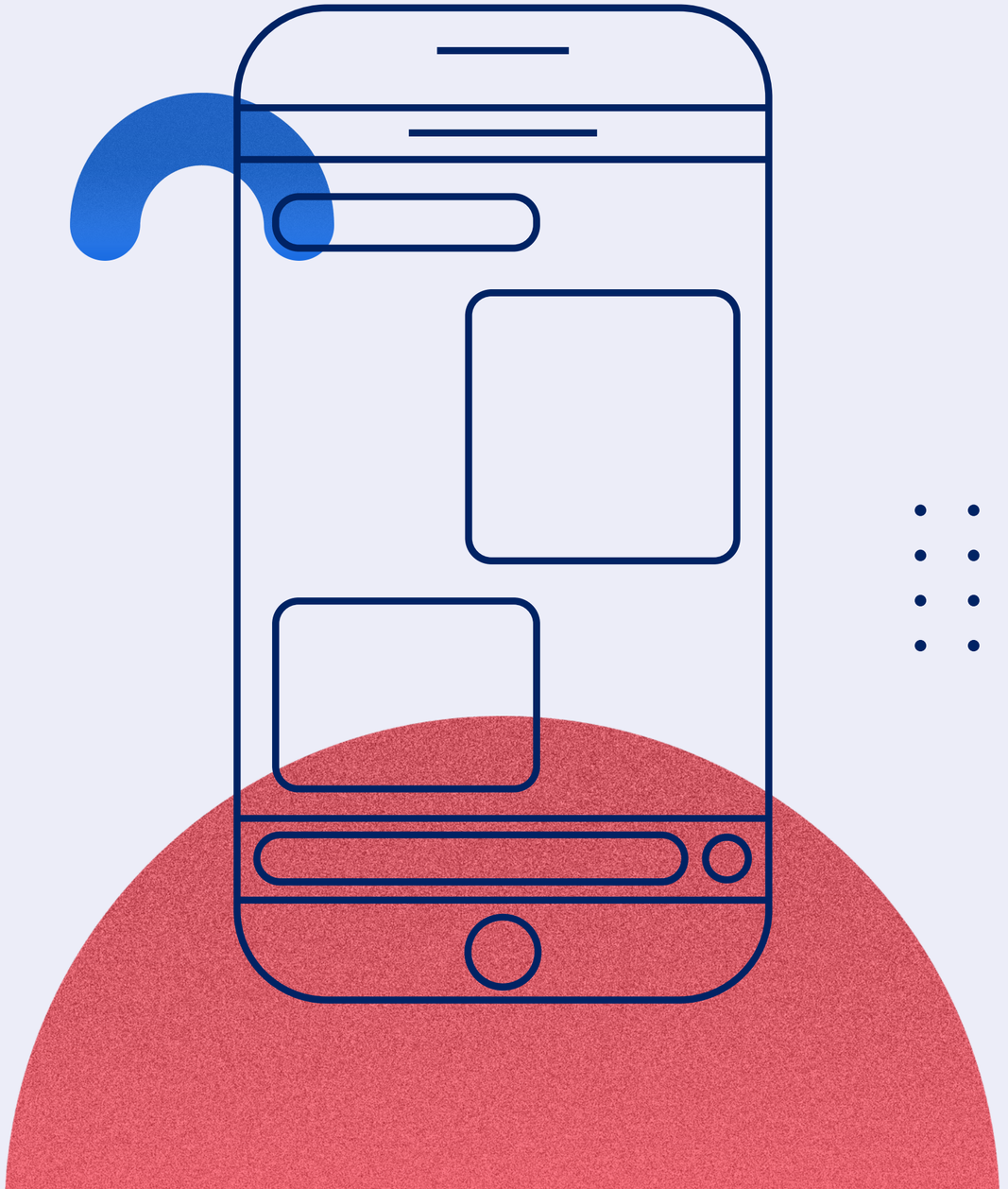
Projet ChatBot
Deep Learning



Marianne / Eden / Jonathan

SOMMAIRE

Contexte du projet

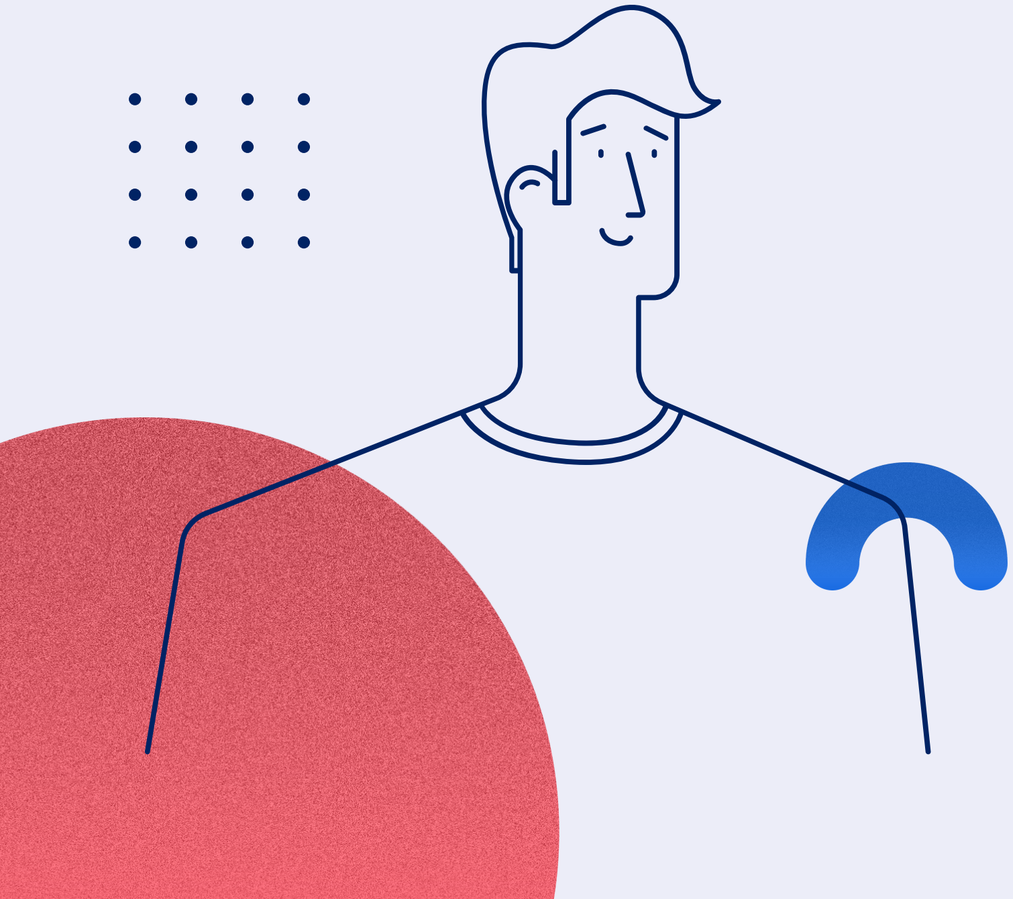
Interfaces utilisateur

EDA

Modèles utilisés

Bilan

LE CONTEXTE



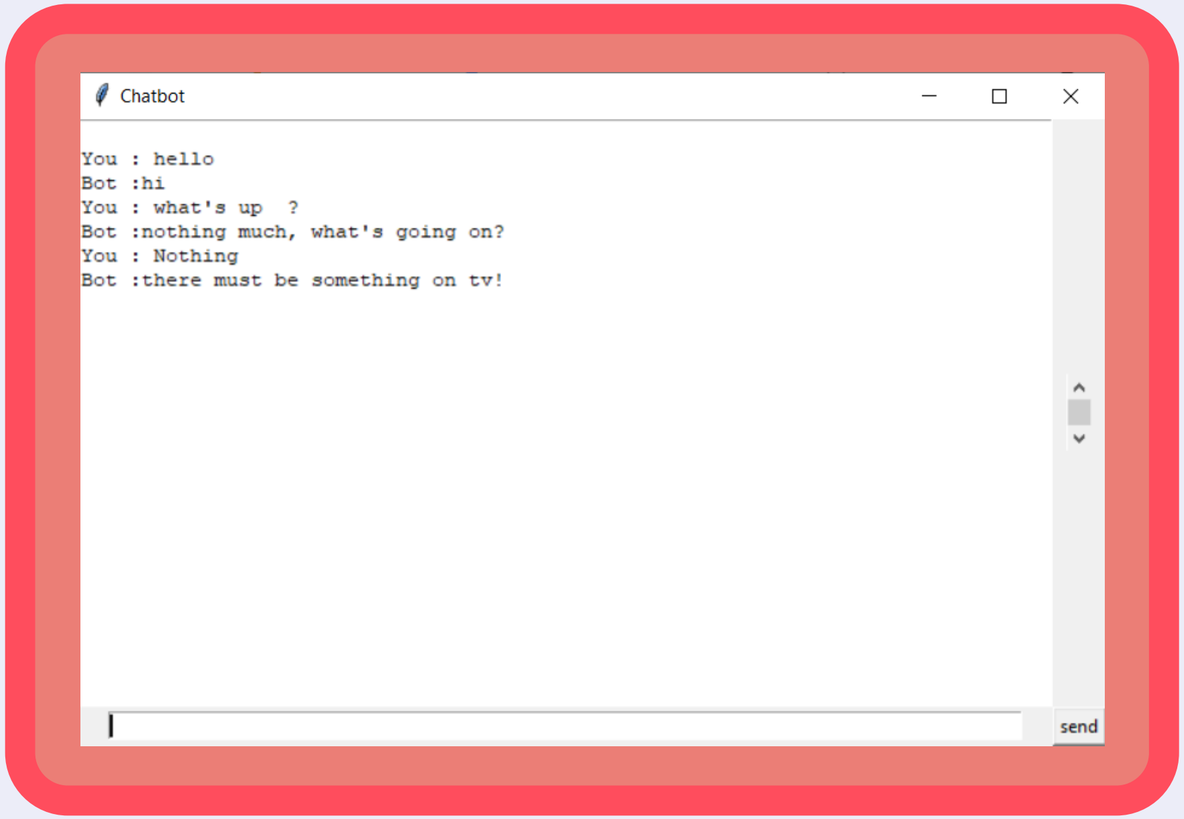
Notre entreprise a mandaté notre équipe pour réaliser un chatbot à destination d'un de ses clients.

Dans un premier temps, l'objectif principal de cet outil sera d'assurer un conversation superficielle avec les utilisateurs, afin de tromper la solitude liée à l'isolement de certaines personnes. Il sera utilisé sur un site de soutien contre l'isolement.

Un MVP devra être prêt à la fin de la 2ème semaine. Il nous a été demandé d'utiliser les données fournies par le client pour entrainer un modèle avec du machine learning, puis du deep learning.

