**Dédicaces**

*Nous dédions ce modeste travail à nous même,*

*Et à Monsieur Ferid Helali qui nous beaucoup encouragé et soutenue.*

*A tous ceux qui nous connaissent et contribuent du près ou de loin à entamer ce travail et auxquels nous devons être reconnaissant.*

*Ahlem Mhamdi & Rihem Zairi*

**Remerciements**

*D'emblée nous tenons à remercier DIEU le tout puissant de nous avoir aidé à entamer et à terminer ce travail.*

*Nous devons être reconnaissante en remerciant infiniment nos encadrant universitaire* ***Mr Mohamed Bjaoui*** *ainsi que nos encadrant de stage* ***Mr Ferid Helali*** *qui nous ont beaucoup aidé et soutenu pour que ce travail soit bien organisé en nous mettant sur la bonne voie sur tous les plans et que sans eux ce travail ne pourrait voir le jour.*

*Tous nous remerciements les plus sincères à toute personne qui nous aidé de près ou de loin pour réaliser et présenter ce projet dans de bonnes conditions.*

Sommaire

[INTRODUCTION GENERALE 6](#_Toc82375375)

[Chapitre 1 : 8](#_Toc82375376)

[Cadre Général du Projet 8](#_Toc82375377)

[INTRODUCTION 9](#_Toc82375378)

[I. PRESENTATION DE LA SOCIETE 9](#_Toc82375379)

# INTRODUCTION GENERALE

Les voitures autonomes, ou machines parlant et agissant comme des humains, ont toujours occupé une place privilégiée dans la science-fiction. Aujourd'hui, la réalité rattrape l'imaginaire.

Plusieurs applications que nous pouvons commander par la voix nous aident à trouver le chemin optimal pour éviter embouteillage, ou nous informer des meilleures offres d'un même produit dans toute une région. Ces applications sont le résultat des avancées réalisées dans le domaine de l'intelligence artificielle. Ces avancées ne se sont pas produites du jour au lendemain. Il y a d'abord eu la naissance du Big Data.

En fait, l'évolution d'Internet et de l'Internet des objets (IOT) a ouvert la voie à nombreuses activités commerciales et sociales qui ont ouvert la voie à l'ère du Big Data, à chaque moment, dans le monde, plus d'une centaine de millions d'emails sont envoyés. Chaque minute, les moteurs de recherche comme Google enregistrent des millions de requêtes différentes sur leurs moteurs de recherche, et sur les réseaux sociaux comme Facebook et Twitter autant de posts et de réactions. Commerce électronique des transactions d'une valeur de 85000$ sont effectuées, cette tendance ne cesse d'augmenter. Chaque étape génère des données, utilisables par les entreprises, mais impossibles à réaliser manuellement.

La nécessité d'automatiser l'extraction de valeur à partir de l'énorme quantité de données stimulées recherche dans les domaines de l'apprentissage automatique et de l'apprentissage profond. Avec l'accomplissement de la loi de Moore, qui a déclaré que les progrès informatiques augmenteraient approximativement du double tous les deux années, il est devenu possible d'exploiter le flux de données du Big Data.

L 'IA est devenue le sur-ensemble pour plusieurs domaines, tels que la vision par ordinateur et Traitement du langage naturel. Le premier est préoccupé par la façon dont les machines peuvent voir le monde comme le font les humains, ce dernier par la façon dont les machines comprennent la parole humaine par écrit ou en format parlé.

Dans ce projet, nous visons à développer une solution intelligente pour prendre les séances de thérapies qui s'appuient sur la puissance du Deep Learning afin d'offrir la meilleure recommandation dans une mode plus humaine. . Au final, la solution sera vue comme un agent conversationnel qui est capable de parler à une personne curieux et qui permet de connais son état psychologique et de prend les recommandations nécessaire pour être en bonne santé.

Feuille de route

Dans ce document, nous couvrirons l'ensemble du processus du projet.il est organisé comme suit:

Le chapitre 1 Contexte du projet donne un aperçu du contexte du projet. On commence par présentant l'entreprise d'accueil du stage, nous fournissons ensuite l'énoncé du problème, brièvement décrire la solution et discuter de la méthodologie de gestion de projet de choix ainsi comme technologies de choix. A la fin, nous donnons le calendrier du projet.

