SAE-S1-2021-Projet individuel : Nuages de mots

Aurélie Leborgne, Véronique Richard, Murielle Torregrossa, Eric Wessler

2021-2022

*Vous **ferez le choix** de réaliser ce projet seul ou en binôme. Dans les deux cas, nous attendons le même travail. Par contre, dans le cas d'un travail en binôme, vous serez les seuls à assumer le fonctionnement de votre binôme.*

1 SAE 1.5 : Identifier les besoins métiers des clients et des utilisateurs (responsable : Éric Wessler)

Choisissez l'un des 5 sujets suivants.

Sujet 1 : La 4e de couverture d'un livre (ou son résumé) donne-t-elle un aperçu fidèle du texte ?

Objectif :
Vérifier s'il existe une corrélation entre la présentation d'un livre sur sa 4e de couverture (parfois, c'est un résumé) et le vocabulaire mesuré objectivement dans ce livre.
À cette fin, il faudra comparer cette 4e de couverture avec les résultats de la mesure statistique (présentés sous forme d'un nuage de mots).

L'ouvrage choisi peut être un récit de voyage, un roman, une biographie, etc. Votre choix devra être validé par M. Wessler, qui vous fournira lui-même la 4e de couverture à comparer.

Choix possibles sur :
- **Wikisource :** `https://fr.wikisource.org/wiki/Wikisource:Accueil`
- **Frantext, avec votre authentification Unistra :** `https://frantext.scd-rproxy.u-strasbg.fr/`

Sujet 2 :
Dans quelle mesure les textes d'un même auteur ou d'une même époque obéissent-ils à un déterminisme ? (Comparaison de textes littéraires.)

Objectif :
Analyser plusieurs textes du même auteur ou de la même décennie, et montrer les points communs et les différences dans le vocabulaire utilisé (qui s'expliquent peut-être en fonction du genre, de l'auteur, du sujet, etc.)
Le travail devra se faire sur 3 textes au moins. Il est conseillé de choisir 3 ou 4 textes longs (romans, etc.) ou une cinquantaine de textes courts (poèmes, articles de presse, par exemple).

A cette fin, il faudra donc présenter :
- **Un nuage de mots par texte analysé**
- **Un nuage de mots commun, permettant de visualiser les mots communs aux textes.**

Choix des textes sur Wikisource : `https://fr.wikisource.org/wiki/Wikisource:Accueil`

Sujet 3 :
La différence entre le droit civil et le droit pénal

Objectif :
Observer une différence de vocabulaire entre le code civil et le code pénal, pour mieux expliquer ensuite la différence de nature entre ces deux branches du droit. (Il s'agit d'un point du programme de droit du semestre 2.)
À cette fin, on produira un nuage de mots par code, et on procédera à une comparaison de ces deux résultats.

Les codes (dans leur version en vigueur actuellement) sont accessibles sur le site Legifrance : `https://www.legifrance.gouv.fr/codes/texte_lc/LEGITEXT000006070721?etatTexte=VIGUEUR&etatTexte=VIGUEUR_DIFF` **et** `https://www.legifrance.gouv.fr/codes/texte_lc/LEGITEXT000006070719?etatTexte=VIGUEUR&etatTexte=VIGUEUR_DIFF`

Sujet 4 :
Les vœux des Présidents de la République depuis 1974

Objectif :
Dire quels sont les thèmes majeurs, quelles sont les idées dominantes, dans les messages de vœux de Nouvel An adressés aux Français par les présidents de la République chaque année.
A cette fin, on produira un nuage de mots pour l'ensemble. Mais, pour les étudiants volontaires, il est également possible d'introduire un paramètre de temps : un nuage de mots par période (découpage libre), ou un seul nuage de mots qui se modifie automatiquement pour refléter l'évolution temporelle.

Textes disponibles sur Vie-publique : `https://www.vie-publique.fr/discours-dans-lactualite/269998-les-voeux-des-presidents-de-la-republique-depuis-1974`

Sujet 5 :
Sur quoi notre attention se focalise-t-elle quand nous écoutons quelqu'un ?

Objectif :
Analyser la manière dont un énoncé oral est perçu (par ceux qui l'écoutent) : les éléments retenus sont-ils un reflet fidèle des éléments que cet énoncé contient objectivement ?
À cette fin, on procédera de la façon suivante :

- **Audition d'un énoncé audio par au moins 4 auditeurs**
- **Questionnaire à remplir par les auditeurs après l'écoute**
- **Transcription textuelle de l'énoncé audio**
- **Mesure statistique du vocabulaire présent dans cet énoncé**
- **Comparaison entre le nuage de mots ainsi obtenu et les réponses au questionnaire**

Le choix d'un énoncé audio doit être validé par M. Wessler. La durée de l'énoncé doit être comprise entre 3 et 5 minutes, et le locuteur doit être unique.

Après avoir choisi un des sujets présentés ci-dessus, vous utiliserez la SAÉ 1.2 pour extraire les mots les plus fréquents d'un texte.

