Tale Box

Projet Traitement automatique du texte en Intelligence Artificielle

Ben Fredj Yasmine

2020 - 2021

Plan

MASTER INFORMATIQUE

UNIVERSITÉ CÔTE D'AZUR

I. Introduction
II. Les Taches principales
    - Collecte des données
    - Prétraitement des données
    - Modélisation et prédiction du sujet
    - Génération du texte
III. Analyse des erreurs
IV. Amélioration Possibles
V. Démonstration
VI. Conclusion

I. Introduction

MASTER INFORMATIQUE

UNIVERSITÉ CÔTE D'AZUR

Tale Box

Jeu Textuel développer en Python et relevant des concepts NLP (Natural Language Processing). Il peut :

- Discuter avec nous (NLU / NLG)
- Nous aider à raconter une histoire (NLG)
- Deviner le thème de l'histoire (NLU)

Importé les données → Pré-traitement → Entrainer les modèles

II. Les Taches principales :

MASTER INFORMATIQUE

UNIVERSITÉ CÔTE D'AZUR

Collecte des données

Pour reconnaitre le thème d'une histoire comme pour pouvoir en rédiger, nos modelés ont besoin d'un grand nombre de données.

Ce données consiste à 300 contes, histoires et légendes que j'ai collecter personnellement.

II. Les Taches principales :

MASTER INFORMATIQUE

UNIVERSITÉ CÔTE D'AZUR

Prétraitement des données

Supprimer les caractères spéciaux

Avec librairie « re.sub »

Tokenizer

Avec « nltk.word_tokenize »

Création des Bigrames et Trigrammes

Avec « gensim.models.Phrases »

Filtre les mot d'arrêt

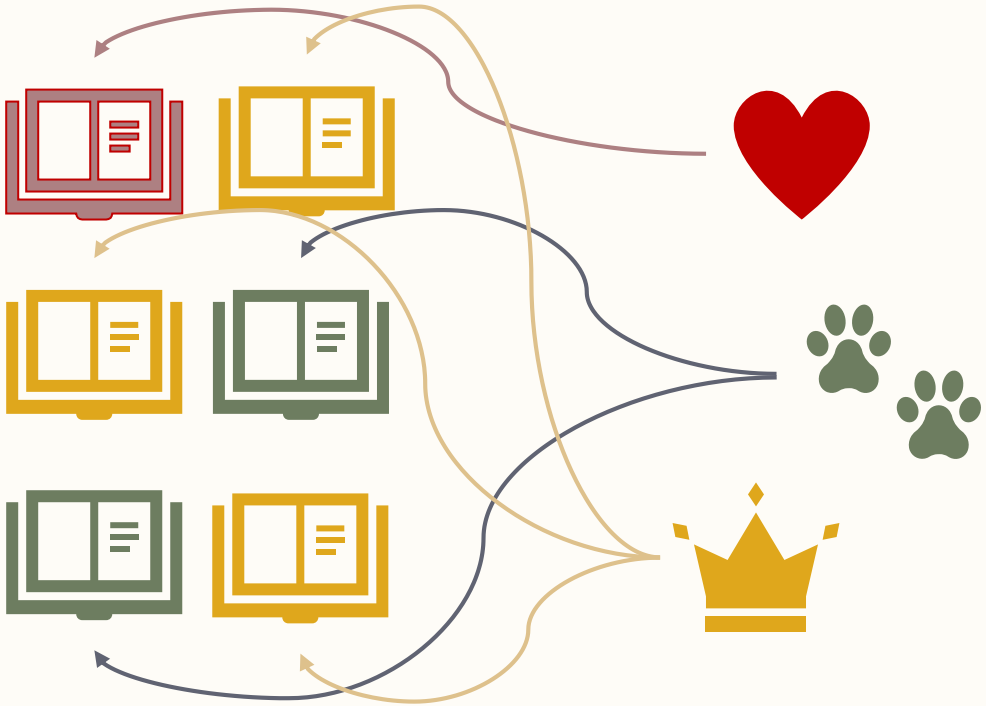
Mots de « nltk.corpus » (en français)

Lemmatiser

Avec librairie « spacy »	Adverbe, Adjectif, Nom, Verbe

II. Les Taches principales :

Modélisation



Allocation de Dirichelet latente (LDA) :

- Un modèle génératif probabiliste qui permet de regroupé des données autours d'un nombre définie de thèmes par le liens de ressemblance.
- Ce modèle nous permet d'avoir K listes de N mots et chaque liste est un thème.
- Ensuite à nous d'essayer d'analyser les mots pour savoir à quelle thème elles correspondent.

