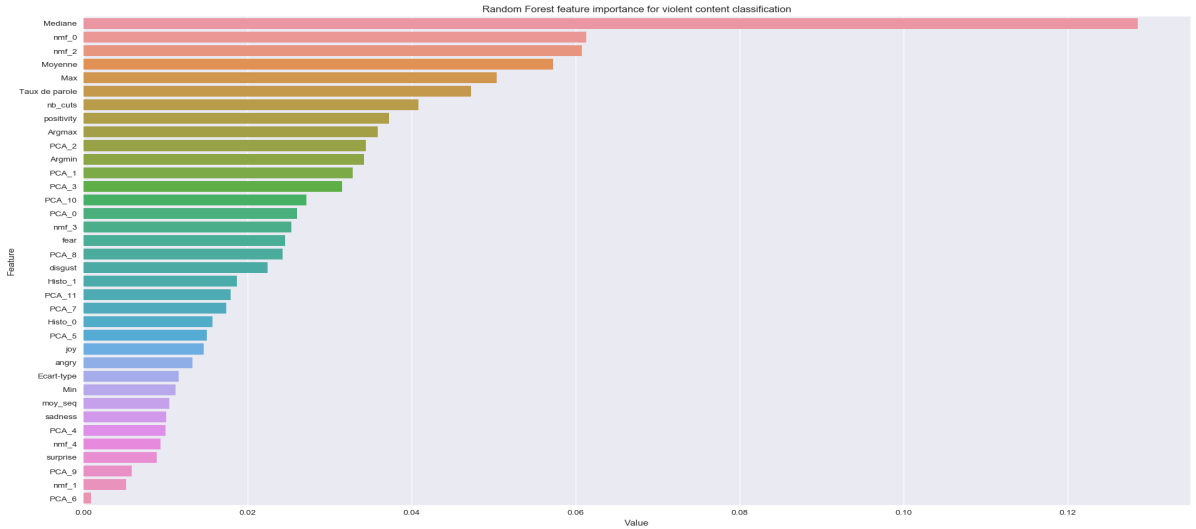
Fusion

- Classification supervisée : Random Forest

Résultats

Violent/Non violent : Score de précision 66.2%
Intérieur/Extérieur : Score de précision 57.5%

Importance des descripteurs pour Violent/Non violent:



Bilan

Mise en place d'un système d'indexation à l'aide du texte, de l'audio et de la vidéo. Le texte nous a permis d'indexer selon des thèmes. L'analyse de l'audio de détecter les enregistrements avec beaucoup de parole. La vidéo d'indexer selon la luminosité de la séquence. La fusion des modalités texte, audio et vidéo permet dans une moindre mesure d'indexer selon la violence.

Perspectives

Audio

La fréquence fondamentale moyenne :
Le F0 moyen diffère entre les voix d'hommes et les voix de femmes.
Valeurs moyennes hommes = 120 Hz
Valeurs moyennes femmes = 240 Hz
= création d'un detecteur: H/F
Il faudrait également étudier l'âge du locuteur pour voir si il joue un rôle sur le F0 moyen des hommes et des femmes.

Perspectives

Vidéo

La détection d'objets a permis de n'identifier que 20 séquences extérieures sur 111 (=18%). Quand plus d'objets seront détectables (entrainer un réseau avec des images de nuages ou d'arbres par exemple) cette méthode pourra être très performante.

Perspectives

Texte

Utiliser un TF-IDF en faisant varier les bi-grams aurait pu améliorer les performances car certaines combinaisons de mots ont un sens complètement différent.

Utiliser des méthodes telles que les embeddings aurait pu nous permettre d'extraire des descriteurs nous apportant de nouvelles informations sur la représentation des mots/répliques/extraits dans un espace sémantique.