Le chapitre 2 Contexte couvre les concepts de base et les principes fondamentaux des sujets d'intérêt et les concepts et technologies de pointe. Nous commençons par explorer le concept de systèmes de recommandation, puis nous passons en revue les technologies d'apprentissage en profondeur et sémantiques, et terminons le chapitre par une revue du traitement du langage naturel. Au Enfin, nous présentons nos choix pour le modèle qui convient à nos exigences.

Le chapitre 3 Solution proposée décrit la solution proposée du chat bot et comment il travaux.

Le chapitre 4 Implémentation selon Scrum Framework décrit les étapes de tous les étapes pour créer la solution

Dans la conclusion, nous évaluons nos réalisations, discutons de la marge d'amélioration et travail futur.

# 

# Chapitre 1 :

# Cadre Général du Projet

# INTRODUCTION

L'objectif de ce chapitre est de définir le contexte global du projet. Premièrement, nous présenter la société de stage Alfa Computer & Consulting, nous donnons l'énoncé du problème et comment le résoudre, y compris la méthodologie de gestion de projet que nous adoptons pour gérer le processus et présenter le calendrier du projet.

# PRESENTATION DE LA SOCIETE

Dans le cadre de notre stage de perfectionnement, nous avons réalisé notre stage au sein de la société Alfa Computer & Consulting dans le but d’enrichir nos connaissances théoriques par l’aspect pratique.

Alfa Computers est une agence des services numériques spécialisé en solutions interactives, en création des sites internet, e-commerce, multimédia, design, développement et hébergement situé à l’avenue Bechir Sfar 5100 Mahdia, Tunisie. Alfa computers a débuté son activité l’année 2009.

# CADRE DE PROJET

# Preuve de concept

# Dans cette section, nous décrivons le projet à long terme ainsi que le module que nous allons construire. Ensuite, nous formulons l'énoncé du problème et fixons les objectifs.

# Emily : Présentation

# Le projet à long terme est de construire un psychologue qui peut agir en tant que mentor ou conseiller d'ami virtuel dans de nombreuses situations, allant de l'assistance pour apporter des solutions aux problèmes émotionnels.

# Le premier pas vers la construction d’Emily, est de construire un système de conversation qui sémantiquement comprend la signification de l'entrée de la personne curieux.

# Énoncé du problème

# Le succès des logiciels de conversation, également appelés systèmes de chat bot, dépend fortement de les règles de saisie et la qualité des réponses de la machine. Le but est d'avoir un système qui permet à l'utilisateur de formuler ses pensées, ses souhaits ou simplement de parler sans règles de formatage, tout en étant capable de détecter ce qu'il veut et de donner une réponse précise. En ordre pour capturer le sens caché d'une entrée utilisateur, les techniques de traitement du langage naturel être utilisé.

# Objectif du projet

# Notre projet consiste à créer un chat bot qui agit comme un psychologue pouvant fournir des conseils à des personnes. La pierre angulaire du chat bot est l'analyse sémantique de la saisie de la personne curieuse. Lorsque la personne curieuse tape un message, le système détecte ce qu'il veut et suggère la meilleure recommandation. Ceci doit être réalisé grâce à un traitement du langage naturel à la pointe de la technologie. La personne curieuse doit être à l'aise avec la communication avec le système, n'a pas besoin d'utiliser de règles pour exprimer ses idées, mais plutôt de parler, et obtient une réponse précise.

# Étapes de réalisation du projet

# Pour la mise en œuvre de la solution, nous ferons une étude approfondie sur les solutions existantes et sur les techniques les mieux appréciées afin d'atteindre notre objectif. Le choix de la technique se fera en créant un prototype avec plusieurs techniques, et la décision de choix se fera sur la technique qui aura livré les meilleurs résultats.

# Les étapes peuvent être résumées sous forme de flux :

# Trouver la meilleure méthodologie pour mener le projet

# Étude comparative des solutions de pointe pour tirer parti des résultats pleinement atteints ;

# Construire des prototypes avec différentes techniques, y compris des approches de pointe pour voir ce qui donne le meilleur résultat dans la portée du projet ;

# Construire une ontologie pour le domaine d'application ;

# Construisez le prototype final avec les outils de votre choix

# Processus de gestion de projet

# Pour mener notre projet, nous comparons dans cette section trois processus de gestion de projet qui sont couramment appliqués aux projets d'analyse de données, et nous en adopterons un en fonction de son adéquation à nos besoins. Étant donné que ces processus sont itératifs, nous adapterons la méthodologie de choix à Scrum agile pour mieux nous adapter aux changements de décision.

