Apprentissage Profond

Génération de proverbes

Groupe L34

Élèves :

THEVENET Louis

LEBOBE Timothé

TENE Zacharie

SABLAYROLLES Guillaume

31 Mars 2025

Table des matières

1. Génération de proverbes en anglais 3
2. Consitution de la base de données 3
   2.1. Acquisition des données 3
   2.2. Script de chargement des données 3
3. Création et entraînement du modèle 4
   3.1. Création du modèle 4
   3.2. Entraînement 5
   3.3. Génération 5
4. Analyse des résultats 6
   4.1. Entraînement sur les proverbes anglais 6
   4.2. Entraînement sur les proverbes traduits 7
   4.3. Expérience avec des humains 9
5. Conclusion 9

1. Génération de proverbes en anglais

Nous avons choisi de créer un modèle de génération de proverbes. La base de donnée est trouvable dans le dossier `raw_data/` à la racine de ce dépôt GitHub

Voici un exemple de proverbes de notre base d'entraînement :

- He that brings good news, knocks hard.
- Anger and haste hinder good counsel.
- Big thunder, little rain.
- Romeo must die in order to save the love.
- The point is plain as a pike staff.

2. Consitution de la base de données

2.1. Acquisition des données

Puisqu'il est plus simple de trouver des données en langue anglaise, nous avons choisi de nous limiter à cette langue et avons utilisé des scripts Python de scrapping pour récolter des données sur différents sites internet.

Nous avons 3200 proverbes originaux anglais et 35000 proverbes en incluant des proverbes traduits d'autres langues et prévoyons de tester le modèle sur ces deux bases.

Nous également constitué une base de données de proverbes annotés avec leurs thèmes.

Voici quelques exemples :

```
{
  "topics": ["Advantage"],
  "proverb": "When a rich man caresses a poor man, he's going to take advantage of him."
}
{
  "topics": ["Advantage"],
  "proverb": "Every advantage has its disadvantage."
}
{
  "topics": ["Advantage", "Vain"],
  "proverb": "He is wise in vain who does not use his wisdom for his own advantage."
}
```

Liste 1. – Extrait des proverbes annotés

2.2. Script de chargement des données

Nous avons réalisé un script de téléchargement et traitement des données. Un exemple d'utilisation est donné dans le fichier `main.ipynb`, il suffit d'appeler le fonction `make_dataset.load_data()` qui renvoie les proverbes classés par sources.

```
proverbs_db.txt: 34142 proverbs
proverbs_db_only_english.txt: 2208 proverbs
proverbs_digest.txt: 1000 proverbs
Total: 37350 proverbs
Total length: 1853527 characters


Examples of proverbs:
Money's for buying and a horse is for riding.
A set of white teeth does not indicate a pure heart.
Time discloses the truth.
The earth has ears, the wind has a voice.
The fox will catch you with cunning, and the wolf with courage.
```

Liste 2. – Proverbes non annotés

De même, la fonction `make_dataset.load_data_with_topics()` renvoie un dataset de données annotées.

3. Création et entraînement du modèle

3.1. Création du modèle

Nous allons faire du fine-tuning à partir du modèle `facebook/opt-125` qui un modèle type GPT-3 basé sur l'architecture Transformers.

Nous testerons également de partir du modèle `TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T` qui est plus récent et contient près de dix fois plus de paramètres.

On charge le modèle et son Tokenizer à l'aide des fonctions `AutoModelForCausalLM.from_pretrained()` et `AutoTokenizer.from_pretrained()` de la librarie `transformers`.

On décide des sources de proverbes que l'on va utiliser, puis on les fusionne en une seule liste.

```python
selected_proverbs_groups = [
    "proverbs_db.txt",
    "proverbs_digest.txt"
]

proverbs = []
for group in selected_proverbs_groups:
    proverbs.extend(all_proverbs[group])
```

Liste 3. – Fusion des listes sélectionnées

On utilise ensuite la librarie `datasets` pour préparer ces données à l'entraînement. La tokenisation du dataset consiste à calculer la taille du proverbe le plus long et ajouter du padding aux autres pour uniformiser les tailles.

