|  |
| --- |
| **Algorithmique et fluctuation** |



**Situation :**

Si vous lisez la phrase « *A magyar valóban olyan nyelv, mint a többi!*», il y a peu de chance que vous la compreniez. En effet, celle-ci est probablement rédigée dans une langue étrangère que vous ne connaissez pas. La traduction de cette phrase est donc assez complexe car, non seulement vous ne connaissez pas la langue dans laquelle la phrase est écrite, mais vous ne savez même pas en quelle langue elle est rédigée. La question est donc la suivante :

**Est-il possible, dans un texte comprenant un grand nombre de mot dans une langue totalement inconnue, de deviner la langue dans laquelle il est écrit ?**

Et bien en théorie, c’est possible et il existe une méthode, même si elle ne marche pas à tous les coups. Elle repose sur le principe des fréquences d’apparitions des lettres, et peut être décrite en deux idées :

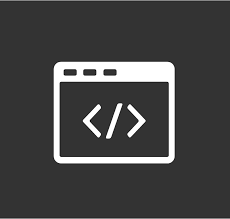
* **Idée 1 : certaines lettres n’existent que dans certaines langues**

Il existe pour chaque langue des lettres qui n’existe pas dans d’autres langues. Par exemple, le « ö » dans la phrase d’introduction n’existe pas en français, ni en espagnol ou en italien. Il existe cependant dans la langue allemande, hongroise ou encore turc. Le fait que la phrase mystère contienne un ö permet donc de réduire les possibilités lors de la recherche.

* **Idée 2 : toutes les langues n’utilisent pas les lettres de la même manière**

Chaque langue utilise plus ou moins certaines lettres. Par exemple, les textes en allemand contiennent environ 17,3 % de « e ». Cette fréquence tombe à 7,6 % pour le polonais. Ainsi, un texte contenant 16 % de « e » aura donc plus de chances d’être allemand que polonais.

A partir de ces deux idées, il est donc possible d’avoir une idée de la langue dans laquelle est rédigé un texte d’origine inconnue. C’est ce que nous allons faire dans la suite de cette activité, à l’aide de deux fonctions Python qui nous aideront dans notre travail : la fonction lettretotal et la fonction contelettre. Mais avant d’attaquer le travail, il faut d’abord comprendre leur fonctionnement.

[**Pour accéder à l’espace de code, clique ici**](https://trinket.io/python/9b4d57a519)**:**

**Partie 1 – Appropriation des fonctions lettretotal(fichier) et contelettre(fichier,lettre)**

1. Fonction lettretotal(fichier)

La fonction a simplement pour but de compter le nombre total de lettres dans un fichier texte. Elle prend en entrée le nom du fichier texte et retourne le nombre de lettre qu’il contient.

1. A l’aide de la fonction lettretotal, compter le nombre de lettre dans le fichier « test.txt ».
2. Compter le nombre de lettres dans les fichiers « vegedream.txt » et « misérables.txt »
3. Fonction contelettre(fichier,lettre)

La fonction compte le nombre d’une certaine lettre dans un fichier. Elle prend en entrée le nom du fichier texte et la lettre à compter et retourne le nombre de cette lettre dans le fichier.

1. A l’aide de la fonction contelettre, compter le nombre de « a » dans le fichier test.txt
2. Compter le nombre de « e » et de « u » dans les fichiers « vegedream.txt » et «baudelaire.txt »

**Partie 2 – Identification d’un texte en langue inconnue**

Vous disposez d’un fichier rédigé en langue inconnu. Nous allons tenter d’identifier sa provenance à l’aide des deux fonctions évoquées ci-dessus et du tableau suivant, contenant les fréquences d’apparitions de certaines lettres dans différents langages :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Fréquence d'apparition des lettres dans quelques langues** | | | | |
|  | **a** | **e** | **ö** | **h** |
| **Turc** | 11,90% | 8,91% | 0,77% | 1,21 |
| **Espagnol** | 12,52% | 13,60% | x | 0,7 |
| **Danois** | 6,02% | 15,45% | x | 1,62 |
| **Allemand** | 6,51% | 17,30% | 0,20% | 4,57 |
| **Français** | 7,60% | 14,70% | x | 0,73 |

1. Compter le nombre total de lettres du fichier mystere.txt
2. Compte le nombre total de « ö » dans le fichier. A l’aide du fichier Excel associé, identifier deux langues possibles pour ce fichier.
3. Calculer la fréquence d’apparition de la lettre « e ». En quel langue ce texte semble être rédigé ?
4. Trouver la langue dans laquelle est rédigée le texte mystere2.txt

**Partie 3 – Limites du modèle**

1. Quelle est la fréquence de « e » dans le texte fr.txt ?
2. Pourquoi notre méthode ne fonctionne pas bien pour identifier un texte court ?
3. Compter la fréquence de « e » dans le texte « disparition.txt ». Comment a été rédigé ce texte ?
4. Pourquoi notre méthode ne fonctionne pas bien pour identifier un texte de ce type ?
5. Rédiger un court paragraphe expliquant pour quel type de texte ces fonctions sont efficaces, et pour lesquels elles ne fonctionnent pas. Pensez à argumenter !

**Pour en savoir plus :**

[Wikipédia - La fréquence d’apparition des lettres dans les différents langages](https://fr.wikipedia.org/wiki/Fr%C3%A9quence_d%27apparition_des_lettres)

[La correction de l’exercice](https://codimd.apps.education.fr/s/WKis712g_)