Travail pratique # 2

Traitement automatique du langage avec des Transformers

Automne 2025

Proposé par Luc Lamontagne et Vincent Martineau

OBJECTIFS :

- Utiliser des modèles de langue *Transformer préentraînés* de la librairie *HuggingFace*.
- Mener des expérimentations avec des encodeurs T*ransformers* pour la classification de textes.
- Comprendre comment sont entraînés les décodeurs *Transformers* (modèles de langue génératifs) et mener des expérimentations sur une tâche de question-réponse.
- Évaluer la performance des modèles sur des jeux de données et faire l'analyse des résultats.

INSTRUCTIONS :

- Matériel disponible le 21 octobre 2025.
- Ce travail sera noté sur 100 et vaut 30% de la note du cours.
- Remise : À remettre le 21 novembre sur MonPortail, le tout compressé en format ZIP.
- Contenu de la remise :
    - Les *notebooks* Jupyter bien documentés avec des analyses détaillées des résultats.
    - Les fichiers de résultats des tâches de génération (c.-à-d. les textes générés).
- Références : Chapitres 7, 8, 9, et 10 de la 3e édition du livre de Jurafsky et Martin.
- Seul langage de programmation autorisé: Python 3.
- Ressources disponibles sur le site du cours:
    - Des fichiers de textes pour mener vos expérimentations.
    - Des *notebooks* pour démarrer chacune des 4 tâches.
- Librairies autorisées :
    - Modèles *Transformers* : ceux sur *Hugging Face* seulement.
    - Tokenisation : Voir les consignes de chacune des tâches.
    - Normalisation de textes : Aucune normalisation sauf les lettres en minuscule.
- Note : Les tâches 2, 3 et 4 sont liées entre elles et il est important de comparer les résultats de chacune de ces tâches.
- Il est recommandé pour ce travail d'avoir un environnement avec une carte graphique GPU compatible avec HuggingFace/Pytorch. Si votre ordinateur n'en possède pas, vous pouvez utiliser *Google Colab* pour exécuter les *notebooks*.

TÂCHE 1 – PRÉDICTION DU SCORE NOVA DES PRODUITS ALIMENTAIRES AVEC DES *TRANSFORMERS*

On reprend, comme au premier travail, la tâche de prédiction du score NOVA de produits alimentaires. Cependant, de nouveaux jeux de données ont été produits. Le corpus de textes contient 3 partitions :

- Un fichier d'entraînement - *data/t1_nova_train.json*
- Un fichier de validation - *data/t1_nova_dev.json*
- Un fichier de test - *data/t1_nova_test.json*

Utilisez la librairie *HuggingFace* pour accomplir cette tâche. On vous demande plus spécifiquement d'utiliser 2 modèles : le modèle ***bert-base-uncased*** et un **modèle encodeur multilingue** préentraîné de

votre choix. Le 2e modèle peut être une variante de Bert, mais cela n'est pas exigé. Me consulter en cas de doute pour valider votre choix. Utilisez le *notebook* ***t1_classification_nova.ipynb*** pour mener vos expérimentations. Toutes les consignes liées à cette tâche sont décrites dans l'en-tête du *notebook*. Assurez-vous de présenter clairement les résultats obtenus pour chacun des modèles et faites-en l'analyse. Comparez les résultats obtenus avec les 2 modèles.

TÂCHE 2 – QUESTION-RÉPONSE AVEC LE MODÈLE GPT-2 PRÉENTRAÎNÉ SANS ADAPTATION

Cette tâche consiste à utiliser un modèle de langue génératif déjà préentraîné et à observer ses limites lorsqu'on lui pose des questions sur un domaine hors sujet de ses données d'origine - l'univers de *Sherlock Holmes*. Le modèle que nous avons sélectionné est **GPT-2 Medium** qui contient 355 millions de paramètres. Pour cette tâche, vous utilisez ce modèle uniquement en mode inférence (génération de textes), ce que consiste à construire un *prompt* simple, à générer des réponses avec le modèle, et à évaluer la pertinence des résultats. Aucun entraînement de modèle n'est effectué dans cette tâche. Un fichier de questions de test (***questions_sherlock.json***) est rendu disponible pour mener vos expérimentations. Les consignes pour cette tâche sont décrites plus en détail dans le *notebook* ***t2_llm_gpt2.ipynb*** que vous utiliserez pour mener vos expérimentations.

TÂCHE 3 – QUESTION-RÉPONSE AVEC POURSUITE DU PRÉENTRAÎNEMENT DE GPT-2

L'objectif de cette tâche est de poursuivre le préentraînement du modèle GPT-2 Medium utilisé à la tâche précédente et d'évaluer son impact sur les réponses générées par le nouveau modèle. Une liste de livres à utiliser pour le préentraînement ainsi qu'une fonction pour monter leur contenu en mémoire est disponible dans le *notebook* ***t3_llm_pretraining.ipynb***. Vous trouverez des consignes plus détaillées dans ce *notebook*. À partir du jeu de questions de test rendu disponible (le même que celui de la tâche précédente), décrivez vos observations sur les textes générés par le nouveau modèle. Comparez les réponses à celles de la tâche précédente et interprétez le comportement des modèles. Veuillez noter qu'il est important de sauvegarder le modèle préentraîné dans cette tâche (ainsi que son tokenizer) car ils sont réutilisés pour la prochaine tâche.

TÂCHE 4 – QUESTION-RÉPONSE AVEC UNE VERSION DE GPT-2 AFFINÉE PAR INSTRUCTIONS

La dernière étape de cette trilogie consiste à affiner par instructions le modèle de langage préentraîné que vous avez construit dans la tâche 3 et à évaluer la qualité des réponses du nouveau modèle. À partir du notebook ***t4_llm_instruction.ipynb***, vous faites le post-entraînement avec des instructions indiquant au modèle comment accomplir des tâches simples. Nous utilisons un sous-ensemble du jeu de données Alpaca (5000 instructions) pour mener le post-entraînement. La fonction pour créer ce jeu d'instructions est rendue disponible. Faites l'analyse des réponses du modèle et présentez vos observations par rapport aux réponses des tâches précédentes. Expliquez ce que vous retenez de l'impact du préentraînement et du post-entraînement du modèle GPT-2

ÉVALUATION DU TRAVAIL

T1 – Classification NOVA avec des *transformers* encodeurs	25%
T2 – Question-réponse avec GPT-2 préentraîné sans adaptation	20%
T3 – Question-réponse avec poursuite du préentraînement de GPT-2	25%
T4 – Question-réponse avec une version de GPT-2 affinée par instructions	30%