2 SAE 1.2 : Appréhender et construire des algorithmes (responsable : Murielle Torregrossa)

Consigne : Les organigrammes et les commentaires sont attendus. Merci d'écrire vos propres fonctions.

Contexte

Pour se faire une idée d'un article de presse, d'un livre, etc., il est souvent utile de se référer au résumé ou aux mots-clés. L'objectif de ce projet est de réaliser un programme d'extraction automatique de mots clés, à partir d'un fichier en format txt.
Un ensemble de fichiers de test vous est fourni dans le répertoire *fichier*. Vous y trouverez aussi les fichiers utiles pour les différentes étapes de résolution.

Question 1 :

Écrire un ensemble de fonctions, permettant, à partir d'un fichier .txt d'obtenir l'ensemble des mots du texte ainsi que le nombre de fois où ils apparaissent.

Question 2 :

On se rend compte que beaucoup de mots n'apportent pas d'informations sur le contenu lui même : *le*, *une*, *est*, etc. Ces mots *vides* sont sauvegardés dans le fichier *mot_vide.txt*, fourni dans le répertoire fichier.
Écrire une version 2 de votre projet, supprimant ces mots de vos mots-clés.

Question 3 :

Dans le résultat précédent, certaines formes d'un même mot apparaissent : la forme singulier ou pluriel, par exemple *Olympique* et *Olympiques*. Cependant, si on juge l'importance du mot dans le texte par son nombre d'apparitions, qu'il soit au singulier ou au pluriel, peu importe ! Il est bien présent ! Il est donc intéressant de mettre en place des règles permettant de relier les mots entre eux sur la base d'un radical commun. Ceci suppose bien entendu l'existence d'un lien sémantique fort entre des mots dont le radical est le même et qui diffèrent seulement dans leurs terminaisons. Il est toutefois important

de noter ici que le radical trouvé n'est pas nécessairement le radical au sens linguistique.
Lorsqu'on analyse un mot, pour en extraire son radical, on va chercher à supprimer les terminaisons superflues, par exemple *Olympiques*, après une première étape deviendra *Olympique*. Ce qui permettra de compter ensemble les occurrences de ces 2 mots. La démarche proposée ici se base sur l'algorithme suivant. Il y a 3 étapes pour obtenir le *pseudo*radical. Les règles proposées dans ce projet sont enregistrées dans 3 fichiers : *etape1.txt*, *etape2.txt* et *etape3.txt*.
Exemple :

```
0 ellement el
0 issement epsilon
0 alement al
0 eraient epsilon
```

ou bien

```
1 tion epsilon
1 el epsilon
0 i epsilon
```

Les règles se présentent comme suit :

- **Chaque ligne du fichier correspond à une règle.**
- **Il y a 3 colonnes.**
  1. **On ne s'occupera pas dans un premier temps de la première.**
  2. **La deuxième colonne contient la terminaison à analyser.**
  3. **La troisième colonne contient la modification à apporter au mot, si on rencontre la terminaison.**

Exemple
Si la règle est :

```
0 alement al
```

ça signifie que la terminaison *alement* doit être remplacée par *al*.
Après application de cette règle, *finalement* deviendra *final*.

Le mot *epsilon* correspond à la disparition de la terminaison.
Exemple
Ainsi, après application de la règle

```
1 tion epsilon
```

***finition* deviendra *fini*.**

Version a :

Ecrire une version 3a de votre projet, permettant d'obtenir l'ensemble des racines du texte ainsi que le nombre de fois où elles apparaissent.

Version b :

Pour certains mots (par exemple *tissaient*, qui deviendrait *t* après l'application de la règle), la racine obtenue n'est pas suffisamment explicite pour être interprétée. Il va donc falloir vérifier la validité de la racine.

Un peu de théorie :

Si nous posons que C est une consonne ou une suite de consonnes, que V est une voyelle ou une suite de voyelles, et que les crochets ([]) marquent un élément optionnel, alors chaque mot du français peut être réduit à cette formule : $[C](VC)^m[V]$, où (VC) est répété un nombre m de fois.

	Mot	[C]	(VC)	(VC)	VC)	[V]	m
	mange	m	ang			e	1
Exemple :	mignon	m	ign	on			2
	arbre		arbr			e	1
	alphabet		alph	ab	et		3

Ici, m est le nombre de fois où (VC) est répété.

Exemple : Analyse du mot *Magnifique*, en suivant la règle

1 **0 que c**

Son radical serait *Magnific*. Est-il valide ?
Je recherche la première voyelle. Tout ce qui apparaît avant cette voyelle, fait partie de [C], donc ici *M*.
Une fois cette voyelle trouvée (ici *a*), je cherche la prochaine consonne (ici *g*). Dès que je l'ai touvée, je cherche la prochaine voyelle et je remplis la première alternance voyelles/consonnes (VC) avec les lettres trouvées. Ici, j'obtiens : *agn*.
Dès que j'ai trouvé la prochaine voyelle (ici *i*), c'est que je démarre une nouvelle alternance ! Je cherche alors la prochaine consonne (ici *f*). Dès que je l'ai touvée, je cherche la prochaine voyelle et je remplis la deuxième alternance voyelles/consonnes (VC) avec les lettres trouvées. Ici, j'obtiens : *if*.
Et ainsi de suite, je recherche la prochaine voyelle, (ici *i*) puis la consonne, puis je remplis l'alternance, etc...
Dès que je ne trouve plus de voyelle ou bien de consonne, c'est que l'analyse est terminée.
Je trouve ici 3 alternances (VC). donc m=3.