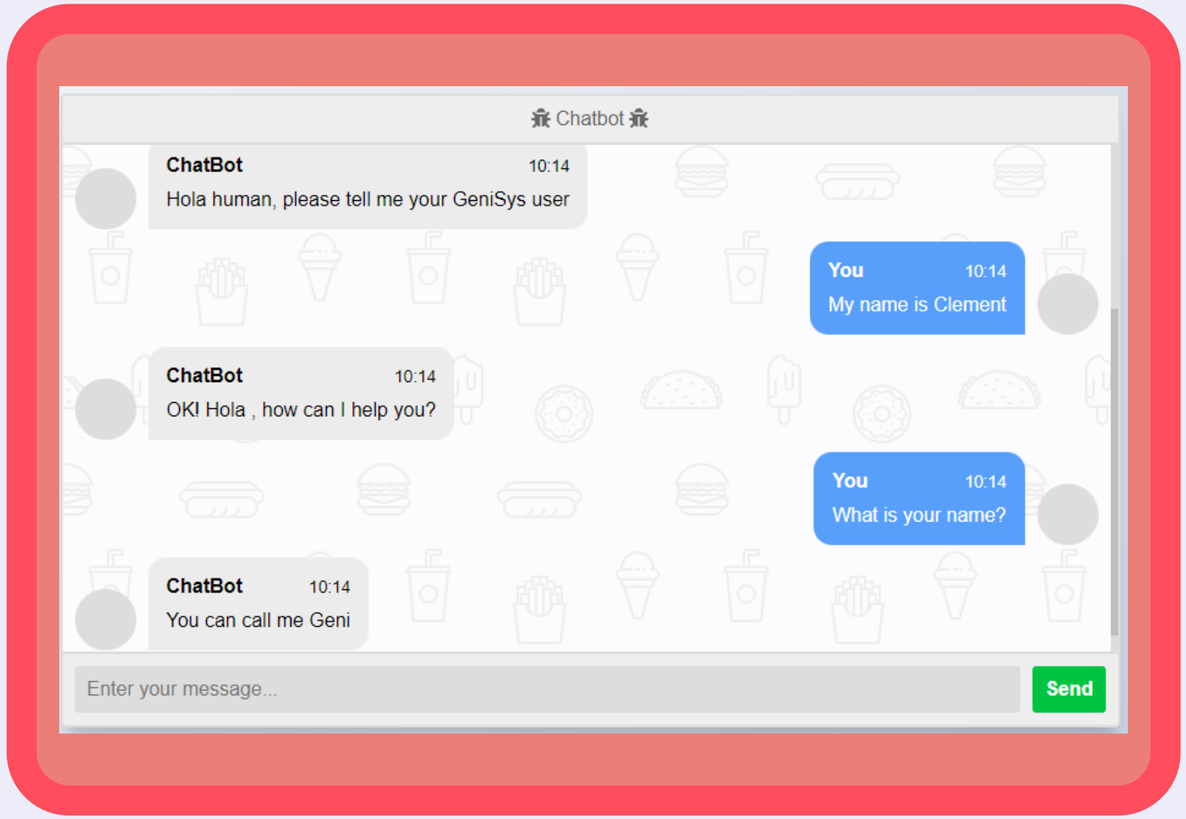
INTERFACE LOGICIEL TKINTER

- Utilisation de le librairie tkinter sur python
- Interface graphique pour le chatbot via un logiciel
- Possibilité d'installer le chatbot via un executable
- Utilisation du bouton "send" pour envoyer la question
- Réponse retournée par le chatbot
- Logiciel de base, amélioration UI/UX nécessaire avant mise à disposition pour les utilisateurs



INTERFACE VIA WEBAPP

- Utilisation du micro framework Flask
- Interface graphique pour le chatbot via web app
- Déploiement effectuée avec Heroku (free tier)
- Utilisation du bouton "send" pour envoyer la question
- Réponse retournée par le chatbot
- Interface graphique en HTML/CSS
- https://maredjo-chatbot.herokuapp.com/



CHATBOT SIMPLIFIÉ DATASET

Aperçu du dataset : un dataset simple

	Questions	Answers
0	hi, how are you doing?	i'm fine. how about yourself?
1	i'm fine. how about yourself?	i'm pretty good. thanks for asking.
2	i'm pretty good. thanks for asking.	no problem. so how have you been?
3	no problem. so how have you been?	i've been great. what about you?
4	i've been great. what about you?	i've been good. i'm in school right now.
...	...	...
3724	but i do all my writing with my right hand.	start typing instead. that way your left hand ...
3725	Hi	hello
3726	Hello	hi
3727	how are you	i'm fine. how about yourself?
3728	how are you doing	i'm fine. how about yourself?

3729 rows × 2 columns

CHATBOT SIMPLIFIÉ

Le premier chatbot réalisé, très simplifié :
- pas de model à faire
- pas de pickle

Ajout à nos données existantes des questions de bases, comme différentes façon de 'saluer' ou de demander 'comment ça va ?'

Création d'un Pipeline contenant :

- CountVectorizer transforme un texte en vecteur sur la base de tout les mots du texte

- TfidfTransformer : transforme une metrique et la normalise en term frequency ou term frequency inverse document frequency (term frequency = mesure de l'importance du terme dans l'ensemble du corpus)

- DecissionTreeClassifier

```
"let's go see spiderman"
```

```
'that sounds like fun.'
```

CHATBOT RESEAU DE NEURONES

On a essayé plusieurs modèles:

premier modèle : Embedding , simpleRNN .

deuxième modèle : Embedding , LSTM.

