LYSL001 - MACHINE CREATIVITY

CONSIGNES DU PROJET FINAL

En guise de projet final, vous êtes invité.e.s à contribuer au wiki collaboratif du cours (*Wikmatic*) en y ajoutant une page wiki que vous aurez générée automatiquement. Ce document reprend les attendus, les détails pratiques du rendu, les ressources sur lesquelles vous pouvez vous baser., et l'évaluation globale du projet.

Si des questions persistent, n'hésitez pas à les poser dans le forum FAQ dédié sur la page iCampus.

I. Attendus

Votre objectif est de parvenir à générer du texte. La génération de votre texte est cependant soumis à une contrainte : il faut que votre texte imite autant que possible le format de pages Wiki, c'est-à-dire qu'il se base sur la syntaxe wikicréole.

La syntaxe wikicréole est le langage utilisé pour formater les pages de bon nombre de wikis en ligne (Wikipedia, Wiktionnaire…). Voyez cela comme une sorte d'équivalent au HTML, mais destiné à être intéprété par les interfaces Wiki, permettant de générer toute une série de formatage : titre, gras, italique, etc. Pour un détail de ces traitements, reportez-vous sur la page correspondante dans le wiki sur le cours iCampus, ou cliquer sur le bouton "Modifier" ou "Modifier le code" de n'importe quelle page de wiki.

Le texte que vous allez générer (c'est-à-dire la pseudo page, ou le pseudo article) n'a pas besoin d'être parfait dans l'absolu : le texte n'a pas à être strictement cohérent, les liens réels, les balises correctes, etc. L'important est que vous parveniez à générer automatiquement ce texte. Ainsi, la quasi intégralité du texte que vous mettrez dans le wiki (voir ci-dessous à la section III) doit avoir été générée par votre système. Si vous jugez certaines modifications nécessaires, par exemple pour augmenter sa lisibilité une fois intégrée au Wiki sur iCampus (voir section III), vous pouvez opérer des modifications de l'output, mais ces modifications doivent être minimes, et elles doivent être précisément renseignées dans le dossier qui accompagne votre rendu.

Pour générer un texte intégrant cette syntaxe, vous allez devoir entraîner votre système sur un corpus intégrant cette syntaxe. La section II (Ressources) revient sur les ressources que vous pouvez utiliser si vous le souhaitez.

Indépendamment de la page Wiki elle-même que vous allez produire, il est attendu de vous un dossier succinct d'analyse (max. 15p, sauf exception). Ce dernier est une trace du travail que vous avez effectué pour générer le texte. Il doit retracer et décrire les différentes étapes de production de la page. Il évoque ainsi la constitution du corpus utilisé pour entraîner votre système de génération de texte (comment l'avez-vous constitué, quels pré-traitements avez-vous appliqué le cas échéant, quelles sont ses caractéristiques ? etc), le choix du système de génération (quel algorithme et quels paramètres avez-vous choisi ?), ainsi que toute intervention manuelle sur le texte généré (avez-vous corrigé/rajouté des balises fermantes à certains endroits ? etc). Le dossier doit offrir une réflexion critique sur la démarche globale : choix et limites du corpus, limite du système, problèmes

identifiés, solutions qu'il aurait été souhaitable de mettre en place… L'important est d'avoir un regard critique sur votre production, de montrer que vous avez compris ce que vous faisiez.

Le Wikmatic n'a pas besoin d'être cohérent dans sa globalité, et peut donc contenir des pages de type encyclopédique, dictionnaire, nouvelles, sur des sujets variés. Les pages elles-mêmes n'ont pas besoin d'être crédibles. L'important est de montrer que vous savez générer du texte en reprenant certaines contraintes – ici formelles, liées au wikicréole.

II. Ressources

Le corpus d'entraînement doit de fait contenir du wikicréole, et pourra être tiré des ressources suivantes (liste non exhaustive) :

- Wikipedia

- WikiNews

- Wiktionnaire

- WikiQuote

- Wikiversité

Vous pouvez aussi utiliser tout wiki en ligne, de type fandom.com ou autre, dont les articles reposent sur du wikicréole. Vous pouvez vérifier l'utilisation du wikicréole en cliquant sur l'onglet "Modifier" généralement accessible sur les pages de ces wiki.