On utilise ensuite la librarie `peft` afin d'utiliser la technique LoRA (Low-Rank Adaptation). Ainsi, on ajoute un petit nombre de nouveaux paramètres entraînables au modèle afin de l'adapter à la nouvelle tâche.

```
LoraConfig(
    r=8,
    lora_alpha=16,
    target_modules=["q_proj", "v_proj"],
    lora_dropout=0.05,
    bias="none",
    task_type="CAUSAL_LM"
)
```

Liste 4. – Configuration LoRA

La méthode `get_peft_model` permet d'obtenir un nouveau modèle à partir de cette configuration de notre modèle initial.

3.2. Entraînement

A l'aide de la librarie `transformers`, on définit les paramètres d'entraînement:

```
TrainingArguments(
    output_dir="./results",
    per_device_train_batch_size=8,
    per_device_eval_batch_size=8,
    num_train_epochs=1,
    logging_dir='./logs',
    logging_steps=10,
    eval_strategy="no",
    save_strategy="epoch",
    report_to="none"
)
```

Liste 5. – Paramètres d'entraînement du modèle

Finalement, on met en commun notre modèle, nos paramètres d'entraînement, notre dataset tokenisé et notre tokeniser via la classe `Trainer` et on peut lancer l'entraînement avec la méthode `train()`.

3.3. Génération

Après entraînement du modèle, on crée un pipeline de génération.

```
generator = pipeline("text-generation", model=model, tokenizer=tokenizer)
```

Liste 6. – Pipeline de génération

On peut maintenant utiliser le modèle pour terminer un début de proverbes:

```
prompt = "A"
results = generator(prompt, max_length=max_length, num_return_sequences=3,
do_sample=True, temperature=0.7)

for i, result in enumerate(results):
    print(result['generated_text'])
```

Liste 7. – Génération de proverbes

Quelques proverbes obtenus:

- A nice dog is a dog.
- A man who keeps his family safe can never be found.
- A child that has no teeth is a coward.

4. Analyse des résultats

Les données de sortie étant des proverbes, elles sont difficilement évaluables et comparables, et sont sujettes à l'appréciation humaine.

Ainsi nous donnerons quelques résultats non filtrés à titre d'exemple, issus du fine tuning de différents modèles que nous commenterons. Puis dans une seconde partie, nous détaillerons une expérience que nous avons réalisé.

4.1. Entraînement sur les proverbes anglais

On utilise ici les deux sources :

- `proverbs_db_only_english.txt`
- `proverbs_digest.txt`

Pour rappel, cette base de données représente 3208 proverbes originaux en anglais.

4.1.1. Modèle de départ `facebook/opt-125m`

Proverbe	Commentaire
A man's only dead when he eats his own.	Intéressant
A man can make a woman forget her brother's death.	Pas un proverbe
A man's heart is full of gold.	Pas un proverbe
A good thing is a good thing.	Pas un proverbe
A woman's dream is to be a widow.	Etrange
A man dies in the wind, a horse dies in the wind	Pas un proverbe
A good example of a good example of a good example of a bad example of a bad example of a bad example.	Incohérent
A few days before you start a new one, you will be remembered for a long time.	Incohérent
A little of light can be a good thing	Pas un proverbe

Tableau 1. – En parant de « A »

Proverbe	Commentaire
Some people have a hard time keeping a family man.	Intéressant
Some people are good, some people are good.	Incohérent
Some people are lucky to live in a tree.	Intéressant
Some of them are to be proud of their first friend.	Intéressant
Some times the best is when you are there to be with.	Incohérent
Some people have a disease.	Etrange
Some people are better than others.	Intéressant
Some of us are better than others	Intéressant
Some people are too silly to forget.	Intéressant
Some of those is nothing compared to the sum of them	Intéressant

Tableau 2. – En parant de « Some »

Intéressants 9
Etranges 2

Pas un proverbe 5
Incohérents 4

On constate que certains résultats sont incohérents ou étranges. Certains proverbes ressemblent plus à des vérités générales et on trouve quelques proverbes intéressants.

4.1.2. Modèle de départ `TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T`

Les données sélectionnées ne sont pas assez importantes pour obtenir des résultats satisfaisants avec ce modèle.

