Distance entre deux mots

Pour savoir si un mot est mal orthographié, il suffit de vérifier s'il existe dans un dictionnaire, par contre pour suggérer une correction orthographique, il faut proposer un mot proche. Pour cela, il faut définir une distance entre deux mots.

Activité 1 (Distance de Hamming)**.**
Prends deux mots de même longueur. La *distance de Hamming*, c'est le nombre d'endroits où les lettres sont différentes.
Par exemple :

JAPON **SAVON**

Les premières lettres sont différentes, les troisièmes aussi. La distance de Hamming entre ces deux mots vaut donc 2.

1. Calcule la distance de Hamming entre les mots suivants :

LIGNE	**LIANE**
BOOLE	**MOORE**
POLICE	**PILOTE**
PASSION	**RATIONS**
CRANE	**ECRAN**

2. Pour chacun des mots de la liste suivante, calcule sa distance de Hamming avec le mot **SIGNE** :

SUITE **LIGNE** **SINGE** **DIGNE** **MIXTE**

Activité 2 (Distance de Jaccard)**.**
La *distance de Jaccard* mesure la proximité de deux mots, indépendamment de l'ordre des lettres.
La formule est :

$$\text{distance} = 1 - \frac{\text{nombre de lettres communes}}{\text{nombre total de lettres}}$$

La distance est comprise entre 0 et 1. Plus la distance est proche de 0, plus les mots ont les mêmes lettres.

Exemple 1. PAIR et **SAPIN**.
Les lettres du premier mot sont **[A,I,P,R]**, celles du second sont **[A,I,N,P,S]**. Les lettres communes sont donc **[A,I,P]**, il y en a donc 3. Toutes les lettres sont **[A,I,N,P,R,S]**, il y en a donc 6. La

distance de Jaccard entre **PAIR** et **SAPIN** est donc :

$$d = 1 - \frac{3}{6} = 1 - \frac{1}{2} = 1 - 0,5 = 0,5.$$

Exemple 2. LETTRE et **TARTE**.
Les lettres du premier mot sont **[E,E,L,R,T,T]**, celles du second **[A,E,R,T,T]**. Les lettres **[A,E,E,L,R,T,T]** permettent d'écrire chacun des mots, il y a donc un total de 7 lettres. Les lettres communes sont **[E,R,T,T]**, il y en a donc 4. La distance de Jaccard entre **LETTRE** et **TARTE** est donc :

$$d = 1 - \frac{4}{7} \simeq 1 - 0,57 \simeq 0,43.$$

Calcule la distance de Jaccard entre les mots suivants :

PLACE	**CRAIE**
AVRIL	**LAIT**
ESPOIR	**PROIE**
STATUE	**ETAT**
NOIR	**BLANC**
OBTENIR	**ROBINET**

Calcule la distance de Jaccard entre **CHIEN** et **NICHE**. Quand est-ce que deux mots ont une distance de Jaccard égale à 0 ?

Activité 3 (Distance de Levenshtein)**.**
On définit trois opérations qui permettent de passer d'un mot à un autre :

1. suppression d'une lettre,
2. ajout d'une lettre,
3. remplacement d'une lettre.

Voici un exemple de chaque type :

1. **PLANTE** vers **PLATE** (la lettre **N** est supprimée),
2. **RAPE** vers **RAMPE** (la lettre **M** est ajoutée),
3. **RAMER** vers **RALER** (la lettre **M** est remplacée par la lettre **L**).

La *distance de Levenshtein* entre deux mots est le nombre minimum d'opérations à effectuer afin de passer du premier mot au second.
Par exemple, la distance entre **PORT** et **PAR** vaut 2.

PORT devient **POR** puis devient **PAR**

On a appliqué l'opération 1, puis l'opération 3. Et pour cet exemple, on ne peut pas faire moins de deux opérations.