# Cadre Scrum

# Scrum est une méthodologie agile itérative et incrémentale. Il repose sur un travail d'équipe et un suivi continu à travers des réunions quotidiennes et périodiques. De plus, Scrum permet d'organiser parfaitement le travail, notamment lorsqu'il s'agit de projets de développement. Cette se fait par la distinction entre les membres de l'équipe en attribuant à chacun un rôle et la répartition des tâches à réaliser en sprints.

Dans un contexte SCRUM, je vais devoir utiliser quelques termes propres à cette méthodologie. En voici les plus pertinent:

Tableau 1 : Réunions SCRUM

|  |  |
| --- | --- |
| **Réunions** | **Description** |
| Planification de sprint | Il s'agit d'une réunion tenue le premier jour du sprint et en lequel le Product Owner présente la liste des fonctionnalités à effectué au cours de ce sprint. Puis l'équipe et la mêlée maître choisir parmi ces fonctionnalités celles qu'ils s'engager à effectuer jusqu'à la fin du sprint par fixant la durée et la capacité nécessaire à chaque fonction nationalité. |
| Mêlé quotidienne | Le mêlé quotidienne est une réunion intermédiaire qui intervient en cours de sprint. Elle réunit les membres de l'équipe de développement. Objectif : faire le point sur les tâches réalisées la veille, et celle prévue pour la journée. |
| Rétrospective Sprint | La rétrospective de sprint intervient dans la foulée de chaque sprint. Les utilisateurs métier n'y sont généralement pas invités. Elle offre à l'équipe de développement un espace d'échange pour tirer les enseignements du sprint, planché sur des axes d'amélioration des processus et outils. C'est aussi l'occasion de revenir sur les relations entre les membres de l'équipe et les problèmes éventuellement rencontré. |

Tableau 2 : L'équipe d'un projet SCRUM

|  |  |
| --- | --- |
| **Rôle** | **Mission** |
| SCRUM Master | Supervision de l’avancement du projet et des activités de l’équipe |
| Propriétaire du produit (Product Owner) | Présentation des caractéristiques et des fonctionnalités du produit à développer et approbation du produit à livrer |
| L’équipe de développeurs | Réalisation des user stories et élaboration des sprints |

Tableau 3 : Artéfact SCRUM

|  |  |
| --- | --- |
| **Artéfact** | **Description** |
| **Backlog du produit** | La définition des besoins fonctionnels sous forme de (user story) |
| **Backlog du Sprint** | la liste des tâches à implémenter dans un sprint, classées par importances et état |
| **Produit partiel** | la liste des tâches à implémenter dans un sprint, classées par importances et état |

# KDD

La découverte des connaissances dans les bases de données (KDD) est le processus de recherche de connaissances dans un grand ensemble de données avec des méthodes d'exploration de données. Cette méthodologie est largement utilisée dans les domaines liés à l'IA comme l'apprentissage automatique. Les étapes de la méthodologie KDD sont :

* *Compréhension commerciale :* Couvre la compréhension du domaine d'application, l'exploration des réalisations connexes et l'identification des exigences et des besoins des utilisateurs.
* *Target Dataset :* Dans cette étape, nous effectuons une sélection des données sur lesquelles travailler. Comme un en règle générale, plus il y a de données, meilleurs sont les résultats.
* *Nettoyage des données :* après avoir sélectionné l'ensemble de données, nous devons le prétraiter et traiter anomalies, telles que des valeurs manquantes, du bruit ou de mauvaises valeurs
* *Transformation des données :* l'ensemble de données nettoyé doit être apporté dans le format adéquat avant de l'utiliser pour l'exploration de données. Cette étape implique des techniques telles que l'extraction de caractéristiques, ou réduction de dimension pour les tâches de visualisation.
* *Choix de la tâche :* dans cette phase, nous décidons du type de tâche que nous voulons réaliser, sous prise en compte du cas d'utilisation.
* *Algorithme :* Il existe des tâches qui peuvent être réalisées par plusieurs algorithmes. Un exemple est la tâche de classification. Par conséquent, KDD définit une étape où nous devons adopter le meilleur algorithme pour accomplir notre tâche.
* *Data Mining :* Nous appliquons l'algorithme que nous avons adopté à l'étape précédente et obtenons un modèle pour notre tâche, qu'il s'agisse de classification ou de regroupement.
* *Interprétation :* Nous examinons le résultat de l'application du modèle et vérifions si nous avons extrait la bonne information pour notre tâche.
* *Consolidation des connaissances :* La dernière étape consiste à utiliser les informations extraite Résoudre des problèmes.

