

LYCEE CHARLEMAGNE

Février

M.P.S.I.2



2024

2025

I.T.C.

L'objectif de ce T.D. est (après un peu d'entraînement sur un fichier contenant un dictionnaire francophone) de réaliser un correcteur orthographique.

Objectif informatique :

- ouverture et lecture de fichiers issus du disque dur
- manipulations de listes
- recherche d'éléments réalisant un maximum
- manipulation de dictionnaires (au sens informatique du terme)
- distance de LEVENSHTAIN (théorie des graphes)
- indice de JACQUARD

Un fichier, à placer dans le répertoire actif contient une quantité importante de mots du dictionnaire français (suffisant pour tout ce qui va être fait ici).

Allez en chercher un sur Internet ¹.

Dans tout ce qui suit, il s'appelle **FRENCH.TXT** (il est donc de type « texte brut »). Si vous l'avez téléchargé sous un autre nom, agissez en conséquence.

Chargez le (pas juste dans le répertoire téléchargements, mais dans le même répertoire que votre fichier de travail).

Si vous avez du mal à le charger ou si elle pose des problèmes à cause des accents, vous pouvez vous en sortir en tapant en début de votre programme `#coding latin` ou `#coding latin` ou `#coding utf-8` ou un autre encodage adéquat, échangez entre vous pour voir ce qui passe avec le dico que vous aurez choisi.

On va l'ouvrir en mode lecture, et se livrer à quelques jeux OuLiPiens.

VOCALISES

ACCENTUATION, AUTOCUISEUR, SOUVERAIN, SAOUDITE, SAPEUR-POMPIER, APOCALYPTIQUE, AÉRODYNAMIQUE, YOUGOSLAVIE et TYPOGRAPHIQUE

Testez le programme suivant :

```
import os
Dico = open('FRENCH.TXT', 'r', encoding = '...') #choisissez l'encoding suivant le dico chargé
Scooby = []
for mot in Dico:
...if 'a' in Mot and 'e' in Mot and 'i' in Mot and 'o' in Mot and 'u' in Mot:
.....Scooby.append(Mot)
print(Scooby)
```

A quoi sert la ligne `import os` ? Et pourquoi pas `from os import *`, ou `import os as oss117` ?

.
 . Oui, c'est un vieux réflexe de programmeur de la première heure.

Sans cette ligne `import os`, que se passe-t-il ?

.
 .

1. « `github french-wordlist` » est pas mal

Quelle est l'erreur en ligne 4 de ce script ? L'aviez vous corrigée ou avez vous recopié sans réfléchir ?

Combien de mots trouvez vous ?

Et si vous voulez plusieurs lettres a ?

Et si vous voulez une seule fois chaque voyelle ?

Au fait, et le y là dedans ?

Pourquoi ne trouve-t-il pas le mot OISEAU quand vous essayez

```
Dico = open('FRENCH.TXT', 'r')
Scooby, ScoobyDoo = [], []
for Mot in Dico :
    ...if 'a' in Mot and 'e' in Mot and 'i' in Mot and 'o' in Mot and 'u' in Mot :
        .....Scooby.append(Mot)
print(Scooby)
for Mot in Dico :
    ...if 'a' in Mot and 'e' in Mot and 'i' in Mot and 'o' in Mot and 'u' in Mot and 's' in Mot :
        .....ScoobyDoo.append(Mot)
print(ScoobyDoo)
```

EXCES

INDIVISIBILITÉ, CORROBORATION, ANTICOMMUNISME, BAOBAB, ABRACADABRANT, TARTELETTE, CONCUSPISCENCE

```
import os
Fichier = open('FRENCH.TXT', 'r')
Scooby, Record = ' ', 0
for Mot in Fichier :
    ...Compt = Mot.count('i')
    ...if Compt > Record :
        .....Scooby = Mot
        .....Recrod = Compt
print('record ', Record, Scooby)
```

Quelles lettres ont été testées dans les exemples ?

Et si vous n'avez pas droit à la fonction count ?

Pourquoi avez vous recopié Recrod = Compt ?

Et si vous cherchez les e, mais aussi les é, les è, les ê ?

```

.
.
.
.

```

Mais si plusieurs mots réalisent le record, que faut il faire ?

```

.
.
.
.
.

```

Et comment trouver le mot qui réalise le record quand on tente toutes les lettres de l'alphabet (le record des records, c'est pour a, e, t ?)

```

.
.

```

MOT CACHE

RESTRUCTURATION AUTODESTRUCTRICE, SOUSCRIPTEUR, SUPERCHERIE, SOUSCRIPTEUR, SURENCHÉRIT

Le malin m'a inspiré le programme suivant :

```

Dico = open('FRENCH.TXT', 'r')
for mot in Dico :
    ...if Mystere(mot, 'satan') :
        .....print(mot)
print('Fini')
Dico.close()

```

```

def Mystere(Mot1, Mot2) :
    ....compt = 0
    ....for lettre in Mot1 :
        .....if lettre == Mot2[compt] :
            .....compt+=1
            .....if compt==len(Mot2) :
                .....return True
    ....return compt == len(Mot2)

```

A quoi sert ma ligne `print('Fini')` ?

```

.
.

```

A quoi sert ma ligne `Dico.close()` ?

```

.
.

```

Pouvez vous me dire si on verra à l'affichage les mots SOUS-TRAITANCE, SATURATION, SPECIALISATION, CONSTATATION, TRANSPARATION, DÉMOCRATISATION.

```

.
.