II. Les Taches principales :

MASTER INFORMATIQUE

UNIVERSITÉ CÔTE D'AZUR

Modélisation : Prédiction des thèmes

Création de sacs des mots (indice, fréquence)

Avec « CountVectorizer » de « sklearn.feature_extraction.text »

Chercher les meilleurs hyperparamètres pour le model LDA

Avec la méthode « GridSearch » de « sklearn.model_selection »

Créer et entrainer le model LDA avec les données prétraités

Avec « sklearn.decomposition »

II. Les Taches principales :

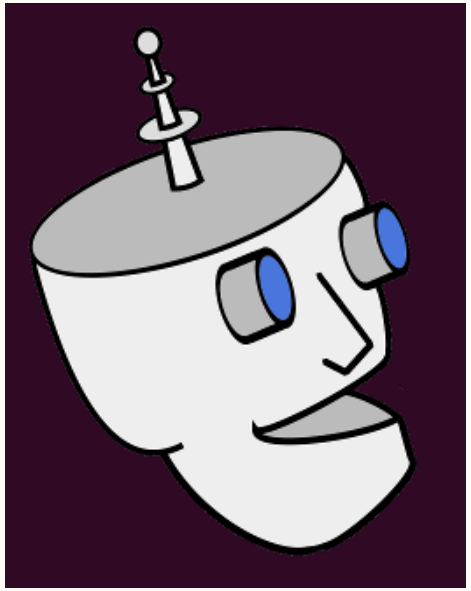
MASTER INFORMATIQUE

UNIVERSITÉ CÔTE D'AZUR

Générations de texte

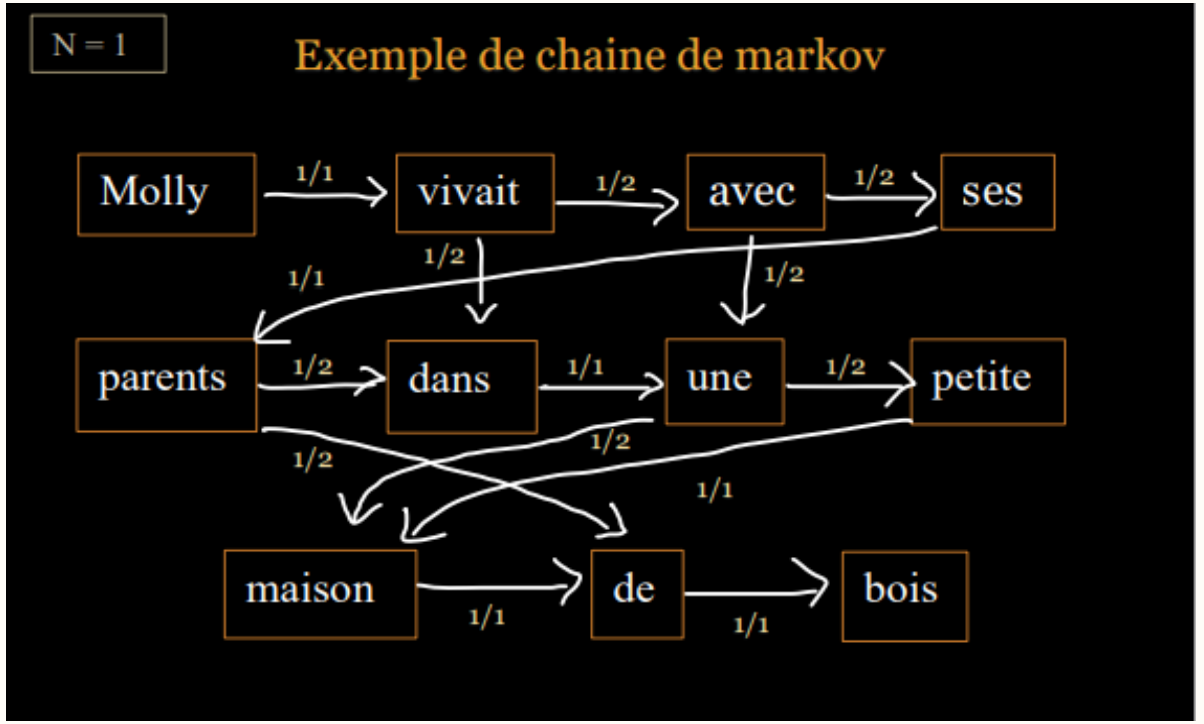
Répondre à des question basique

ChatterBot



Continuer la rédaction d'une histoire

Modèle de markov



III. Analyse des erreurs

MASTER INFORMATIQUE

UNIVERSITÉ CÔTE D'AZUR

- Le model de Markov peut ne pas reconnaitre un état s'il ne l'a jamais rencontrer dans la phase de l'entrainement .