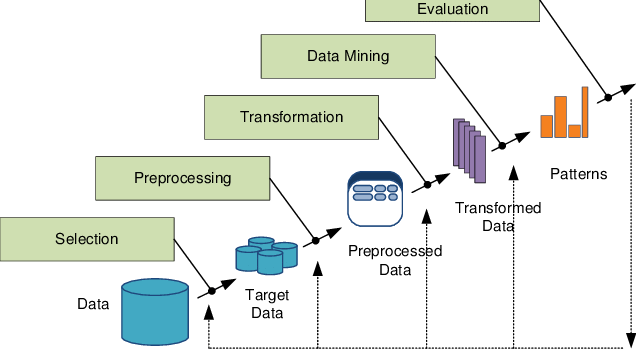


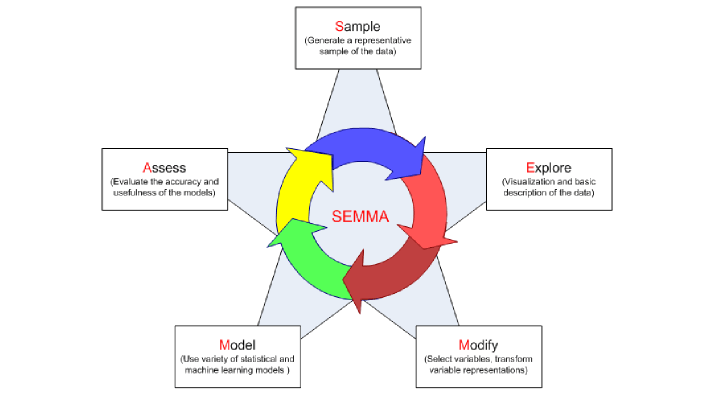
Figure1 : Les étapes de la méthodologie KDD [1]

1. **SEMMA**

L'acronyme SEMMA (sample, explore, modify, model, assess) qui se traduit en français par : échantillonne, explore, modifie, modélise, évalue) se rapporte au noyau du processus de conduite de l'exploitation de données. [2]

Le processus se décompose en son propre ensemble d'étapes. Ceux-ci inclus:

* *Échantillon :* Cette étape consiste à choisir un sous-ensemble du jeu de données de volume approprié à partir d'un vaste jeu de données qui a été fourni pour la construction du modèle. L'objectif de cette étape initiale du processus est d'identifier les variables ou les facteurs (à la fois dépendants et indépendants) influençant le processus. Les informations collectées sont ensuite triées en catégories de préparation et de validation.
* *Explorer :* Au cours de cette étape, une analyse univariée et multivariée est menée afin d'étudier les relations interconnectées entre les éléments de données et d'identifier les lacunes dans les données. Alors que l'analyse multivariée étudie la relation entre les variables, l'analyse univariée examine chaque facteur individuellement pour comprendre sa part dans le schéma global. Tous les facteurs d'influence susceptibles d'influencer les résultats de l'étude sont analysés, en s'appuyant fortement sur la visualisation des données.
* *Modifier :* Dans cette étape, les enseignements tirés de la phase d'exploration à partir des données collectées dans la phase d'échantillonnage sont dérivés avec l'application du logique métier. En d'autres termes, les données sont analysées et nettoyées, puis transmises à l'étape de modélisation et explorées si les données nécessitent un raffinement et une transformation.
* *Modèle :* Une fois les variables affinées et les données nettoyées, l'étape de modélisation applique diverses techniques d'exploration de données afin de produire un modèle projeté de la façon dont ces données atteignent le résultat final souhaité du processus*.*
* *Accès :* Dans cette dernière étape SEMMA, le modèle est évalué pour son utilité et sa fiabilité pour le sujet étudié. Les données peuvent maintenant être testées et utilisées pour estimer l'efficacité de ses performances. [3]