```

Écrivez les commentaires qui manquent dans le programme. Décrivez² l'évolution des variables `lettre`, `compt`, `lettre == Mot2[compt]` quand s'exécute la boucle `for lettre in Mot1` de la procédure `Mystere(mot, 'satan')` pour l'un des six mots ci-dessus (*de sous-traitance à démocratisation*).

```

.
.
.
.
.
.
.

```

```

if compt==len(Mot2) :
    ....return True

```

Expliquez le rôle des lignes

2. (par un tableau, évidemment, on est en sciences, pas en dissertation d'histoire-géo, et encore, même en histoire géo, c'est tellement plus clair avec des tableaux et de l'infographie)

.

.

.

.

1. J'ai remplacé SATAN par un prénom de la classe et trouvé plein de mots dont CATALYSEUR, TACHYCARDIE, ÉTAYERAI, TAYLORISEREZ... c'était qui ?
2. Et pour ARTICHAUTIÈRE, MARTIN-PÊCHEUR ?
3. Et pour GLOSOLALIE, PISTOLET-MITRAILLEUR, INSOLVABLE ?
4. Et pour FOULARD ? Et qui pour ANTICONSTITUTIONNEL, ANTICONCEPTUEL ?
5. Et KILOCYCLES, NITROGLYCERINE, RHINOPHARYNGITES, HIMALAYENS, DIALYSÉES ? Et FRÉQUENTATION, QUATRE-INGT-QUINZE ?
6. Et pour FORMIDABLE ? Et BARRISSEMENT ? Et DÉSAMIDONNERAI, SHAMPOUINERAS, PSALMODIERAS, ESTAMPILLERAI ?
7. Et pour NYMPHOMANIES, TYRANIQUES ?
8. Et pour OTO-RHINO-LARYNGOLOGIE, GONDOLANT ?
9. J'ai enlevé un des mots [embelliront, enorgueilliront, envieilliront, feuilletteront, helléniseront, intellectualiseront, javelliseront, parcellariseront, parcelliseront, recristalliseront, satelliseront, treillisseront, tressailliront, vieilliront], lequel ? Mais pourquoi pas SAPERLIPOPETTE ?
10. Qui a le record avec plus de 7000 mots : Malo, Jean, Léon, Louis, Côme, Lucas, Léo ?

.

.

Pour SOLÈNE³, un joli mot réagit positivement pour LOUISE⁴, et une floppée pour son autre fille ANAÏS (*si on ne met pas le tréma*).

Allez, pour LOUISE, c'est quoi ?

.

.

Et qui est le mot caché du titre ?

..

Les trois exercices suivants ne sont pas obligatoires.

Vous pouvez, si vous préférez, passer tout de suite au correcteur orthographique.

ULCERATIONS

PSYCHODRAME TRANSLUCIDE PROBLÉMATIQUES

Écrivez un programme qui cherche le mot le plus long du dictionnaire, mais dont toutes les lettres sont distinctes (ce n'est donc pas "ANTICONSTITUTIONNELLEMENT", ni même un quelconque "DIMÉTHYLPHÉNOL...ETC").

L'Oulipo de Queneau, LeLyonnais, Perec et les autres avait "ULCÉRATIONS". "PSYCHODRAME TRANSLUCIDE SYMPATHIQUE" ne m'ont pas l'air mal. Et si les lettres accentuées sont différentes entre elles, j'ai "CINÉMATÈQUES DÉMOGRAPHIQUES PROBLÉMATIQUES".

Écrivez un programme qui cherche le mot du dictionnaire qui utilise le plus de lettres différentes ("PORTEZ CE VIEUX WHISKY AU JUGE BLOND QUI FUME" est une phrase, pas un mot, mais "PSYCHOLOGIQUEMENT" est pas mal).

CRESCENDO DECRESCENDO

ABEILLE, BOUSSOLE, MORSURE, CEINTURON

Toujours notre fichier FRENCH.TXT. On doit trouver cette fois les mots comme ABEILLE, CEINTURON, IMPOLI, BOUSSOLE, MOUSSE et MORSURE.

3. insolence, somnolence, obsolescence, sous-alimenter

4. non, pas glandouillasse ni hospitalo-universitaire

Distance (et étapes) de SUCRI à CHRIST.

Justifiez que la distance de chien à niche est de 4.

Distance (et étapes) de CHIEN à NICHE.

L'algorithme est le suivant⁸ :

```
def Leven(MotA, MotB) : #distance de Leveshtein
...LA, LB = len(MotA), len(MotB)
...D = [[0 for j in range(LB+1)] for i in range(LA+1)]
...for i in range(LA+1) :
.....D[i][0]=i
...for j in range(LB+1) :
.....D[0][j]=j
...for i in range(1, LA+1) :
.....for j in range(1, LB+1) :
.....if MotA[i-1] == MotB[j-1] :
.....cout = 0
.....else :
.....cout = 1
.....D[i][j]=min(D[i-1][j]+1, D[i][j-1]+1, D[i-1][j-1]+cout)
.....#print('Etape ', i)
.....#for k in range(LA+1) :
.....#...print(D[k])
...return D[LA][LB]
```

On ne demande pas de justifier que ce script calcule effectivement la distance (*ce serait du domaine de l'option informatique, du T.I.P.E. et pas de l'I.T.C.*).

