Rapport de résultats Projet 2

Groupe: Antoine JULIEN, Gwenegan BERTHO, Laura LACAMBRA

Partie 1

Données

On utilise le dataset SQuAD v2. On extrait environ 5000 échantillons pour le dataset de d'entrainement, et 6000 échantillons pour le dataset de validation.

Choix des modèles QA

On cherche à comparer l'évaluation des deux modèles suivants : un basique ***distilbert-base-uncased*** (qu'on fine-tune), et le modèle ***mvonwyl/distilbert-base-uncased-finetuned-squad2.***

Mesures des performances

On retrouve ces mesures dans NLP_Sujet2_Part1.pdf, page 16 et page 21.

```
{'HasAns_exact': 31.59675236806495,
 'HasAns_f1': 36.900098704626316,
 'HasAns_total': 2956,
 'NoAns_exact': 59.71150620597115,
 'NoAns_f1': 59.71150620597115,
 'NoAns_total': 2981,
 'best_exact': 50.27791814047499,
 'best_exact_thresh': 0.0,
 'best_f1': 50.36720472152805,
 'best_f1_thresh': 0.0,
 'exact': 45.713323227219135,
 'f1': 48.353830515559274,
 'total': 5937}
```

Evaluation du distilbert-base-uncased

```
{'HasAns_exact': 62.31393775372125,
 'HasAns_f1': 70.89900675793425,
 'HasAns_total': 2956,
 'NoAns_exact': 58.705132505870516,
 'NoAns_f1': 58.705132505870516,
 'NoAns_total': 2981,
 'best_exact': 60.55246757621695,
 'best_exact_thresh': 0.0,
 'best_f1': 64.80778786010966,
 'best_f1_thresh': 0.0,
 'exact': 60.50193700522149,
 'f1': 64.77639615571039,
 'total': 5937}
```

Evaluation du distilbert-base-uncased-finetuneed-squad2

On voit que le modèle 2 est bien meilleur, que ce soit au niveau du f1 score ou du nombre de réponses exactes.

Exemples de mauvaises classifications

On retrouve des exemples de prédiction associée à leur vérité de terrain page21-22.

Prenons deux exemples de mauvaises prédictions.

```
Prediction  i = 3
Expected : Gaussian primes
Obtained :  ['4k + 3', '4k + 3', 'Z']
```

Exemple 1.

On peut supposer que les chiffres et les opérations faussent les résultats.

```
Prediction  i = 9
Expected : Saul Bellow
Obtained :  []
```

Exemple 2

Il se peut que cet exemple soit mal classifié car notre modèle a vu trop de titre vide associé à un même genre de texte, dont celui-là. Il faudrait augmenter la diversité du dataset pour réduire les erreurs de ce type.

Partie 2

- Chargement des données

Le jeu de donnée de validation de SQUAdV2 contient trop peu de contextes, j'utilise alors ceux de DBPedia pour ajouter plus de diversité. Au total 10 000 contextes de DBPedia sont ajoutés pour s'approcher des 10 000 contextes demandés.

- Projection et création des embeddings

Une fois le bruit ajouté l'ensemble du dataset est encodé à l'aide du modèle ms MARCO : msmarco-roberta-base-v3 avec une similarité cosin.

Sur l'ensemble des questions disponibles ce modèle nous permet d'avoir une **MMR de 85.7%.**

- Nearest Neighbour

Pour l'implémentation de l'approximation nearest neighbour nous avons choisi d'utiliser la bibliothèque annoy. On utilise alors les différents contextes comme moyen d'accélérer alors notre recherche d'une réponse. Les contextes utilisés sont ceux que l'on trouve dans le dataset SQuAD v2 qui a été enrichie avec des contextes issus de DBpedia.

Lorsque l'on fournit un nouveau contexte à notre algorithme nearest neighbour, celui-ci va nous retourner les N meilleurs résultats dans notre dataset. Cela nous permet alors de faire un premier filtrage par contexte. On voit ci-dessous le contexte soumis à notre algorithme ainsi que les 10 meilleures correspondances dans notre dataset avec leur score associé.

```
Input question: What is the original meaning of the word Norman?
Results (after 0.059 seconds):
        0.950   What does Norman mean in the article when saying "Near East"?
        0.929   What is the Norman term for "waterfowl?"
        0.923   What word is derived from the medieval English word?
        0.916   After the Norman Conquest, Latin words entered English via what language?
        0.912   What book did Sir Henry Norman write after traveling to the Far East?
        0.912   How did the Norse borrow words from Old English?
        0.911   What is the name of the English language's earliest form?
        0.911   A formula in Anglo-Norman Law Greek indicates what?
        0.910   To what word does the Anglo Saxon translate?
        0.910   What were examples of morphology in Old English and Old Norse?
```

Nous n'avons pas pu correctement connecter les différentes parties mais idéalement nous aurions voulu faire un premier encodage de la query sous forme de contexte, approximer les contextes les plus similaire avec le nearest neighbour et enfin utiliser le QA model pour affiner la recherche et fourni la réponse la plus adéquat dans le minimum de temps.

Partie 3

Le code de la partie 2 est repris et adpaté pour faire un système de QA. Le script QA.py permet de poser des questions dans un invité de commande et de tester la performance du modèle.

Pour améliorer la performance, l'embedding précédemment calculé a été sérialisé et est chargé directement au démarrage du script.

Conclusion et commentaires :

A cause de moyen limités les embedding n'ont pu être calculés que sur un nombre restreint de contextes. Ceci a pour effet de rendre le système assez performant lorsqu'il rencontre des questions déjà posés mais le fait dérailler assez vite qu'on aborde un topic avec peu de lien aux données d'entrainement.