L'embedding :

- L'embedding nous empêche de convertir des mots dans une taille fixe , Le vector restant est danse et ou il ajoute de vrais valeurs et pas que des 0 et 1

LSTM :

- Le LSTM a des connexions de rétroaction. Il peut traiter non seulement des points de données uniques (tels que des images), mais également des séquences entières de données (telles que la parole ou la vidéo)

simpleRNN :

- Un réseaux de neurone récurrent est un type spécial de réseaux de neurone il est adapté pour travailler sur les donnée des séries temporelle ou qui contient des séquences

EDA : GÉNÉRAL

0	Greeting
1	GreetingResponse
2	CourtesyGreeting
3	CourtesyGreetingResponse
4	CurrentHumanQuery
5	NameQuery
6	RealNameQuery
7	TimeQuery
8	Thanks
9	NotTalking2U
10	UnderstandQuery
11	Shutup
12	Swearing
13	GoodBye
14	CourtesyGoodBye
15	WhoAmI
16	Clever
17	Gossip
18	Jokes
19	PodBayDoor
20	PodBayDoorResponse
21	SelfAware

Liste des intents

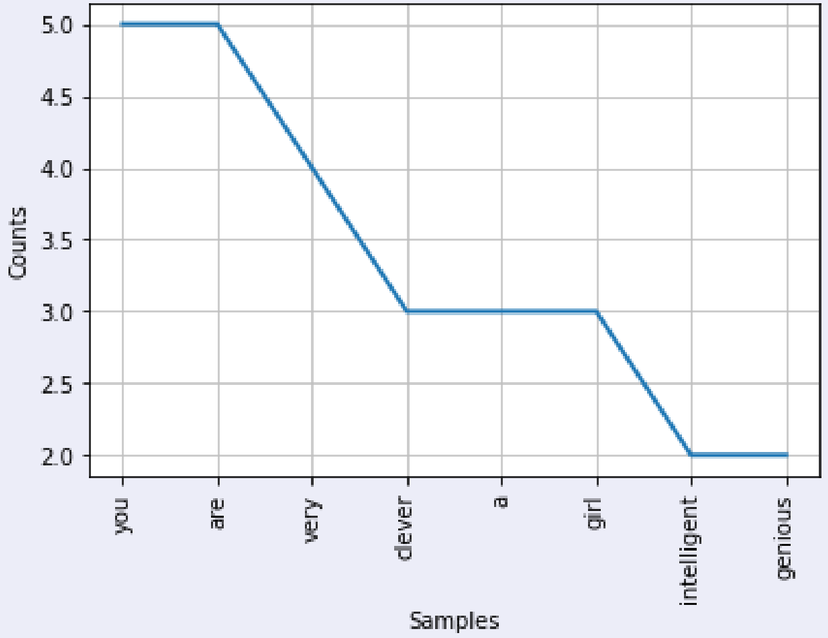
Nombre d'intents : 22

Nombre d'inputs moyen par intent : 6.5

Nombre de responses median par intent : 3
(Gossip et Jokes > 1 000)

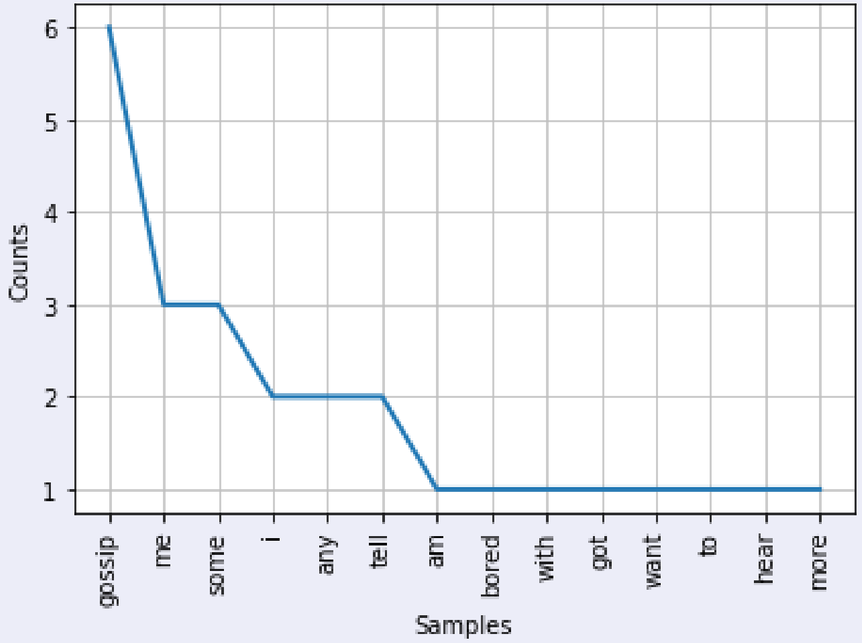
EDA : DATASET DE BASE

Clever



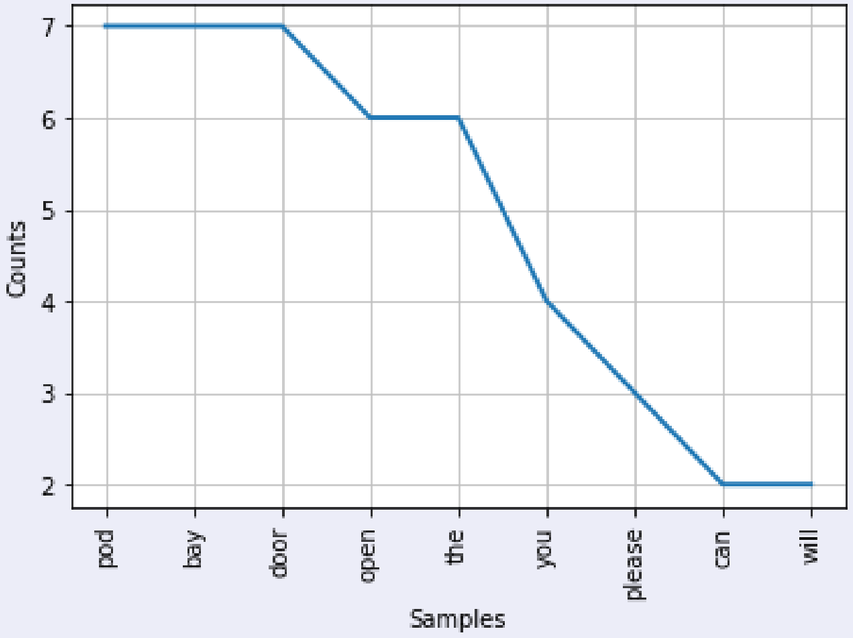
--> mots courants (x10)

