Université Mohammed V - Rabat
École Nationale Supérieure d'Informatique et d'Analyse des Systèmes

Processus de decision markovien

Filière

Ingénierie de l'Intelligence Artificielle (2IA)

Sujet :

Analyse du texte par une approche markovienne

Réalisé par :

ZIYAD ABIDATE

Année Universitaire 2024-2025

19 février 2025

0.1 Introduction

L'objectif de ce TP est d'explorer l'utilisation des chaînes de Markov pour l'analyse et la génération de texte. Une chaîne de Markov est un modèle probabiliste qui permet de modéliser des systèmes dynamiques où l'état futur ne dépend que de l'état actuel. Dans notre cas, nous utilisons ce modèle pour apprendre les relations entre les caractères ou les mots dans un texte donné et générer du texte en conséquence.

Nous avons implémenté trois modèles distincts :

— Un modèle de Markov de premier ordre basé sur les transitions entre lettres.
— Un modèle de Markov basé sur les triplets de lettres afin de capturer davantage de contexte.
— Un modèle de Markov basé sur les transitions entre mots pour une meilleure cohérence sémantique.

Chaque modèle a été évalué en fonction de sa capacité à reproduire une structure textuelle cohérente et significative a la base de l'approche d'évaluation discutée dans le support de TP.

0.2 Modélisation et génération de texte

0.2.1 Modèle de Markov d'ordre 1 (Lettres)

Dans cette première approche, nous avons construit un modèle de Markov simple où chaque caractère est prédit en fonction du caractère précédent. L'idée est de représenter le texte sous forme d'une matrice de transition où chaque ligne correspond à une lettre et chaque colonne indique la probabilité de transition vers une autre lettre. Après avoir entraîné le modèle sur un corpus de texte, nous avons généré du texte en sélectionnant les lettres avec les probabilités maximales.

L'évaluation a montré que cette approche produit du texte relativement aléatoire, souvent difficilement compréhensible. Bien que certaines séquences de lettres correspondent à des mots réels, la majorité du texte généré manque de structure et de cohérence.

Résultats et évaluation

— Score total : 68.3811
— Précision globale : 34.3624%

Le texte généré contient des fragments de mots, mais reste en grande partie incompréhensible.

0.2.2 Modèle basé sur les triplets

Afin d'améliorer la qualité du texte généré, nous avons adopté une approche basée sur des triplets de lettres. En considérant trois lettres à la fois, le modèle est capable de mieux capturer les structures courantes des mots et d'améliorer la fluidité des transitions.

Les résultats ont montré une nette amélioration par rapport au modèle précédent. Le texte généré contient quelques mots corrects et semble plus naturel, bien que des erreurs persistent.

Résultats et évaluation

— Score total : 115.9826

— Précision globale : 58.8744%

Ce modèle génère du texte qui ressemble davantage à une langue naturelle et présente une meilleure structuration des mots sauf qu il lui manque de la sémantique.

0.2.3 Modèle basé sur les mots

Dans cette dernière approche, nous avons choisi d'appliquer la chaîne de Markov sur les mots plutôt que sur les lettres. L'idée est d'apprendre la probabilité de transition entre des mots entiers, ce qui permet de générer des phrases ayant une meilleure cohérence syntaxique et sémantique.

Les résultats montrent que bien que la précision de ce modèle soit inférieure à celle du modèle basé sur les triplets, le texte produit est beaucoup plus fluide. Cela souligne une faiblesse dans la méthode d'évaluation qui privilégie la correspondance de caractcres sans prendre en compte la signification globale du texte.

Résultats et évaluation

— Score total : 66.9682

— Précision globale : 33.6524%

Malgré un score inférieur, le texte généré est de meilleure qualité sur le plan sémantique comparé aux deux premiers modèles .

0.3 Discussion et Conclusion

Ce TP nous a permis d'explorer différentes approches pour la génération de texte en utilisant des chaînes de Markov. Les résultats obtenus montrent que les modèles basés sur les lettres sont limités et produisent souvent du texte incohérent. En revanche, les modèles basés sur les triplets améliorent significativement la structure du texte, et les modèles basés sur les mots offrent une plus grande fluidité et une meilleure cohérence sémantique.

Nous avons également constaté que la méthode d'évaluation utilisée présente certaines limites. En effet, elle se base uniquement sur des critères de correspondance de caractères et ne prend pas en compte la qualité sémantique du texte. Par exemple, un texte généré avec de nombreuses répétitions peut obtenir un score élevé alors qu'il n'a pas de véritable sens. Cette observation met en lumière l'importance de développer des métriques d'évaluation prenant en compte le contexte et la signification du texte généré.

En conclusion, les chaînes de Markov sont une approche intéressante pour la génération de texte, mais elles présentent des limites lorsqu'elles sont appliquées à des structures linguistiques complexes. Des approches plus avancées, comme les modèles basés sur les réseaux de neurones, pourraient être explorées pour améliorer encore la qualité du texte généré.