Pouvez vous compacter en une ligne le bloc

```
if MotA[i-1] == MotB[j-1] :
...cout = 0
else :
...cout = 1
```

Compactage du bloc en une ligne.

8. on ne justifiera pas qu'il calcule effectivement la distance, il faudrait un prochain TD

Qu'est ce qu'un {dictionnaire} Python ? Un objet un peu plus complexe que les 'chaînes de caractères' et des [listes].

Un dictionnaire est proche d'une liste (ses éléments sont modifiables), mais les éléments ne sont pas indexés en séquence (l'appel de `L[i]` avec `i` dans `range(len(L))`) ; les éléments enregistrés ne sont pas dans un ordre immuable. L'accès aux éléments se fait par une clef (en anglais KEY), qui sera ici alphabétique.

Par exemple, un dictionnaire nommé MPSI pourra être constitué des noms des enseignants de la classe avec pour clef la matière enseignée :

```
MPSI = {'Maths': ['Choquet', 'Wattiez'], 'Physique': ['Chevalier', 'Chabane'] : 'SII',
['Debernardi'] : 'Philo' : ['Seguret']}
```

On veut savoir qui enseigne la philo : `print(MPSI2['Physique'])` donne `['Chevalier', 'Chabane']`.

On a oublié l'anglais ? Facile : `MPSI['Anglais']=['Jeanbourquin']` (*pas besoin de préciser qu'on allonge par append*).

Et voilà le nouveau MPSI : `{'Maths': ['Choquet', 'Wattiez'], 'Physique': ['Chevalier', 'Chabane'], 'SII': ['Debernardi'], 'Anglais': ['Jeanbourquin'], 'Philo': ['Seguret']}`.

On a oublié que les MPSI1 avaient un autre professeur : `MPSI['Maths'].append('La peluche verte')`.

On veut voir les clefs : `print(MPSI.keys())` donne `['Maths', 'Anglais', 'Physique', 'SII', 'Philo']` (*l'ordre étant imprévisible*).

1. Que donne `MPSI['Maths'].remove('Choquet')` ?
2. Quelle instruction devez vous taper pour ajouter l'informatique et l'I.T.C..
3. Quelle instruction devez vous taper pour que madame CHEVALIER s'appelle CHEVALIER-THERY ?
4. Quel est l'effet de `for A in MPSI : MPSI[A].append('Sucr') ?`
5. La PCSI a aussi un dictionnaire du même type. Tapez une instruction qui affecte l'I.T.C. à la réunion des professeurs de maths, de physique, de chimie et de O.Brunet, si toutefois il n'y a pas déjà un prof d'I.T.C. dans le dictionnaire.

Dictionnaire MPSI.

Un petit test sur les dictionnaires ? Que va donner ce script ?

```
texte = 'longtemps je me suis couché de bonne heure'
table = {}
for lettre in texte :
...if lettre in table.keys() :
.....table[lettre] += 1
...else :
.....table[lettre] = 0
print(table)
```

Citation de Proust.

I.T.C.

Un script du serpent satanique.



On décrypte le script :

<code>import os</code>	on importe le module
<code>Dico=open('FRENCH.TXT','r')</code>	on ouvre un document et on l'affecte à une variable
<code>for mot in Dico :</code>	on lit ligne à ligne le contenu de ce fichier ¹¹
<code>...if Mystere(mot,'satan') :</code>	on fait un test en sollicitant une procédure
<code>.....print(mot)</code>	si le test est valide on affiche le mot
<code>print('Fini')</code>	on prévient que c'est fini
<code>Dico.close()</code>	on fait les choses proprement, bon sang !

On étudie alors la procédure, qui prend en entrée deux variables. Ce sera ici le mot lu dans le fichier texte, et un second mot, 'satan' dans notre exemple. On initialise un compteur à 0, qui ne va pas se contenter de compter des égalités, mais qui va aussi faire avancer caractère à caractère dans le mot Mot2.

<code>def Mystere(Mot1,Mot2) :</code>	On prend donc deux chaînes ou listes en variables.
<code>...compt = 0</code>	On initialise un compteur.
<code>.....for lettre in Mot1 :</code>	On parcourt une à une les lettres du mot Mot1.
<code>.....if lettre == Mot2[compt] :</code>	On regarde si il y a occurrence avec une certaine lettre du mot Mot2.
<code>.....compt+=1</code>	Si il y a occurrence, on avance le compteur. On passera donc à la comparaison à la lettre suivante de Mot2.
<code>.....if compt==len(Mot2) :</code>	On teste si on est allé jusqu'au bout des lettres de Mot2.
<code>.....return True</code>	Si oui, on répond True, sinon, on continue à comparer lettre à lettre.
<code>...return compt == len(Mot2)</code>	On renvoie un test. C'est un booléen, qui est donc évalué et vaut True ou False

Je sais, pour la fin, vous auriez tapé

```
if compt == len(Mot2) :
...return True
else :
...return False
```

ce qui est correct, mais aussi redondant que si vous écriviez `if (a compt==len(Mot2))==True`.

celà dit, je vous connais, et vous m'écrivez effectivement des `if Test(a)==True` : quand `Test(a)` est lui même un booléen. Mais c'est normal, vous avez juste un bac S, pas un bac scientifique.

Que fait ce script ? Exactement ce que vous feriez à la main avec par exemple les mots `constatation` et `satan`, qui sont deux chaînes de caractères.

Vous prenez le `s` de `satan` c'est `Mot2[compt]` avec `compt=0`.

