ENSEEIHT

Apprentissage Profond
Rapport

Projet Immorthon - Base de données

Élèves :
Axel Bechu
Laerian Bontinck
Clément Demazure
Vianney Hervy
Yige Yang

Enseignant :
Axel Carlier

31 mars 2025

Table des matières

1 Introduction

Lorsque nous entendons pour la première fois un mot en français, il nous est parfois possible d'en comprendre le sens à l'aide du contexte ou de l'étymologie. C'est cette seconde méthode que nous allons essayer de reproduire à l'aide d'un LLM.

2 Description du sujet

Ce projet est donc un projet de génération de texte. Il sera réalisé sur un Jupyter Notebook hébergé sur Google Colab. Nous utiliserons principalement les bibliothèques `pandas` (pour le traitement préalable des données), `tensorflow.keras` (pour la création et l'entraînement du modèle) et `numpy` (pour le traitement général des données).

L'entrée du modèle sera un mot unique inventé [1] qui ressemble au français. La sortie attendue sera une courte définition de ce mot.

Le nom "Immorthon" est un mot-valise entre "immortel" et "python". "Immortel" est le surnom [2] donné aux académiciens de l'Académie Française, dont une partie de la responsibilité est de définir les mots de la langue française.

3 Liens

Notebook :
`https://colab.research.google.com/drive/1JHivpddBodExl7UOi0xoBnRhk8Y1bTVQ?usp=sharing`

Base de données :
`https://www.kaggle.com/datasets/kartmaan/dictionnaire-francais`

Repositoire git :
`https://github.com/sully-vian/immorthon`

1. avec ce projet par exemple `https://github.com/sully-vian/creabulaire`
2. `https://fr.wikipedia.org/wiki/Acad%C3%A9mie_fran%C3%A7aise#%C2%AB_Immortalit%C3%A9_%C2%BB`

4 Base de données

Nous prévoyons de partir d'un modèle déjà entraîné sur le français et de l'affiner sur un corpus de défnitions. Nous avons trouvé sur le site de Kaggle [3] un dictionnaire français [4] sous forme de `csv` qui allie plus de 800 000 mots et leurs définitions.

L'avantage principal de cette base de données est que l'hébergeur Kaggle dispose de sa propre API ainsi que de sa bibliothèque python, ce qui nous permet de récupérer les données directement sans les mettre sur un git nous-mêmes.

Nous pensons découper les mots en entrée, soit en tokenisant par syllabe (cohérent avec l'approche éymologique), soit en utilisant directement les lettres comme tokens. Nous pensons que la première approche est plus pertinente, car elle nous permet de mieux comprendre la structure du mot et d'en déduire son sens.

4.1 Structure de la base de données

La base de données utilisée est un tableau contenant deux colonnes principales :
— **Mot** : Cette colonne contient les mots en français.
— **Définitions** : Cette colonne contient les définitions associées à chaque mot.
Voici un exemple de données issues de la base de données :

Mot	Définitions
Toulouse	Commune de France située au bord de la Garonne, chef-lieu du département de la Haute-Garonne et de la région administrative Midi-Pyrénées.
Python	Langage de programmation interprété, orienté objet et de haut niveau, avec une syntaxe simple et lisible.

Table 1 – Exemple de structure de la base de données.

4.2 Traitement sur la base de données

La base de données possède toutefaois quelques défauts majeurs :
— Tous les mots de la première colonne commencent par une majuscule. Ce qui est réglable par une fonction de `lower()` sur la première colonne.
— Ensuite, une part non-négligeable de la base de donnée concerne les verbes conjugués et les variantes (plurielles ou féminines) de noms et adjectifs. Il nous faut donc trouver un moyen de les identifier pour ne pas les inclure dans nos données d'entraînement ou de test.
— De même, nous avons décidé d'exclure les noms de villes, de pays et de personnes et les noms propres en général. En effet, ces mots sont souvent très spécifiques et ne sont pas représentatifs de la langue française dans son ensemble.
— Nous envisageons également d'exclure les mots qui sont "explicitement" issus de langues étrangères, dont l'étymologie est clairement à part. Ces mots seront considérés comme "outliers" ne suivant pas la "logique" que nous souhaitons inculquer à notre modèle.

4.2.1 Exemple de données valides

```
1    Mardi ["Deuxième jour de la semaine"]
```

3. https://www.kaggle.com/
4. https://www.kaggle.com/datasets/kartmaan/dictionnaire-francais

4.2.2 Exemples de données invalides

```
    # variation d'un nom
    Allemands ["Masculin pluriel de Allemand"]
    # verbe conjugué
    Ont ["Troisième personne du pluriel de l'indicatif présent du verbe avoir."]
    # nom de ville
    Toulouse ["commune de France située au bord de la Garonne,  chef-lieu du
    ↪  département de la Haute-Garonne et de la région administrative
    ↪  Midi-Pyrénées."]
```