Proverbe	Commentaire
This is the story of an old woman.	Pas un proverbe
This is the way of the world.	Pas un proverbe
This is the best.	Pas un proverbe
This is the way to get your money's worth.	Pas un proverbe
This is the day.	Pas un proverbe
This is what you call a new dress.	Pas un proverbe
This is a story about a little girl.	Pas un proverbe
This is the day the Lord hath spoken. It is the day that the Lord hath spoken.	Pas un proverbe
This is what happens when we start out on the wrong path.	Pas un proverbe
This is my favourite drink.	Pas un proverbe

Tableau 3. – En parant de « This »

Intéressants 0
Etrange 0
Pas un proverbe 10
Incohérents 0

Le modèle de base étant plus important et donc déjà plus entraîné, il ne crée pas de résultat incohérent comme le précédent, mais le dataset utilisé est trop petit pour créer des proverbes intéressants. Peu de proverbes commencent par « Some » dans le dataset. ($\frac{4}{3208} \approx 0.12\%$)

4.2. Entraînement sur les proverbes traduits

4.2.1. Modèle de départ `facebook/opt-125m`

On entraîne d'abord sur **20000** proverbes (57.14%).

Proverbe	Commentaire
Some men are good at it, some are bad at it, and some are good at it.	Incohérent
Some day people will get one.	Pas un proverbe
Some days a man wears a leather jacket and a woman wears a leather garment.	Pas un proverbe
Somehow the worst person in the world can make the best at things.	Étrange
Somehow, I am not a thief.	Pas un proverbe
Some people are more beautiful than others.	Pas un proverbe
Some day, the sun will rise on a mountain.	Incohérent
Someones got a good idea and they're good at it.	Pas un proverbe
Some men have a heart to die for.	Intéressant
Somehow it looks like a human's ear.	Incohérent

Tableau 4. – En parant de « Some »

Proverbe	Commentaire
A dog is a beast of a mind.	Incohérent
A man who makes his wife cry is a thief.	Étrange
A man's heart is not a dog's tongue.	Incohérent
A girl may be a good girl, but she has a better chance of getting married.	Étrange
A man who has never seen a woman is a man.	Incohérent
A house that is just a house is not a house that is a house.	Incohérent
A man who will not sleep for an hour, will not sleep for an hour.	Incohérent
A man cannot take a wife.	Pas un proverbe
A man who eats meat will not be a judge.	Étrange
A bit of good luck in life is better than nothing.	Intéressant

Tableau 5. – En parant de « A »

Pas ouf ces résultats, faudrait réentrainer sur moins de données et savoir l'expliquer

---

4.2.2. Modèle de départ `TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T`

La RAM disponible ne nous a permis que de sélectionner un maximum de 5000 proverbes.

Cependant, on obtient quand même des résultats plus intéressants qu'avec 3208 proverbes

Proverbe	Commentaire
Some men are born to lead, and some to follow.	Intéressant
Some men are so proud of their looks that they never look at the rest of their faces.	Étrange
Someone will not be able to find a wife whom his father chooses.	Pas un proverbe
Some things are good for the body, but not for the stomach.	Intéressant
Some are born with the gift of knowledge, and some with the gift of ignorance.	Intéressant
Some folks make their own pies, and some do not.	Pas un proverbe
Some people call the wind the sun's enemy.	Incohérent
Some are wiser than they know. Humor as a way of life is also a part of the world I have to live in.	Incohérent
Some are born great, but some become great by their wit.	Intéressant
Someone is not what you think about him; you think about him.	Incohérent

Tableau 6. – En parant de « Some »

Proverbe	Commentaire
A man's name is his life.	Intéressant
A man who has never taken a step should never be trusted.	Intéressant
A great king cannot be a great man.	Intéressant
A woman is hard to handle, but easy to cheat.	Étrange
A dog's nose is better than a man's eyes.	Intéressant
A woman who has a good head cannot be a fool.	Intéressant
A bird from a city is a nesting place for many.	Incohérent
A little knowledge is better than a great ignorance.	Intéressant
A man who speaks of a hundred will be thought a hundred.	Incohérent
A man should not be too ambitious.	Intéressant

Tableau 7. – En parant de « A »

Intéressants 11
Etranges 2
Incohérents 5
Pas un proverbe 2

Comme précédemment, ce modèle produit peu de résultat incohérents. On constate qu'on obtient des résultats plus intéressant avec ce dataset plus important, plus adapté à la taille du modèle de base.

4.3. Expérience avec des humains

A titre d'expérience, nous avons rassemblé quelques dizaines de proverbes originaux et générés par nos modèles. (`facebook/opt-125m` sur $\approx$ 15000 proverbes)

5. Conclusion