Gossip



--> 1 mot clé (x7)

PodBayDoorResponse

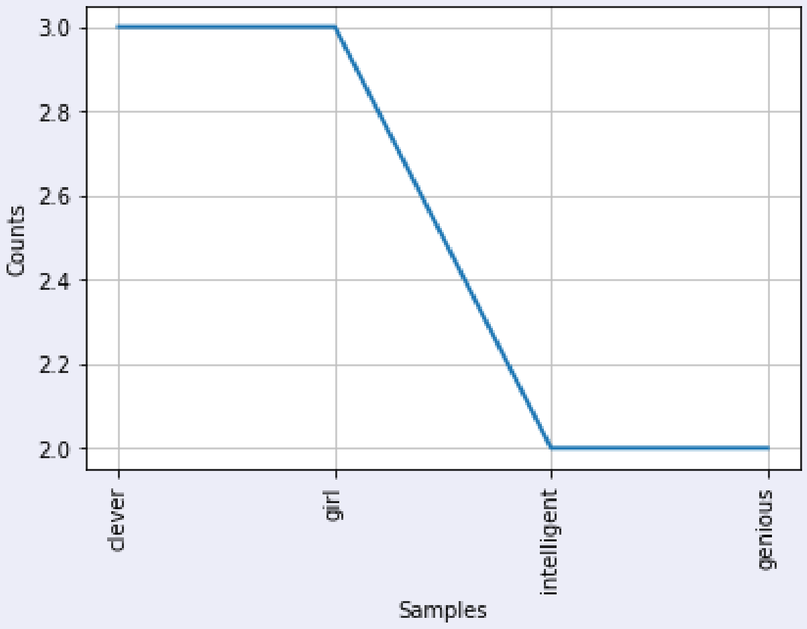


--> plusieurs mots clé (x5)

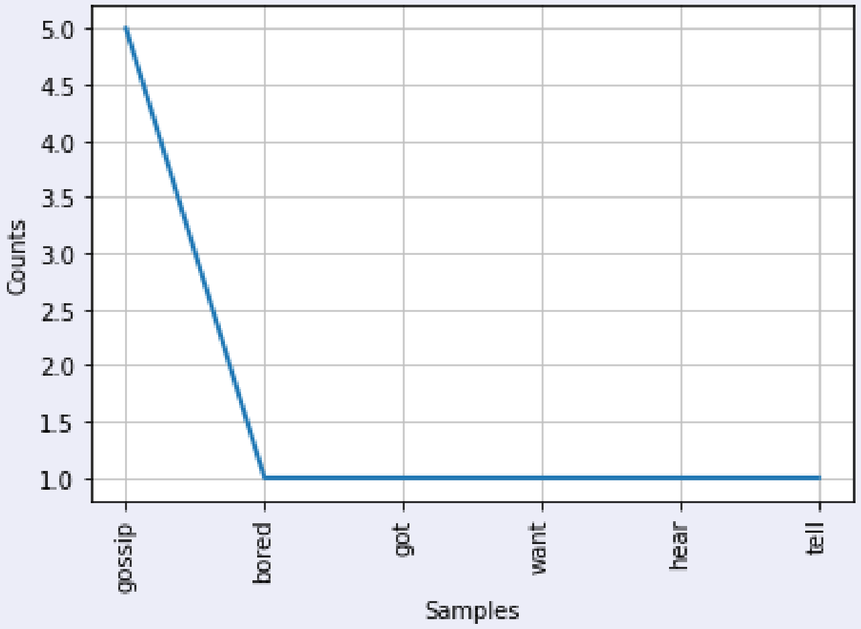
RQ : mot clé "thanks" majoritaire dans plusieurs intent

EDA : STOP WORDS

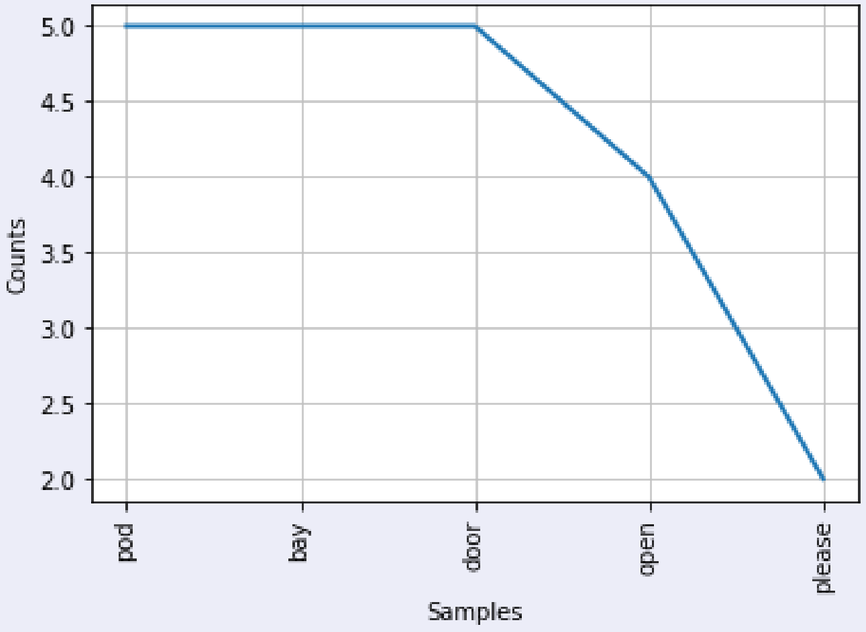
Clever (mots courants)