Vous prenez une à une les lettres de `constatation` : '`c`', puis '`o`', puis '`n`' (*sans malice, ça fait con*).

Pour l'instant, aucune n'est égale à '`s`', vous traversez le test sans rien faire.

Vous arrivez à '`s`' de `conStatation` ; là, il y a égalité entre lettre et `Mot2[0]` ; le teste est valide, `compt` passe à 1. Les tests se feront alors entre `lettre` et `Mot2[1]`, c'est à dire '`a`'.

La lettre suivante ('`t`') ne donne pas un test valide, on continue à avancer.

La lettre suivante ('`a`') donne un test valide, on passe `compt` à 2. Il n'a pas encore atteint la valeur critique `len(Mot2)`, on continue donc.

lettre	c	o	n	s	t	a	t	a	t	i	o	n
Mot2[compt]	s	s	s	s	a	a	t	a	n	n	n	n
test	F	F	F	T	F	T	T	T	F	F	F	F
compt	0	0	0	1	1	2	3	4	4	4	4	5

A la fin, `compt` vaut 5, on a trouvé, dans l'ordre toutes les lettres du mot `satan`, le programme répond `True`.
Avec DÉMOCRATISATION ça finit moins bien.

lettre	d	e	m	o	c	r	a	t	i	s	a	t	i	o	n
Mot2[compt]	s	s	s	s	s	s	s	s	s	s	a	t	a	a	a
test	F	F	F	F	F	F	F	F	F	T	T	T	F	F	F
compt	0	0	0	0	0	0	0	0	0	1	2	3	3	3	3

Bilan : on regarde si on retrouve une à une les lettres de `Mot2` cachées dans `Mot1`, dans le bon ordre. C'est une sorte d'inclusion d'un mot dans l'autre.

On retrouve alors les prénoms cachés :

Prochain challenge : on donne les mots de la colonne de gauche à un programme et il cherche à retrouver le mot de la colonne de droite (*lettres communes à nos mots, mais en plus dans le même ordre, ça doit pas être mal à programmer...*).

Le test `if compt==len(Mot2):`
`...return True` est là pour quand, comme dans `ELISA` et `ReLUISANT`, on a égalité de la der-

nière lettre de `Mot2` (ici, le `a`) avec une lettre du milieu de `Mot2`. Si on ne sort pas à cet instant en rompant la boucle `compt` a atteint la valeur `len(Mot2)` (ici 5) alors on fait le test `lettre==Mot2[5]` alors que `Mot2[5]` n'existe pas¹².

Le cadeau pour Solène : **LOUISE** et **EBLOUISSEMENT**, mais pour **ANAIS**, j'ai cannabis, démangeaison, clef-anglaise, française, vandalisme, vantardisme, organisations, anachronisme, et même azerbaïdjanais si on veut un `i` (ah non, pas au bon endroit).

I.T.C.

Le mot le plus long.



On va ouvrir le fichier en mode lecture simple et donner un nom au fichier ouvert. Ensuite, on le parcourera ligne à ligne.

```
import os
Dico = open('FRENCK.TXT', 'r') #simple mode lecture
for Mot in Dico: #comme ça on les aura tous
...oui tiens, on fait quoi?
```

On va devoir tester si toutes les lettres d'un mot sont différentes.

Pour cela, on crée une procédure de test :

```
def PasDeLettreEnDouble(Mot):
...for k in range(len(Mot)): #on regarde les lettres une à une
.....for i in range(k): #on regarde les lettres qui précèdent
.....if Mot[k] == Mot[i]: #si il y a une égalité de lettres
.....return False #on sort et on dit "pas bon"
...return True #il n'y a eu aucun doublon, c'est bonnard !
```

On crée ensuite une variable `Record`, qui va mémoriser la longueur du mot le plus long trouvé jusqu'à présent :

¹². on rappelle que le dernier élément d'une liste de longueur `n` est `L[n-1]`

```
Record = 0
for Mot in Dico :
...if PasDeLettreEnDouble(mot) : #on ne regarde que les mots sans doublons
.....if len(Mot) > Record : #chaque fois qu'on a trouvé mieux
.....Record = len(Mot) #on mémorise la longueur
.....MotRecord = Mot #on mémorise le mot
Dico.close() #par mesure de sécurité
```

On peut aussi se dire qu'on ne testera que les mots longs :

```
Record = 0
for Mot in Dico :
...if len(Mot) > Record : #déjà, il faut un nouveau record
.....if PasDeLettreEnDouble(mot) : #chaque fois qu'on a trouvé mieux
.....Record = len(Mot) #on mémorise la longueur
.....MotRecord = Mot #on mémorise le mot
Dico.close() #par mesure de sécurité
```

Sinon, il faut aussi une procédure qui compte combien il y a de lettres différentes dans un mot.

On fait simple. Chaque fois qu'on nous donne un mot, on le recopie lettre à lettre ; mais quand une lettre a déjà été utilisée, on ne la prend pas :

```
def NbLettres(Mot) :
...NouveauMot = ""
...for lettre in Mot :
.....if not(lettre in NouveauMot) :
.....NouveauMot += lettre
...return len(NouveauMot)
```

Par exemple, avec `Mot = 'PSYCHOLOGIQUEMENT'`, on arrive à `NouveauMot = 'PSYCHOLOGIQUERMNT'`, tandis qu'avec `Mot = 'ANTICONSTITUTIONNELLEMENT'`, on a `NouveauMot = 'ANTICOSUELM'`, ce qui est décevant. Ensuite, on crée un parcours avec record qu'on pousse au fur et à mesure :

```
Dico = open('FRENCK.TXT', 'r')
record = 0
for Mot in Dico :
...if NbLettres(mot) > record :
.....record = NbLettres(mot)
.....MotRecord = Mot
print(MotRecord)
```

I.T.C.