**Figure 2 :** les étapes de la méthodologie SEMMA [4]

1. **CRISP-DM**

CRISP-DM, qui signifie Cross-Industry Standard Process for Data Mining, est une méthode mise à l'épreuve sur le terrain permettant d'orienter vos travaux d'exploration de données. [5]

Les étapes de CRISP-DM sont :

• *Compréhension commerciale :* la première étape consiste à comprendre ce qu'il faut accomplir à partir d'un point de vue commercial. Cela signifie qu'une liste d'objectifs et de contraintes doit être correctement réglée. À travers cette étape, nous visons à saisir les facteurs qui peuvent avoir un impact sur le résultat du projet. Sans cette étape, cela signifie essayer de répondre aux mauvaises questions.

*• Compréhension des données :* La deuxième étape implique la collecte de données, l'analyse par l'outil de visualisation et d'intégration si les données ne sont pas dans le même format.

• *Préparation des données :* cette étape couvre le nettoyage des données et l'extraction des fonctionnalités pour la prochaine étape. Le nettoyage des données consiste à traiter les valeurs manquantes, etc.

*• Modélisation :* Dans cette phase, nous sélectionnons la technique de modélisation que nous appliquerons. Si plusieurs techniques sont nécessaires, cette étape doit être accomplie pour chacune.

• *Évaluation :* Dans cette étape, nous évaluons le modèle que nous avons créé à l'étape précédente en analyser des métriques telles que la précision de la prédiction, etc.

*• Déploiement :* Dans la dernière étape du processus CRISP-DM, nous prenons le modèle et utilisez-le une application du monde réel.

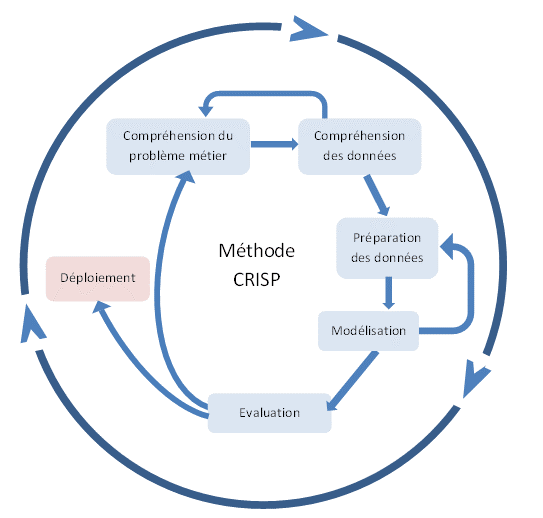


Figure 3 **:** les étapes de la méthodologie CRISP-DM

1. **Comparaison et méthodologie de choix**

On note une nette similitude entre les méthodologies KDD et CRISP-DM en termes de description des étapes. CRISP-DM a moins d'étapes que KDD, en fait KDD a trois étapes distinctes et bien étapes définies de la phase de modélisation, tandis que CRISP-DM définit abstraitement une étape. Dans ce sens, l'abstraction ne nous lie pas à un ordre ou à un processus particulier de construction de modèles. Dans ce point, CRISP-DM exprime simplement le processus souhaité mais avec moins de détails. SEMMA est défini par une société de logiciels pour guider les projets réalisés avec leur outil, n'est donc pas un standard, plutôt une méthodologie spécifique à un outil. Les normes offrent aux différents acteurs un moyen commun de gérer leurs projets. Les principaux avantages des normes sont l'interopérabilité qui conduit à efforts conjugués et innovation. Comme son nom l'indique, CRISP-DM est un standard ouvert norme [7]. Par conséquent, notre méthodologie de choix est la méthodologie CRISP-DM.