Gossip (1 mot clé)



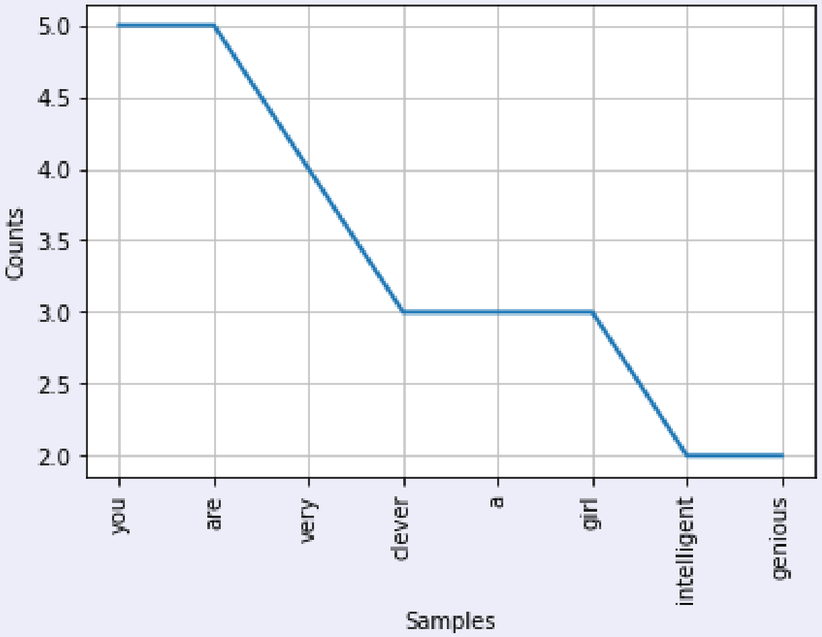
PodBayDoorResponse (plusieurs mots clé)



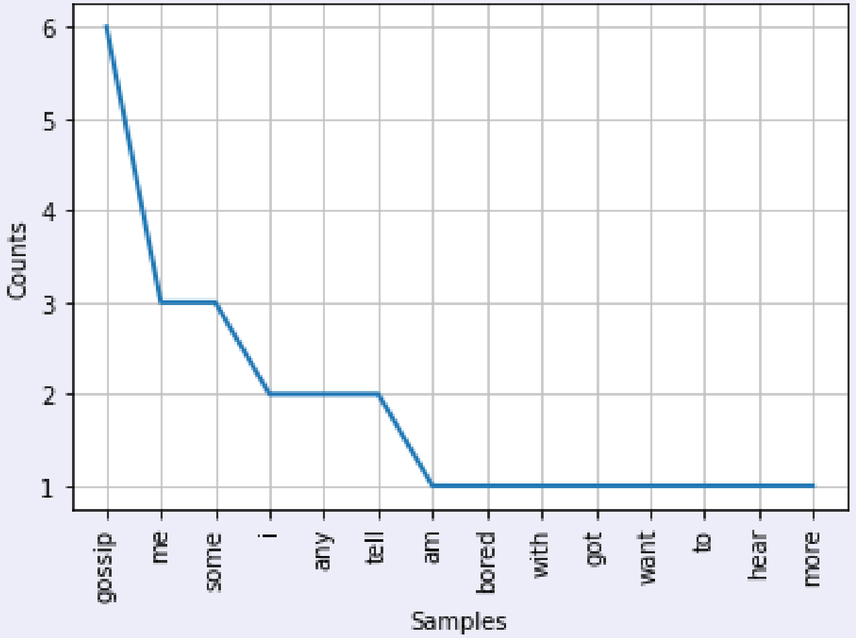
--> amélioration des "mots courants" sans dégration des autres catégories = piste à explorer

EDA : LEMMATIZATION

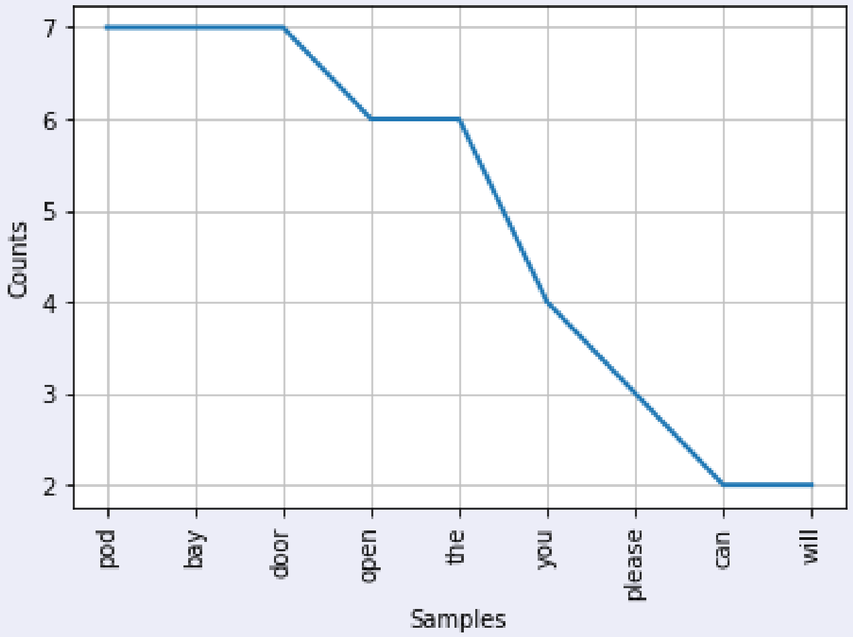
Clever (mots courants)