Crescendo decrescendo.



On crée une procédure qui prend un mot et regarde si ses lettres sont classées par ordre croissant, décroissant ou change de sens de variation.

On mesure donc le signe de $\text{Mot}[k+1] - \text{Mot}[k]$ et de $\text{Mot}[k+2] - \text{Mot}[k+1]$.

Si les deux sont de même signe, la suite est monotone.

Si les deux sont de signe opposé, la suite a un extremum.

	$\text{Mot}[k] < \text{Mot}[k+1]$	$\text{Mot}[k] > \text{Mot}[k+1]$
$\text{Mot}[k+1] < \text{Mot}[k+2]$	croissante	minimum local
$\text{Mot}[k+1] > \text{Mot}[k+2]$	maximum local	décroissante

Comme $\text{Mot}[k] < \text{Mot}[k+1]$ et $\text{Mot}[k+1] < \text{Mot}[k+2]$ sont des booléens, ils valent 0 ou 1, le tableau se résume en

somme des booléens	True	False
True	2	1
False	1	0

et on constate que le changement de sens de variation se lit par

$$(\text{Mot}[k] < \text{Mot}[k+1]) + (\text{Mot}[k+1] < \text{Mot}[k+2]) == 1$$

(on pouvait aussi jouer avec un ou exclusif $(\text{Mot}[k] < \text{Mot}[k+1]) \wedge (\text{Mot}[k+1] < \text{Mot}[k+2])$).

On doit donc savoir si il y a trop de changements de sens de variation :

```
def MotCroissantDecroissant(Mot) :
...#if len(Mot)<2 : #à quoi bon tester les mots trop courts pour varier
...#...return False
...change=0 #on initialise
...for k in range(len(Mot)-2) : #on va aller regarder Mot(k+2) on s'arrête avant la fin
.....if (Mot[k] < Mot[k+1]) + (Mot[k+1] < Mot[k+2]) == 1 :
.....change += 1 #on a détecté un extrémum local
...return change < 2 #il n'y a eu que 0 ou 1 changement de variation
```

Je vous connais, vous auriez conclu avec

```
...if change < 2 :
.....return True
...else :
.....return False
```

Mais quel gâchis ! change < 2 est déjà un booléen qui vaut True ou False. C'est comme vos test en if (a==b) == True. C'est mignon... et ridicule.

On n'a plus qu'à utiliser ce petit programme :

```
Dico=open('French.txt','r')
L = [ ]
for Mot in Dico :
...if MotCroissantDecroissant(Mot) :
.....L.append(Mot)
print(L)
Dico.close()
```



```

Dico = open('FRENCH.TXT', 'r')
DicoMots = [ ] #pour les mots proprement écrits
DicoSorted = [ ] #pour les mots triés

for mot in Dico: #on lit le dictionnaire
...DicoMots.append(mot) #on remplit le dictionnaire des mots lisibles
...MotListe = sorted(list(Mot)) #on crée l'anagramme trié
...DicoSorted.append(MotListe) #on le mémorise
Dico.close( ) #on est propre, on ferme

for k in range(len(DicoSorted)): #on va lire le dictionnaire trié
...Mot1 = DicoSorted[k] #on prend le mot pour ne pas le lire et le relire
...for i in range(k): #on ne regarde que les mots qui précèdent
.....Mot2 = DicoSorted[i] #on lit le mot
.....if len(Mot1) == len(Mot2): #on ne regarde que si ils ont la même longueur
.....if Mot1 == Mot2: #on compare les mots triés
.....print(DicoMots[i], DicoMots[k]) #on affiche les mots sous forme lisible

print('Et voila, fini !') #on prévient qu'on a fait le tour

```

I.T.C.

Correcteur orthographique.



Transformation de FRENCH.TXT en Dico.

```

import os.
fichier = open('FRENCH.TXT', 'r')
Dico = [ ]
for mot in fichier:
...Dico.append(mot[:-2])
fichier.close()

```

On ouvrira même avec **import os, time** car on aura besoin de **time** plus loin.

On n'oublie pas d'affecter l'ouverture à une variable.

De plus, dans l'ouverture, le nom de fichier doit être mis entre guillemets.

Il est d'autre part absurde de penser travailler sur des **FRENCH.TXT.readlines()** car **FRENCH.TXT** n'est pas un objet Python.

La variable **mot[:-2]** prend **mot** et ne le recopie que jusqu'à deux caractères de la fin, car dans le fichier ouvert, chaque mot contient deux caractères de "retour chariot" à la fin de chaque ligne.

On peut aussi lire le fichier entier **Dico = fichier.readlines()** et découper par la méthode **D.split()**, le caractère de découpage étant justement le **'n'** de retour chariot.

Distance (et étapes) de SOLENE à NICOLAS.

On se contente de reprendre en sens inverse le chemin de NICOLAS à SOLENE.

