**1. Text Mining est le processus consistant à (a)\_\_\_\_\_ à partir de (b)\_\_\_\_\_**

(a) extraire des informations ; (b) des sources textuelles

(a) organiser des phrases ; (b) des sources désordonnées

(a) obtenir des nombres ; (b) des textes

(a) construire des modèles ; (b) des documents structurés

**2. À quoi sert l'exploration de texte ?**

Rechercher facilement des données diverses et utiles.

Reconstituer un texte de mauvaise qualité en un texte de qualité.

Rechercher efficacement un document, extraire les mots similaires et souligner les répétitions.

Identifier et extraire des schémas d'une grande quantité de données cachées et non structurées.

**3. Quel est le processus de travail pour le text mining ?**

Collecter, prétraiter, indexer, extraire, analyser.

Exploiter, prétraiter, collecter, indexer, analyser.

Collecter, indexer, prétraiter, extraire, analyser.

Collecter, extraire, prétraiter, indexer, analyser.

# **PROJET TEXT MINING**

# Projet : Analyse de commentaires de films

**PLAN :**

**1. L'analyse exploratoire des données**

* Lecture du Jeu de Données
* Visualisation des mots-clés positifs et négatifs des critiques de films et des nuages de mots
* Les 25 meilleurs mots pour exprimer des sentiments positifs et négatifs

**2. Nettoyage et prétraitement des données**

* Minuscules
* Supprimer les balises HTML
* Supprimer les ponctuations
* Supprimer les StopWords

**3. Décoder les critiques de films : ce que les mots et les expressions révèlent sur les critiques**

* Trigrammes Courants dans les Critiques Positives
* 5-Grammes Courants dans les Critiques Positives
* Trigrammes Courants dans les Critiques Negatives
* 5-Grammes Courants dans les Critiques Negatives

**4. Partition en corpus d’apprentissage et de test**

* Un vectoriseur TF-IDF
* Fréquence des termes

**5. Modélisation à l’aide de la régression logistique**

* Logisitic regression
* Amélioreration de model avec STEMMER
* Amélioration de model avec Lemmatizer

**5. Interprétation des résultats – Influence des termes dans le modèle**

**6. Déploiement du modèle en testant deux nouvelles critiques**