Gossip (1 mot clé)



PodBayDoorResponse (plusieurs mots clé)



--> pas de changements notables = piste non retenue

MODELE SIMPLE RNN

Architecture:

- Sequential()
- Embedding(vocabulary+1,10))
- InputLayer(input_shape=(input_shape,))
- SimpleRNN(30)
- Dense(output_length, activation="softmax")

Compilation :

- loss="sparse_categorical_crossentropy"
- optimizer="adam"
- metrics=["accuracy"]

Fit : 100 epochs

Validation accuracy: 0.52

SIMPLE RNN & STOP WORDS

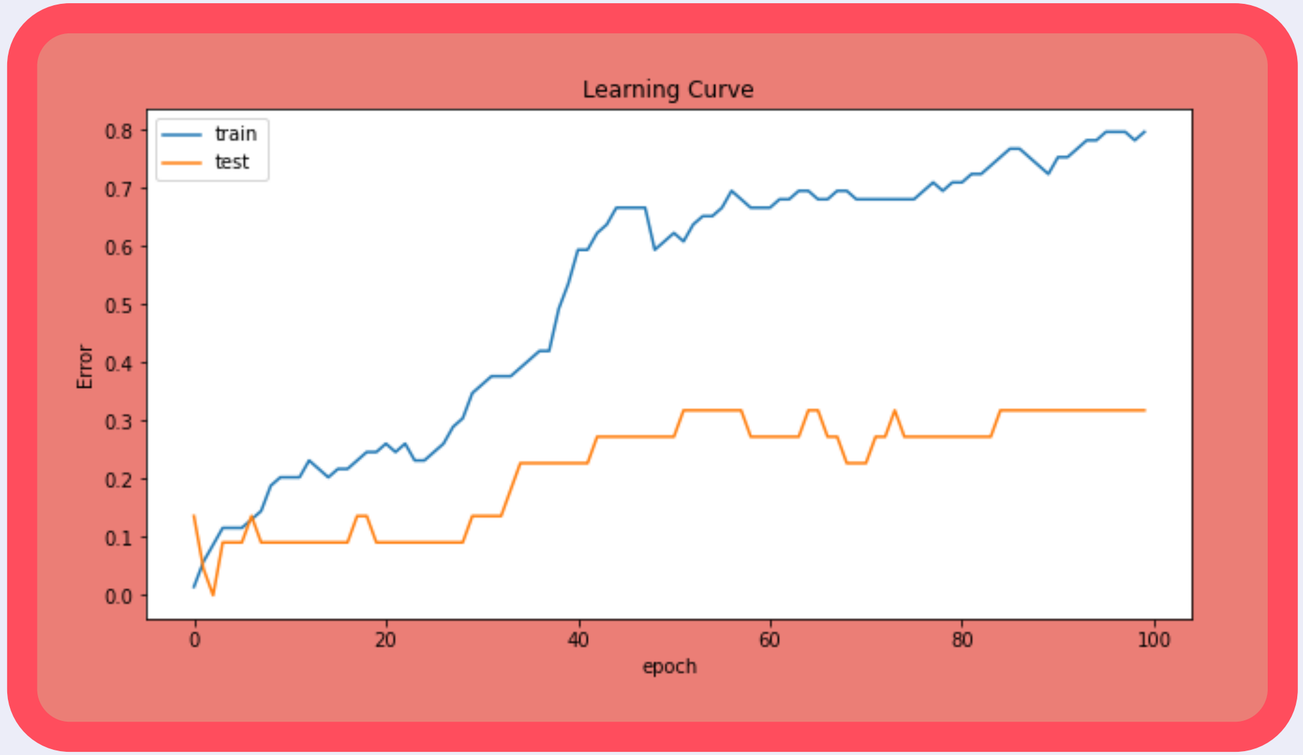
Architecture:

- Sequential()
- Embedding(vocabulary+1,10))
- InputLayer(input_shape=(input_shape,))
- SimpleRNN(30)
- Dense(output_length, activation="softmax")

Compilation :

- loss="sparse_categorical_crossentropy"
- optimizer="adam"
- metrics=["accuracy"]