On disposait de

nicolas	$R_{0,0}$	sicolas	E_1	scolas	E_1	solas	$R_{3,3}$	soles	$R_{4,4}$	solen	$I_{5,5}$	solene
---------	-----------	---------	-------	--------	-------	-------	-----------	-------	-----------	-------	-----------	--------

On peut créer

solene	E_5	solen	$R_{4,6}$	soles	$R_{3,5}$	solas	$I_{1,2}$	scolas	$I_{1,1}$	sicolas	$R_{0,0}$	nicolas
--------	-------	-------	-----------	-------	-----------	-------	-----------	--------	-----------	---------	-----------	---------

Il faudrait quand même prouver qu'il n'y a pas de chemin plus court. Il faudrait argumenter sur le nombre de lettres différentes déjà.

Il n'y a pas unicité du chemin de longueur 6.

SOLENE	R	NOLENE	E	NOLEE	R	NOLAE	R	NOLAS	I	NIOLAS	I	NICOLAS
SOLENE	R	SOLENS	R	SOLEAS	E	SOLAS	I	SIOLAS	I	SICOLAS	R	NICOLAS

Ensuite, je peux me demander si des E_i commutent avec des $R_{j,k}$.

Distance (et étapes) de SUCRI à CHRIST.

On peut faire ça

SUCRI	E_0	UCRI	E_0	CRI	$I_{1,1}$	CHRI	$I_{4,4}$	CHRIS	$I_{5,5}$	CHRIST
-------	-------	------	-------	-----	-----------	------	-----------	-------	-----------	--------

Et j'avoue bien humblement ne pas pouvoir faire plus court.

Il faut au moins effacer ou remplacer S et U. Et ajouter H, S et T.

On me dira qu'il est idiot d'effacer un S pour en remettre un. Mais il n'est pas au bon endroit. De plus, on ne peut pas gagner du temps en remplaçant le U par un T par exemple ; les emplacements ne sont pas les bons. Ce qui est utile, c'est un $I_{k,k}$.

Distance (et étapes) de CHIEN à NICHE.

Là aussi, les solutions sont multiples :

CHIEN	E_2	CHEN	E_3	CHE	$I_{0,1}$	ICHE	$I_{0,0}$	NICHE
CHIEN	I	NCHIEN	I	NICHIE	E	NICHEN	E	NICHE

On ne pourra pas faire plus court, car si on utilise un E, il faudra utiliser une I, et on ne peut pas n'utiliser que des R (sauf à en utiliser 5 car aucune lettre n'est au bon endroit attendu).

Je sais, ce n'était pas évident, il fallait trouver une séquence ordonnée présente dans les deux : le CHiEn et niCHE.

Compactage du bloc en une ligne.

La variable créée est `cout`, elle vaut 0 ou 1, suivant que `MotA[i-1]` est égal à `MotB[j-1]` ou pas.

On peut profiter du fait qu'un booléen vaut 0 pour `False` et 1 pour `True`.

On proposera `cout = (MotA[i-1] != MotB[j-1])`.

Ou `cout = 1-(MotA[i-1] == MotB[j-1])` (ici, le fait de faire un calcul transforme vraiment le booléen en entier).

Exécution de Leven('SUCRI', 'CHRIST').

Le bloc

```
print('Etape ', i)
for k in range(LA+1) :
...print(D[k])
```

 affiche le texte Etape et l'indice i (qui varie de 1 à LA

(inclus) puis affiche ligne par ligne la matrice D

La matrice D a été initialisée par

`D = [[0 for j in range(LB+1)] for i in range(LA+1)]`

elle a LA+1 lignes, formées de LB+1 termes.

LA est la longueur du premier mot passé (ici 'Sucri' LA=5), et LB est la longueur du second mot passé (ici 'Christ' LB=6)

Les deux +1 ne sont pas des erreurs, ni des compensations du principe informatique des `range` ; la matrice est bien plus grande que les mots traités.

Ici, D a 6 lignes de longueur 7.

La matrice initiale est truffée de 0.

Mais ensuite, on modifie quelques termes

`...for i in range(LA+1) :`

`.....D[i][0]=i`

les termes d'indices 0 de chaque ligne valent i (indice de ligne)

`...for j in range(LB+1) :`

`.....D[0][j]=j`

les termes de la ligne d'indice 0 valent j (indice de colonne).

La première matrice est

Ensuite, on modifie ligne par ligne la matrice, en la parcourant, avec des tests plus ou moins clairs. J'applique :

0	1	2	3	4	5	6	0	1	2	3	4	5	6	0	1	2	3	4	5	6
1	1	2	3	4	4	5	1	1	2	3	4	4	5	1	1	2	3	4	4	5
2	0	0	0	0	0	0	2	2	2	3	4	5	5	2	2	2	3	4	5	5
3	0	0	0	0	0	0	3	0	0	0	0	0	0	3	2	3	3	4	5	6
4	0	0	0	0	0	0	4	0	0	0	0	0	0	4	0	0	0	0	0	0
5	0	0	0	0	0	0	5	0	0	0	0	0	0	5	0	0	0	0	0	0
0	1	2	3	4	5	6	0	1	2	3	4	5	6							
1	1	2	3	4	4	5	1	1	2	3	4	4	5							
2	2	2	3	4	5	5	2	2	2	3	4	5	5							
3	2	3	3	4	5	6	3	2	3	3	4	5	6							
4	3	3	3	4	5	6	4	3	3	3	4	5	6							
5	0	0	0	0	0	0	5	4	4	4	3	4	5							

Je vous laisse comprendre sur cet exemple quel mot est formé peu à peu pour passer de `sucri` à `christ`.

