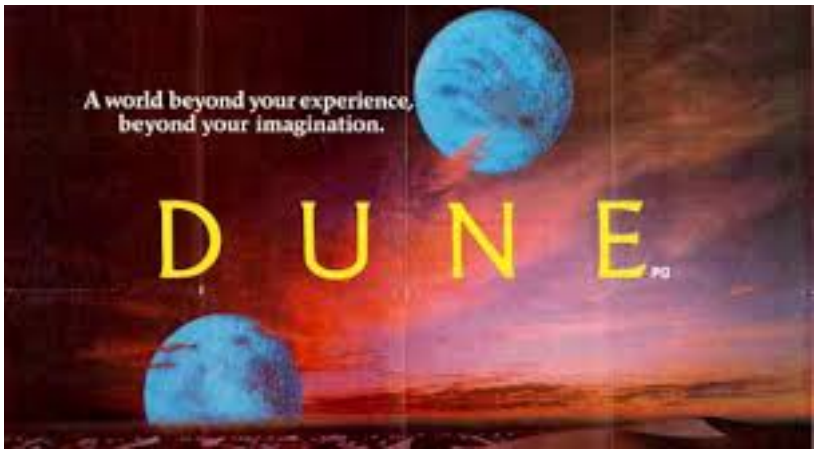
# Projet de Fouille de texte

Résumer automatiquement une œuvre littéraire

Auteur du sujet : J. Velcin




*"Deep in the human unconscious is a pervasive need for a logical universe that makes sense. But the real universe is always one step beyond logic."*
– **Frank Herbert, Dune**

## Objectif principal

Les outils informatiques et statistiques ont montré un réel potentiel pour l'étude d'œuvres littéraires dont le volume rend l'analyse manuelle fastidieuse voire impossible. L'objectif principal de ce projet consiste à proposer une manière de *résumer* une œuvre de manière à donner un bref aperçu du contenu à une personne qui n'aurait pas le temps de la parcourir dans son ensemble. Il s'agit d'une forme de *Reader's Digest* qui pose un certain nombre de questions, parmi lesquelles :

- Quelle forme doit prendre la restitution créée automatiquement ? S'agit-il de conserver une forme d'expression en langage naturel (une suite de phrases, par exemple) ou faut-il imaginer d'autres moyens de réaliser la restitution (nuage de mots-clefs, thématiques, réseau de personnages, etc.) ?
- S'agit-il de créer une méthode valable pour tous les types d'œuvres littéraires, ou certaines méthodes peuvent-elles être conçues de manière ad-hoc pour tel ou tel style ?
- Quel degré de détail doit être choisi pour réaliser cette restitution ? Le résumé peut en effet prendre la forme d'une simple phrase ou d'un résumé composé de plusieurs pages, elles-mêmes intégrant plusieurs points de vue différents sur l'œuvre ?

Le sujet de ce projet est volontairement ouvert et laisse donc la possibilité de faire de nombreuses propositions. Il faut cependant garder à l'esprit trois éléments essentiels :

1. le projet doit être l'occasion d'expérimenter des techniques abordées durant le cours de fouille de textes, que celles-ci aient été vues en détail (modélisation thématique) ou juste aperçues (résumé automatique),
2. le barème prendra en compte le niveau de difficulté choisi dans la réalisation du projet (un nuage des mots-clefs les plus fréquents sera jugé très facile, tandis que générer un résumé totalement original avec des approches génératives sera jugé plus difficile),
3. il ne s'agit pas de développer une méthode ad-hoc pour cet unique livre mais de garder l'idée que ce vous proposez pourra être testé sur un autre livre.

Pour finir, l'œuvre littéraire est imposée : il s'agit de la série Dune, écrite par F. Herbert. Vous aurez à votre disposition le premier ouvrage de l'écrivain, mais le travail proposé sera testé sur d'autres livres de la série (avec le même format en entrée). Les données étant protégées par copyright, elles sont fournies de manière individuelle et ne doivent être diffusées sous aucun prétexte.

## Plusieurs pistes à explorer

Pour atteindre cet objectif, voilà ci-dessous quelques pistes de départ. Attention, il n'est pas attendu à ce que vous poursuiviez toutes ces pistes mais plutôt que vous trouviez une manière cohérente de présenter le résumé à partir de *quelques* techniques.