Tableau 4: Aperçu des méthodologies

|  |  |  |
| --- | --- | --- |
| **KDD** | **SEMMA** | **CRISP-DM** |
| Steps : 8 | Steps : 5 | Steps : 6 |
| Business Domain Understanding |  | Business Understanding |
| Data Selection | Sample | Data Understanding |
| Data Pre-processing | Explore |
| Data Transformation | Modify | Data Preparation |
| Data Mining Task | Model | Modeling |
| Choose Algorithm |
| Apply Algorithm |
| Interpretation / Evaluation | Assess | Evaluation |
| Pattern/Information Usage |  | Deployment |

1. **CRISP-DM Adapté en Scrum : Présentation**

Pour mieux conduire l'exécution des six phases, nous appliquons CRISP-DM dans le Scrum agile, car nous sommes confrontés aux défis suivants : L'un des avantages de Scrum est qu'il s'adapte aux changements fréquents d'exigences des parties prenantes. Puisque nous sommes construire une preuve de concept, nous serons confrontés à des situations où nous devons progresser, tout en ayant à changer constamment de décisions si nous sommes confrontés à des processus de goulot d'étranglement. Afin d’assurer une progression constante du travail nous nous appuierons sur le time-boxing de Scrum, afin que nous puissions diviser les tâches en tâches encore plus petites et nous assurer d'avoir des artefacts à la fin de chaque sprint.

Étant donné que nous allons parcourir les six phases de CRISP-DM, pour effectuer des tâches de réglage fin et ajustements, il est nécessaire de suivre les progrès en adoptant l'aspect incrémental de Scrum. Nous montrons comment planifier les étapes CRISP-DM en tant que sprints dans le tableau suivant :

Tableau 5 : Mappage des étapes CRISP-DM en tant que sprint Scrum

|  |  |
| --- | --- |
| **CRISP-DM Phases** | **Sprint** |
| Business understanding | Sprint 0 |
| Data understanding | Sprint 1 |
| Data preparation | Sprint 2 |
| Modeling and Evaluation | Sprint 3 |
| Deployment | Sprint 4 |

**Conclusion**

Dans ce chapitre, nous avons donné un aperçu du projet. Nous avons présenté l'hôte de stage entreprise, puis nous avons donné une brève description du projet énoncé le problème et proposé la solution que nous adopterons pour le résoudre.

# Chapitre 3:

# Solution proposée

**Introduction**

Dans ce chapitre, nous verrons comment fonctionne le mécanisme de détection d'intention avec le séquentiel model. Tout d'abord, nous présentons deux concepts sur lesquels séquentiel s'appuie, à savoir le concept d'insertion de mots et le modèle de transformateur. Ensuite, nous décrivons comment BERT est construit et comment il peut être affiné pour effectuer différentes tâches.

1. **Présentation de l'application**

Dans cette section, nous décrivons la grande image de la solution proposée. Nous montrons d'abord les composants du chatbot, puis nous illustrons le flux de travail interne.

1. **Architecture globale**

L'application se compose d'un modèle de traitement du langage naturel qui est exposé comme un service Web qui récupère le texte de l'utilisateur via un client Web et renvoie une réponse au client après avoir effectué sur l'entrée comme le montre la figure suivante.

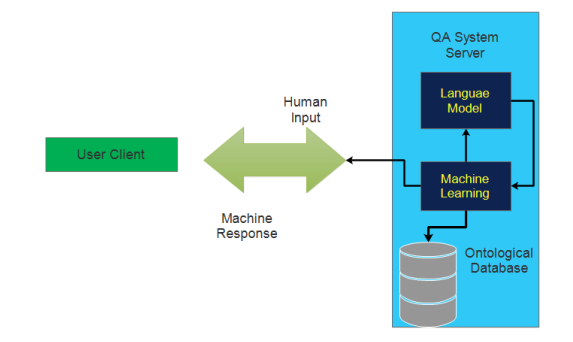


Figure 1 : architecture globale

1. **Représentation de mots dans Chatbot**

Dans cette section, nous introduisons le concept de plongement de mots qui est le cœur concept qui apportera de l'intelligence au chatbot.