Dans l'instruction `D[i][j]=min(D[i-1][j]+1, D[i][j-1]+1, D[i-1][j-1]+cout)`

on reconnaît peut être les trois types d'étapes de la méthode de Levenshtein :

effacement, insertion, remplacement

Mais il faut vraiment se prendre la tête.

Recherche des mots proches.

```
def MostProches(Mot) :
...Tdebut = time.clock()
...if Mot in Dico :
.....return(True, 0, time.clock()-Tdebut)
...Mini = Leven(Mot, Dico[0])
...MotsProches = [Dico[0]]
...for BonMot in Dico :
.....Dist = Leven(Mot, BonMot)
.....if Dist < Mini :
.....Mini = Dist
.....MotsProches = []
.....id Dist == Mini :
.....MotsProches.append(BonMot)
...return False, Mini, MotsProches, time.clock()-Tdebut
```

Avec `Tdebut`, on mémorise l'heure de début, et avec `time.clock()-Tdebut`, on mesure donc le temps écoulé entre l'entrée dans la procédure, et la sortie.

Les comparaisons entre `Mini` et `Dist` sont la classique recherche d'un *minimum* (on a initialisé avec le premier mot du dictionnaire).

La liste `MotsProches` est un tampon que l'on remet à zéro quand la distance diminue (*BonMot* est un record), et qu'on agrandit quand la distance stagne (*BonMot* fait aussi bien que le record précédent), et à laquelle on ne fait rien quand *BonMot* n'est vraiment pas mieux que le record.

Procédure Trigrammes.

Un mot de n lettres a n trigrammes.

En effet, pour chaque lettre du mot jusqu'à l'avant dernière, il y a un trigramme (la lettre et les deux lettres qui suivent, et aussi pour l'avant dernière lettre, il y a deux lettres suivies du symbole \$). Et il y a aussi un trigramme initial formé de \$ suivi des deux premières lettres du mot.

Pour n lettres, il y a $n-1$ trigrammes purs en `def`, un trigramme d'ouverture `$ab` et un trigramme de clôture `yz$`.

```
def Trigrammes(Mot) :
...mot = '$'+Mot+'$'
...Trig = [ ]
...for k in range(len(mot)-2) :
.....Trig.append(mot[k:k+3])
...return Trig
```

On travaille sur `mot` (sans majuscule) pour ne pas abimer `Mot` (avec majuscule), et en en profite pour coller les deux \$ typographiques.

La liste des trigrammes est initialisée à une liste vide qu'on agrandira.

Le range va de 0 à `len(mot)-2`, pour que l'extraction `mot[k:k+3]` s'arrête à la fin de `mot`.

On peut construire la boucle implicitement :

```
...Trig = [mot[k:k+3] for k in range(len(mot)-2)]
```

On peut aussi lire les lettres une à une :

```
...mot = '$'+Mot+'$'
...L1, L2 = Mot[0], Mot[1]
...Trig = [ ]
...for k in range(2 : len(mot)) :
.....L1, L2, L3 = L2, L3, mot[k]
.....Trig.append(L1+L2+L3)
```

Dictionnaire MPSI2.	
1-	<p>Que donne <code>MPSI['Maths'].remove('Choquet')</code> ? On va chercher la liste <code>MPSI['Maths']</code>, qui existe, et de cette liste, on efface 'Choquet'. Dans cette instruction, on n'appelle pas d'élément qui n'existe pas. Attention, si on l'exécute deux fois, on a un message d'erreur, puisque la liste <code>MPSI['Maths']</code> ne contient plus 'Choquet'. (déjà parti à la retraite ?)</p>
2-	<p>Quelle instruction devez vous taper pour ajouter l'informatique et l'I.P.T. <code>MPSI['Informatique']=['Brunet']</code> <code>MPSI['ITC']=['Brunet', 'Choquet', 'Chabane']</code> Il faut deux instruction.</p>
3-	<p>Quelle instruction devez vous taper pour que madame CHEVALIER s'appelle CHEVALIER-THERY ? On peut certes remplacer la liste : <code>MPSI['Physique'] = ['Chevalier-Thery']</code> Mais comme on peut la trouver sur d'autres matières : <pre>for matiere in MPSI.keys():Profs = MPSI2[matiere]:for k in range(len(Profs)):if Profs[k] == 'Chevalier':Profs[k] = 'Chevalier-Thery'</pre></p>
4-	<p>Quel est l'effet de <code>for A in MPSI : MPSI[A].append('Sucri')</code> ? A chaque liste de MPSI, on ajoute Sucri. Il devient prof omnipotent...</p>
5-	<p>La <code>MPSI</code> n'a aussi un dictionnaire du même type. Tapez une instruction qui affecte l'I.P.T. à la réunion des professeurs de maths et de physique, si toutefois il n'y a pas déjà un prof d'I.P.T. dans le dictionnaire. <pre>if not('ITC' in PCSI.keys()):PCSI['ITC']=['Brunet']+PCSI['Maths']+PCSI['Physique']+PCSI['Chimie']</pre> Et peut être même : <pre>Matieres = PCSI.keys() if not('IPT' in Matieres) and 'Maths' in Matieres and 'Physique' in Matieres and 'Chimie' in Matieres:PCSI['IPT']=['Brunet']+PCSI['Maths']+PCSI['Physique']+PCSI['Chimie']</pre> afin de s'assurer qu'il y a bien des Maths et de la Physique dans cette PCSI... Tiens, mais si c'est le même prof en maths, en physique et en chimie ? Non, ne parlez pas d'horreur.</p>

Citation de Proust.