- Une première piste évidente est d'avoir recourir à des modèles thématiques. L'une des difficultés est alors de réfléchir à la manière de structurer les thématiques en fonction de l'organisation du livre. Il s'agit également de rendre les thématiques accessibles via des techniques d'étiquetage, telles que celles vues en cours.
- Une autre piste consiste à prendre en compte les entités qui peuplent le livre, tels que les personnages ou les lieux. Les possibilités sont ensuite très nombreuses, entre réaliser une cartographie des lieux évoqués dans le livre, le réseau social des personnages et de leurs interactions, tout cela en prenant en compte la dimension temporelle.
- Les pistes précédentes privilégient une représentation synthétique sous forme de mots-clefs et de graphes. Une idée pour aller plus loin consiste à extraire un certain nombre de phrases qui paraissent importantes pour la compréhension du romain. Les phrases peuvent être considérées importantes car elles traitent des personnages importants, ou qu'elles abordent ensemble l'intégralité des thématiques découvertes.
- Une piste certainement plus difficile consisterait à essayer de *générer* automatiquement des phrases, basées soit sur des patrons prédéfinis (par ex. une phrase recensant les personnages les plus importants pour chaque partie, une autre sur leurs relations familiales), soit générées à l'aide de modèles génératifs comme des réseaux de neurones artificiels.
- Enfin, les pistes évoquées précédemment ne prennent pas en compte de nombreux aspects qui pourraient apporter un réel plus à votre travail. Il peut s'agit de la prise en compte des émotions dans le texte (très présentes dans le livre), de la distinction entre monologues intérieurs et dialogues, etc. D'autre part, vous avez le droit d'utiliser des connaissances liées à l'œuvre de F. Herbert (voir par ex. le wiki très documenté : http ://dune.wikia.com/wiki/Main_Page).

## Ce qu'il faut rendre

Vous devez rendre un fichier archive avec votre numéro de groupe (ex. : `projet_groupeX.zip`) qui contient les éléments suivants :

1. Un document PDF retraçant les différentes étapes de votre travail.
   (a) Autant que possible, argumentez votre démarche. Si vous avez tenté plusieurs approches avant de fixer votre choix définitif, une description même succincte des différents essais et de ce qui a motivé leur abandon serait apprécié.
   (b) On s'attend à trouver dans votre rapport : (1) une description détaillée de la solution finalement retenue pour construire le résumé ; (2) les résultats obtenus, qui comporteront à minima une visualisation permettant de rendre compte de l'histoire de manière synthétique d'un certain point de vue ; (3) une discussion sur ces résultats, sur leur pertinence, sur ce qui permettrait de pousser l'analyse plus loin.
   (c) Si cela s'avère judicieux, n'hésitez pas à inclure des éléments de bibliographies dans votre document.
   (d) Vous pouvez éventuellement compléter l'article principal avec des annexes (par ex. composé de tableaux et figures additionnels) si vous le jugez utile.

   Ce document doit prendre la forme d'un article scientifique de huit pages maximum, en respectant le format double colonnes IEEE [1].
2. Le programme utilisé pour la réalisation du projet. Le langage choisi est laissé à l'appréciation des étudiants mais il faut s'assurer qu'il ne pose pas de problème de portabilité. Il doit être suffisamment clair et abondamment commenté, étant éventuellement réparti sur plusieurs fichiers pour en faciliter sa compréhension. Il doit être auto-suffisant, c'est-à-dire que le correcteur doit pouvoir reproduire facilement et sans problème les opérations décrites dans le rapport.

1. https ://www.ieee.org/conferences/publishing/templates.html

La note finale dépendra bien sûr de la qualité du travail rendu, que ce soit en terme de rédaction du rapport ou du code, mais également de l'originalité et du niveau de complexité du projet. Il n'est bien sûr pas nécessaire de réaliser toutes les pistes proposées car l'important est de mener à bien celles choisies le mieux possible.

La date de rendu est fixée au **mardi 18 décembre**. Les documents (rapport et code) sont à envoyer à l'adresse suivante : **julien.velcin@univ-lyon2.fr**.

Bonne chance !