1. **Word Embedding**

En PNL, les mots sont représentés comme des vecteurs de caractéristiques. Une caractéristique peut être un sens reliant deux mots. Les deux mots Chats et Chiens peuvent avoir des caractéristiques similaires, si l'on considère des caractéristiques telles que, est animal, âge, couleur, animaux vertébrés. Il y a plusieurs caractéristiques ou points communs qui peuvent apparaître entre les mots. Comme indiqué, les caractéristiques sont des vecteurs de nombres car essayez d'apprendre la distance entre les mots à l'aide de réseaux de neurones d'apprentissage en profondeur. Au final, nous représentons tous les mots d'un vocabulaire donné comme vecteurs, on obtient une matrice 3.1. Une matrice de plongements peut être considérée comme une table de recherche de mots du vocabulaire d'entrée. Pour retourner le vecteur de un mot particulier, on utilise la colonne id du dictionnaire illustré en 3.2, pour trouver la position du vecteur dans la matrice de plongements. Le mot incrustations fait référence au fait que les mots sont intégrés ou enfermés dans un espace avec un nombre fini de caractéristiques. 3.1 montre un enrobage avec dimensions D (nombre de caractéristiques) et mots V (nombre total de mots du vocabulaire.

Tableau 3.1 : Représentation d'une matrice d'inclusion

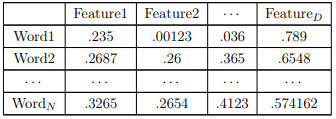
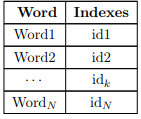


Tableau 3.2 : Dictionnaire



1. **Apprentissage des fonctionnalités**

Pour mieux comprendre comment sont calculées les valeurs des vecteurs de mots, nous prenons un exemple simplifié dans un espace vectoriel à deux dimensions comme indiqué en 3.3. La première étape comme pour toute tâche de PNL, est de fournir un corpus. On prend les deux phrases suivantes : « she is intelligent women » et « she is intelligent queen ». Lorsque le traitement du corpus atteint les mots women et queen, il voit que ces mots sont utilisés dans un contexte très similaire. le terme contexte désigne les mots entourant le mot actuel que l'algorithme examine.

En langage naturel, un contexte peut être comparé à des phrases ou des paragraphes. Basé sur dans un contexte similaire, les algorithmes apprennent à mapper women et queen sur deux vecteurs qui sont géométriquement proches les uns des autres. Des mots qui ne sont pas vus ensemble dans le corps sont séparés par une plus grande distance comme indiqué en figure3.3. Nous voyons < queen> et <women> sont plus proches l'un de l'autre, tandis que <women> et <man> ne le sont pas. On dit alors que la distance reflète la similitude. L'avantage que nous avons lorsque nous transformons des mots en vecteurs est que nous pouvons calculer la distance en utilisant la métrique de distance.

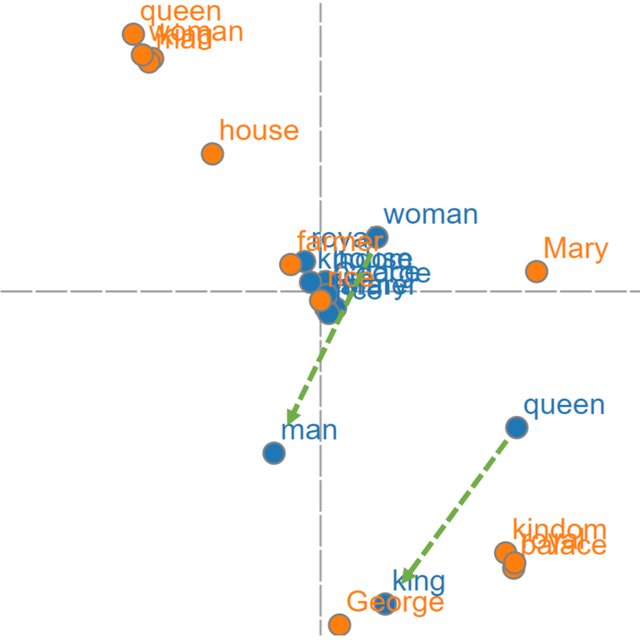


Figure3.3 : Espace vectoriel dimension pour représenter mots

1. **Détection de l'intention linguistique dans les chatbots**

Dans la section précédente, nous avons donné un aperçu de la façon dont les séquences de mots peuvent être représentées comme vecteurs de nombres afin que nous puissions les utiliser dans une tâche d'apprentissage en profondeur. Dans cette rubrique nous décrire le fonctionnement interne du modèle séquentiel. Nous bien voir comment les plongements sont calculés dans ce modèle.