- Erreurs de génération de texte:
  - Phrases non finis
  - Phrases avec un sens ambiguë
  - Phrases sui ont tendance à s'éloigné du contexte

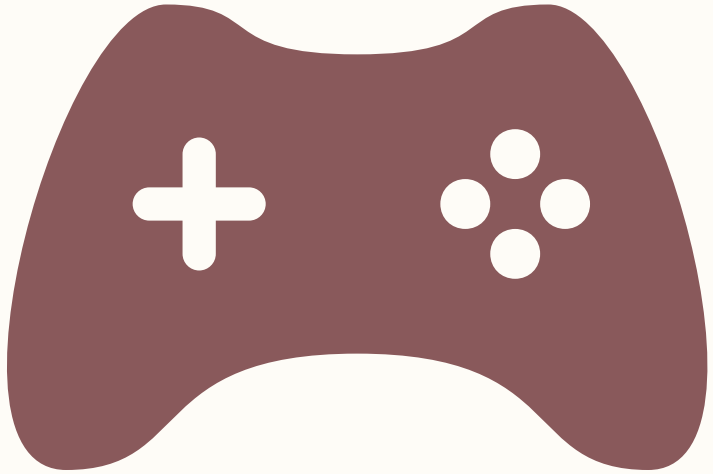
- Prédiction du sujets peut ne pas être précise

- Le bot peut ne pas répondre correctement dans la discutions s'il n'a pas était entrainer sur plusieurs discutions.

V. Démonstration

MASTER INFORMATIQUE

UNIVERSITÉ CÔTE D'AZUR



IV. Améliorations possibles

MASTER INFORMATIQUE

UNIVERSITÉ CÔTE D'AZUR

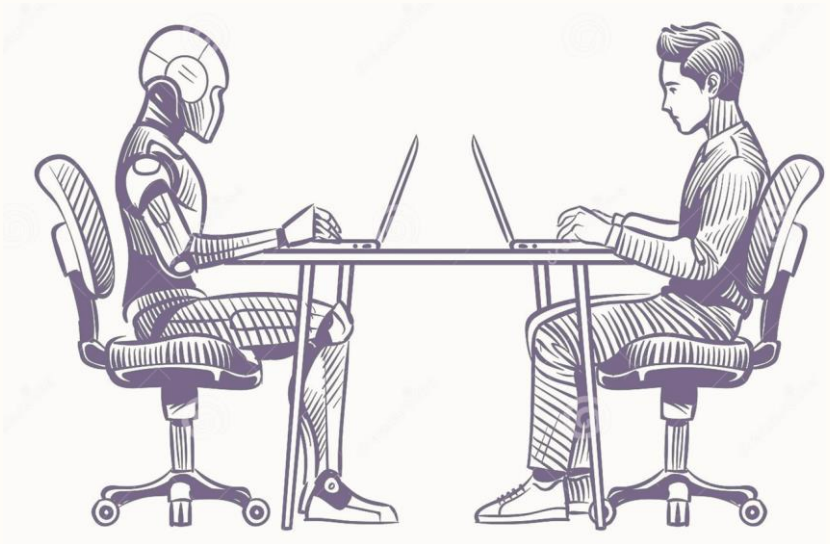
- Avoir un jeu de données plus grand et plus organiser.
- Avoir une génération de texte plus adapter pour avoir des phrases fini et pouvoir garder le contexte de l'histoire tout au long.
- Entrainer notre bot sur plus de discutions.
- Améliorer l'interface graphique.
- Ajouter plus d'option au jeu:
  - Permettre au joueur de faire un choix entre plusieurs suites de texte.
  - Permettre au bot d'analyser l'état d'esprits du joueur avec l'histoire générer.

VI. Conclusion

MASTER INFORMATIQUE

UNIVERSITÉ CÔTE D'AZUR

- Objectifs plus ou moins Atteints
- Découvert de nombreuse librairies et concept de traitement automatique de texte
- Découverte de la difficulté de permettre à l'ordinateur d'agir comme un humain



Merci pour votre attention