Pour extraire les articles qui constitueront votre corpus d'entraînement, vous pouvez évidemment récupérer manuellement le contenu des articles en question. Cette approche est cependant longue, et limite la taille de votre corpus.

Vous pouvez aussi passer par l'outil d'extraction intégré dans les Wiki relevant du projet MediaWiki. Ces derniers comportent des pages spéciales, généralement accessibles à l'adresse https://xxxxxx/wiki/Spécial:Pages_spéciales (où xxxx correspond à l'adresse principale du wiki). Parmi ces pages spéciales, vous trouverez une section intitulée "Outils pour les pages", et notamment la page "Exporter des pages". En suivant les instructions proposées sur cette page, vous pouvez alors récupérer une fichier xml regroupant l'ensemble des pages demandées. Ce fichier regroupe diverses métadonnées sur le wiki et sur les pages, mais surtout le contenu des articles (entre les balises <text></text>). Un programme d'extraction du titre et du texte de chaque article est mis à votre disposition sur iCampus. Ce programme repose sur le module 'mwxml' à installer en amont. Contactez-moi rapidement si vous rencontrez le moindre souci pour son utilisation.

Vous pouvez évidemment choisir d'utiliser les dumps officiellement mis à disposition par MediaWiki (https://dumps.wikimedia.org/backup-index.html) mais ces fichiers sont particulièrement lourds, et demandent un pré-traitement plus conséquent pour en extraire les articles de votre choix.

Concernant la méthode de génération elle-même, vous pouvez utiliser au choix les codes présentés en cours, ou de algorithmes que vous auriez pu tester/rencontrer ailleurs.

III. Détails pratiques

La page et le dossier peuvent être en français ou en anglais.

Il s'agit d'un travail **individuel** ou par **groupe de deux** (maximum). Le rendu du travail contiendra les éléments suivants :

- **La page wiki elle-même**

- **Un dossier succinct d'analyse** (max 15p, sauf exception)

- **Les ressources** (corpus et code)

Le rendu du dossier et des ressources se fera sous la forme d'une archive à déposer sur la plateforme de dépôt dédiée. L'intégration des pages générées au Wikmatic se fera par vos soins selon les consignes renseignées sur le Sommaire du wiki et rappelées ici. L'ensemble doit être rendu pour le 29 janvier 2023, 23h59 au plus tard. Si vous rencontrez des difficultés à respecter la date limite, merci de me contacter en amont.

L'ajout de votre texte généré au wiki du cours se fait de la façon suivante : aller sur le wiki et dans le menu déroulant, sélectionner l'option "Modifier". Cela vous permettra de modifier le wikicode du sommaire. Chaque groupe (ou individu) devra alors intégrer sa page à l'aide de la ligne :

- [[NOM Prénom – n° étudiant]]

Où NOM et Prénom correspond à votre propre nom et prénom. Attention à ne pas oublier les crochets ! Cela permet de créer une nouvelle page (voir la page décrivant le wikicréole sur le Wiki).

Par exemple, si je veux créer la page que j'ai générée, je vais dans la version modifiable de la page de sommaire, et j'ajoute la ligne "- [[WAUQUIER Marine - 12345678]]". Je valide et lorsque je repasse en mode Aperçu de la page, votre ajout apparaît en rouge (sans les crochets). Si vous passez la souris sur les mots en rouge, vous voyez que vous pouvez cliquer. En cliquant, cela vous renvoie vers une page vierge, dans laquelle vous pouvez coller le texte que vous avez généré automatiquement. Pensez à sauvegarder la page ainsi créée.

Si le travail a été en groupe de 2, le format d'inclusion est donc :

- [[NOM Prénom – n° étudiant ; NOM Prénom – n° étudiant]]

IV. Evaluation

Le projet contera pour 60% de la note du cours.

L'évaluation se fera principalement sur le dossier lui-même.

Les ressources (code et corpus) ne seront soumis qu'à titre indicatif, et contribueront à la notation du dossier. Ils garantissent l'originalité et la reproductibilité du travail. Si le code repose sur un notebook en ligne, merci de le rendre accessible et de transmettre explicitement le lien.

La page ne sera pas évaluée sur son contenu à proprement parler - que ce soit en termes de factualité, de diversité lexicale ou de cohérence syntaxique (par exemple). De par la nature de l'exercice, il n'est pas attendu que le texte généré soit cohérent ou propre.