Mot	[C]	(VC)	(VC)	VC)	[V]	m
Magnific	M	agn	if	ic		3

La première colonne des fichiers *etape* contient le critère de validité lié à m.

Ainsi, si la valeur de la première colonne du fichier *etape* contient 0, le radical trouvé pour un mot après application d'une règle doit avoir une valeur m *strictement* supérieure à 0.

Ici, le critère de validité est 0 **et pour le radical,** $m = 3$**. *m* est bien strictement supérieur au critère, donc le radical est validé.**
Exemple : application de la règle :

```
1 0 issaient epsilon
```

Le mot *remplissaient* deviendra *rempl*.

Mot	[C]	(VC)	[V]	m
rempl	r	empl		1

En effet, le *m* pour ce radical est 1. Et $1 > 0$ **! donc la contrainte est bien respectée.**
Par contre, le mot *tissaient* ne deviendra pas *t*. Le radical ne comprend pas au moins une séquence VC (il n'y a pas de voyelle du tout !) et donc $m = 0$**. Et** $0 > 0$ **est faux !**

***Attention* : Parfois le critère de validité est égal à 1 il faudra donc veiller à ce que la racine contienne plusieurs fois (VC) ! (un *m* strictement supérieur à 1 !)**

Ecrire une version 3b de votre projet, permettant, d'obtenir l'ensemble des racines du texte ainsi que le nombre de fois où elles apparaissent, en intégrant la vérification de la validité de la racine.

Version c : la dernière

Lorsque vous obtenez le résultat précédent, vous avez l'ensemble des racines du texte ainsi que le nombre de fois où elles apparaissent. Cependant, ces racines ne sont pas intéressantes pour la compréhension du texte...
Proposer une version 3c, permettant d'avoir l'ensemble des *mots* du texte ainsi que le nombre de fois où ils apparaissent, en tenant compte des racines et de la vérification précédente.

Question 4 :

Remarquez-vous des points faibles dans cet algorithme ? Par quel algorithme pouvez-vous les contourner ?

3 SAE 1.1 : Développer des applications informatiques simples (responsable : Aurélie Leborgne)

À partir de l'analyse réalisée dans la SAE 1.2 et du sujet que vous avez choisi, vous devrez créer une page web présentant les résultats obtenus en soignant le web design, l'accessibilité, *etc.* Cette page web doit contenir au minimum un titre et un nuage de mots. Notez que plus la page web est aboutie, plus le client sera satisfait ; alors osez et ne vous restreignez pas !

objectifs

- **Respecter le sujet et les spécifications techniques décrites ci-après par le client**
- **Choisir les ressources techniques appropriées**
- **Utiliser les principes de base vus en cours et des ressources externes que vous aurez sélectionnées pour gagner en autonomie**
- **Veiller à la qualité du code, à son organisation et à sa documentation**
- **Implémenter des conceptions simples pour avoir un résultat fonctionnel rapidement puis ajouter ou approfondir chaque fonctionnalité**
- **Faire des essais et évaluer leurs résultats au regard des spécifications**

Spécification techniques

- **Utiliser uniquement du HTML et du CSS**
- **Possibilité d'utiliser des cadres CSS comme Bootstrap, Bulma, *etc.* en fonction du rendu que vous visez.**
- **Ne pas utiliser de système de gestion de contenu (CMS) comme WordPress**
- **Automatiser la génération de la page web (utiliser le code C# pour écrire (ou compléter) votre code HTML et/ou CSS afin de ne pas avoir à écrire manuellement une ligne de code entre le lancement de l'analyse du/des document(s) et la visualisation de votre page web.**

Go ! ! ! ! ! Impressionnez-nous ! ! ! ! !

Notation

La notation de votre projet aura lieu lors d'une soutenance à la fin du semestre. Pour ce faire, en amont de la soutenance, vous devrez

créer une vidéo de 7 minutes (pas une seconde de plus) dans laquelle vous présenterez votre travail et montrerez que tous les objectifs ont été atteints. La vidéo sera à déposer sur moodle à l'endroit prévu à cet effet.
Pour enregistrer la vidéo, vous pouvez utiliser OBS. Voici un tuto pouvant vous aider : `https://www.youtube.com/watch?v=7vCivrV9u74&t=2s`
Pour faire le montage, vous pouvez utiliser shotcut. Voici un tuto simple pour commencer à prendre en main ce logiciel : `https://pod.unistra.fr/video/43256-creer-un-montage-video-simple-avec-shotcut/`
NB : Ces deux logiciels sont gratuits.

Lors de la soutenance, vous aurez en charge de montrer la vidéo à votre évaluateur et de vous préparer à répondre à quelques questions. Faites en sorte que la réalisation de votre projet soit facilement accessible.