Fit : 100 epochs



Validation accuracy: 0.31

LSTM

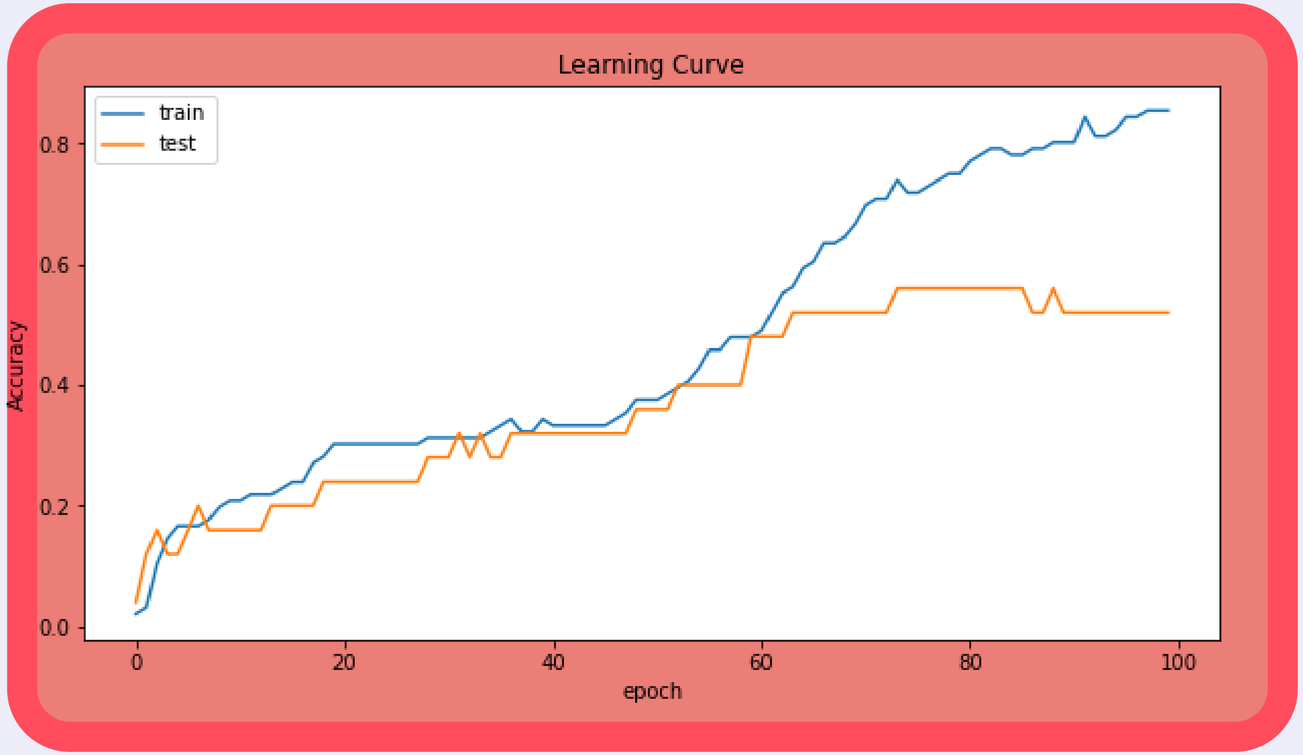
Architecture:

- Sequential()
- InputLayer(input_shape=(input_shape,))
- Embedding(vocabulary+1,10)
- LSTM(10, return_sequences=True)
- Flatten()
- Dense(output_length, activation="softmax")

Compilation :

- loss="sparse_categorical_crossentropy"
- optimizer="adam"
- metrics=["accuracy"]

Fit : 100 epochs



Validation accuracy: 0.52

BLOCAGES ET AMÉLIORATIONS

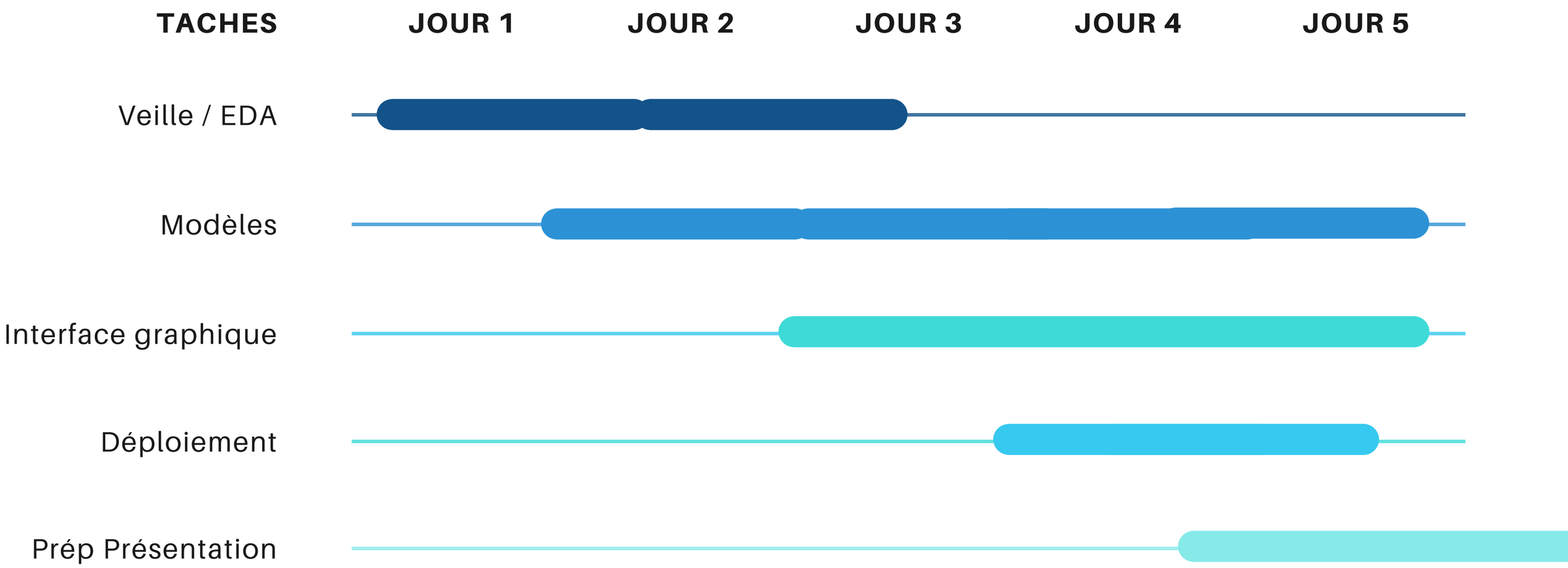
Blocages

- Données insuffisantes
- Solution logicielle non maitrisée
- Difficultés pour généraliser le modèle
- Connaissance limitées sur preprocessing des données pour un chatbot

Améliorations

- Architecture différente à tester (bidirectionnel, ajout couche LSTM)
- Enrichir les données d'entrainement
- Utiliser les réponses des utilisateurs pour améliorer et utiliser une sorte d'active learning.

PLANNING


TACHES
JOUR 1
JOUR 2
JOUR 3
JOUR 4
JOUR 5
Veille / EDA
Modèles
Interface graphique
Déploiement
Prép Présentation