La variable `texte` contient une chaîne de caractères qu'on va explorer lettre par lettre (*y compris les espaces et la ponctuation*).

On initialise un dictionnaire au sens pythonien du terme, qui s'appelle `table` (*on reconnaît au fait qu'on initialise avec des accolades*).

Pour chaque lettre, on regarde si elle est déjà dans les clefs de la `table`.

Si ce n'est pas le cas, on la crée comme clef, avec comme élément associé l'entier 1.

Peu à peu, `table` contient des couples (`lettre`, 1).

On a donc `{'l' : 1}`, puis `{'l' : 1, 'o' : 1}`, puis `{'l' : 1, 'o' : 1, 'n' : 1}` et ainsi de suite, jusqu'à `{'l' : 1, 'o' : 1, 'n' : 1, 'g' : 1, 't' : 1, 'e' : 1, 'm' : 1, 'p' : 1, 's' : 1, ' ' : 1, 'j' : 1}`

La lettre suivante est un 'e' encore. Il est déjà dans `table.keys()`. On sollicite alors `table['e']` et on l'augmente de 1 :

`{'l' : 1, 'o' : 1, 'n' : 1, 'g' : 1, 't' : 1, 'm' : 1, 'p' : 1, 's' : 1, ' ' : 1, 'j' : 1, 'e' : 2}`

Pour chaque symbole nouveau, on met un compteur à 1.

Pour chaque symbole déjà croisé, on augmente son compteur de une unité.

On crée donc un histogramme des lettres de la phrase.

`{' ' : 7, 'c' : 2, 'b' : 1, 'e' : 7, 'é' : 1, 'd' : 1, 'g' : 1, 'i' : 1, 'h' : 2, 'j' : 1, 'm' : 2, 'l' : 1, 'o' : 3, 'n' : 3, 'p' : 1, 's' : 3, 'r' : 1, 'u' : 3, 't' : 1}`

Il ne faut pas avoir oublié ' ' : 7.

Création du dictionnaire BONSTRIGRAMMES.

```
BonsTrigrammes = { }
for Mot in Dico :
    ....Trigs = Trigrammes(Mot)
    ....for trig in Trigs :
    .....if trig in BonsTrigrammes.keys() :
    .....BonsTrigrammes[trig].append(mot)
    .....else :
    .....BonsTrigrammes[trig] = [mot]
```

Les programmeurs pourront même écrire

```
for mot in Dico :
    ....for trig in Trigrammes(mot) :
    .....BonsTrigrammes[trig]=BonsTrigrammes.get(trig, [])+[mot]
```

En effet, la méthode `get` appliquée aux dictionnaires est quelque chose de bien pratique :

`dict.get(a, sinon)` cherche si il y a dans `dict` une clef appelée `a`

- si il y en a une, on prend `dict[a]`
- si il n'y en a pas, on crée la valeur `sinon`.

Ici, donc, avec `BonsTrigrammes[trig]=BonsTrigrammes.get(trig, [])+[mot]`

- si `trig` est une clef de `BonsTrigrammes`, alors `BonsTrigrammes.get(trig, [])` est `BonsTrigrammes[trig]`, et on ajoute `[mot]` (*par addition, la liste s'agrandit*)

- si `trig` n'est pas une clef de `BonsTrigrammes`, alors `BonsTrigrammes.get(trig, [])` est égal à `[]` (*liste vide*). On lui ajoute `mot`, on obtient `[mot]` et on l'affecte à `BonsTrigrammes[trig]`, qui n'existait pas encore.

Génial, n'est il pas ?

Indice de Jacquard.

Que signifie $Jac(A, B) = 0$? A et B ont une intersection vide. Ils sont disjoints.

Que signifie $Jac(A, B) = 1$? Le cardinal de $A \cap B$ est égal à celui de $A \cup B$. Comme l'un est inclus dans l'autre, c'est qu'ils sont égaux. Comme on a $A \cap B \subset A \subset A \cup B$, on arrive à $A = A \cup B$, puis $B = A \cup B$. Comme on s'en doutait : A est égal à B .

Que signifie $Jac(A, B) = 2$? Que le cardinal de $A \cap B$ est plus grand que celui de $A \cup B$. Ce n'est pas possible.

A moins que l'on ne prenne l'ensemble vide, mais $\frac{Card(\emptyset \cap \emptyset)}{Card(\emptyset \cup \emptyset)}$ n'a pas de sens...

Mots d'indice plus grand que 0.2.

```

#liste des trigrammes de Mot
Trig=Trigrammes(Mot)
#creation d'un dictionnaire des mots ayant au moins un trigramme commun avec Mot
#on prend les trigrammes de Mot, et on regarde à quels bons mots du dictionnaire ils
renvoient
Commun={ }
for trig in Trig:
...for MotCorrect in BonsTrigrammes[trig]:
.....if MotCorrect in Commun.keys():
.....Commun[MotCorrect] += 1
.....else:
.....Commun[MotCorrect] = 1
#maintenant, on utilise Commun, et on regarde quels mots ont un bon taux
MotsProches = [ ]
for MotCorrect in Commun:
...NbTrig = float(Commun[MotCorrect])
...taux = NbTrig/(len(MotCorrect)+len(MotA)-NbTrig)
...if taux > 0.2:
.....MotsProches.append(MotCorrect)

```