Natural Language Processing

# Data Science Life Cycle:

1. **Business Understanding (Compréhension du Métier) :**
   * Dans cette phase initiale, l'équipe de science des données collabore étroitement avec les parties prenantes pour comprendre les objectifs, les buts et les exigences de l'entreprise. Il s'agit de définir le problème ou l'opportunité que la science des données peut résoudre.
2. **Data Collection (Collecte de Données) :**
   * L'équipe rassemble des données pertinentes auprès de diverses sources. Cela peut impliquer l'acquisition de données à partir de bases de données, d'API, de jeux de données externes ou d'autres référentiels. L'objectif est de collecter des données complètes et de haute qualité qui sont pertinentes pour le problème à résoudre.
3. Public Data (Données Publiques) :

Les données publiques font référence à des informations accessibles au grand public sans restrictions particulières. Elles sont généralement disponibles gratuitement et peuvent être consultées, utilisées, ou distribuées par n'importe qui. Ces données peuvent provenir de diverses sources, telles que les gouvernements, les organisations non gouvernementales (ONG), les institutions académiques, et d'autres entités qui choisissent de partager des informations dans l'intérêt public. Les ensembles de données publics sont souvent utilisés à des fins de recherche, d'analyse, de développement d'applications, et plus encore.

1. Private Data (Données Privées) :

Les données privées sont des informations qui sont sensibles ou confidentielles et qui ne sont pas destinées à être partagées ou accessibles au public en général. Ces données sont généralement soumises à des restrictions d'accès strictes afin de protéger la vie privée, les droits et les intérêts des individus ou des organisations auxquels elles appartiennent. Les données privées peuvent inclure des informations personnelles, financières, médicales, commerciales, et d'autres types de données sensibles. L'accès à ces données est souvent limité à des personnes autorisées dans le cadre de politiques de confidentialité et de sécurité.

1. Open Data (Données Ouvertes) :

Les données ouvertes sont des informations accessibles au public de manière libre et gratuite, avec peu ou pas de restrictions d'utilisation, de redistribution ou de modification. L'objectif des données ouvertes est de favoriser la transparence, la collaboration et l'innovation. Les gouvernements, les organisations et d'autres entités peuvent choisir de publier des ensembles de données ouvertes pour encourager la participation du public, faciliter la création d'applications, soutenir la recherche, et promouvoir la responsabilité. Les principes des données ouvertes incluent souvent la disponibilité en ligne, la facilité d'accès, la réutilisation, et l'utilisation de formats standardisés pour faciliter l'interopérabilité.

1. **Data Preparation (Préparation des Données) :**
   * Les données brutes sont souvent désordonnées ou incomplètes. Dans cette phase, les données sont nettoyées, transformées et formatées pour les rendre adaptées à l'analyse. Cela peut impliquer le traitement des valeurs manquantes, la suppression des valeurs aberrantes et la conversion des types de données.
2. **Exploratory Data Analysis (Analyse Exploratoire des Données - AED) :**
   * L'AED est le processus d'exploration visuelle et statistique des données pour découvrir des tendances, des anomalies et des relations. Cela permet de mieux comprendre la nature des données et guide les analyses ultérieures.
3. **Data Modeling (Modélisation des Données) :**
   * Dans cette phase, différents modèles d'apprentissage automatique sont sélectionnés, entraînés et testés à l'aide des données préparées. L'objectif est de construire un modèle capable de faire des prédictions ou des classifications précises basées sur les motifs identifiés lors de l'analyse exploratoire des données.
4. **Data Evaluation (Évaluation des Données) :**
   * Les performances du modèle sont évaluées à l'aide de mesures telles que la précision, le rappel et le score F1. Cette étape garantit que le modèle répond aux objectifs de l'entreprise et fournit de la valeur.
5. **Model Deployment (Déploiement du Modèle) :**
   * Une fois que le modèle est validé, il est déployé dans un environnement de production. Cela implique l'intégration du modèle dans les systèmes ou processus existants, le rendant accessible aux utilisateurs finaux, tout en assurant sa surveillance et sa maintenance continues.

Ces phases forment collectivement un processus cyclique, car les insights tirés du déploiement du modèle peuvent conduire à des refinements ultérieurs dans les itérations suivantes du cycle de vie de la science des données.