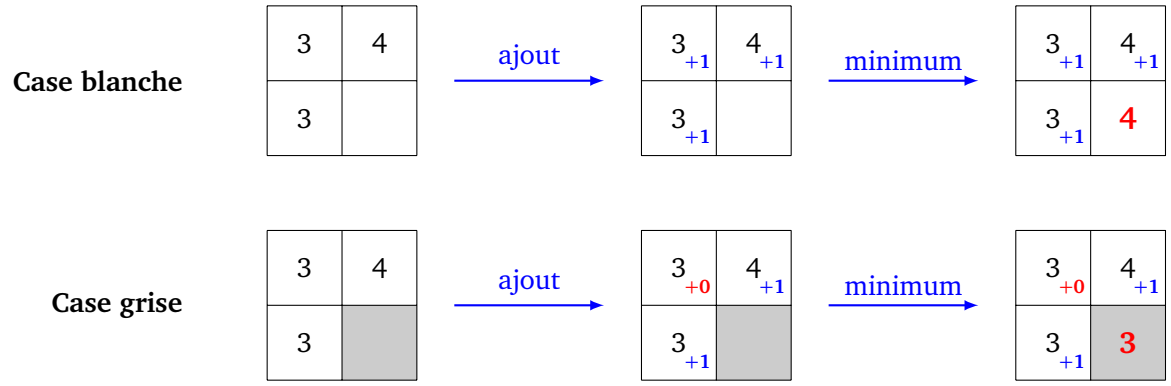
1. Trouve quelle opération permet de passer du premier au second mot :
   - **CLEF** vers **CLE**
   - **PILE** vers **PALE**
   - **MIEL** vers **CIEL**

2. D'après toi, combien vaut la distance de Levenshtein pour les paires de mots suivantes :
   - **SPIRE** et **PITRE**
   - **POMME** et **POIRE**
   - **PILE** et **PLI**
   - **LOUPE** et **POULE**
   - **LAMPE** et **PLACE**
   - **ROIS** et **OIE**
   - **PRISE** et **ROSE**
   - **ANANAS** et **BANANE**

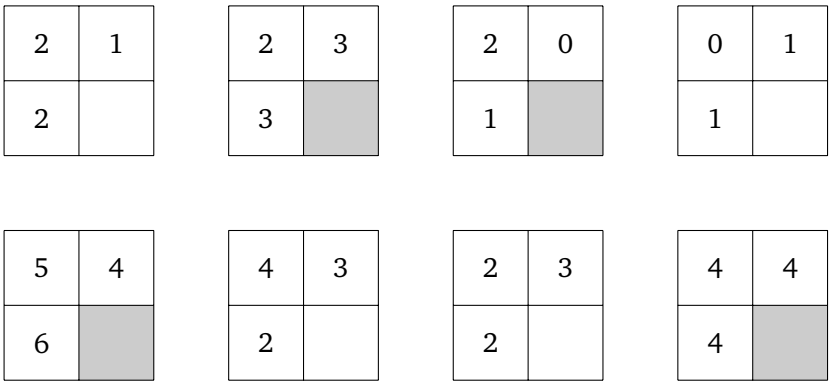
Activité 4 (Algorithme de Levenshtein)**.**
Nous allons voir une méthode systématique pour calculer la distance de Levenshtein et trouver les opérations qui permettent de passer d'un mot à un autre.

1. **Règle du minimum.**
   Nous remplissons la quatrième et dernière case d'un petit tableau (2×2) par la règle suivante :
   - si la quatrième case est blanche : on ajoute +1 à tous les nombres et on prend le minimum des trois nombres ;
   - si la quatrième case est grise : on ajoute +1 seulement aux cases au-dessus et à gauche, puis on prend le minimum des trois nombres.

   Voici deux exemples : avec une case blanche et avec une case grise.



   Complète les tableaux suivants :



2. **Initialisation.**
   Pour calculer la distance entre deux mots, nous allons faire les calculs à l'aide d'un tableau. Avant de commencer les calculs, voici la disposition de départ :

- le premier mot est écrit en colonne, le second en ligne ;
- on remplit une colonne et une ligne d'entiers $0, 1, 2, \ldots$;
- on grise les cases lorsque les deux lettres de chaque mot sont identiques.

Exemple simple avec **PAS** et **PLAT** (il y a une case grise pour le **P** commun et une grise pour le **A** commun).

		P	**L**	**A**	**T**
P					
A					
S					

		P	**L**	**A**	**T**
	0	1	2	3	4
P	1				
A	2				
S	3				

		P	**L**	**A**	**T**
	0	1	2	3	4
P	1				
A	2				
S	3				

Voici un autre exemple avec **VOILE** et **CERISE**.

		C	**E**	**R**	**I**	**S**	**E**
	0	1	2	3	4	5	6
V	1						
O	2						
I	3						
L	4						
E	5						

Trace les grilles, la numérotation et grise les cases pour les paires de mots :

BUS	et	**BRUT**
FRUIT	et	**CRIS**
PETITE	et	**LETTRE**
AVION	et	**BATEAU**

3. **Calcul de la distance de Levenshtein.**

Voici un algorithme pour calculer la distance de Levenshtein entre deux mots.

- Initialise le tableau avec un mot en colonne et l'autre en ligne.
- Remplis une ligne et une colonne à l'aide d'entiers consécutifs en partant de 0.
- Grise les cases lorque deux lettres de chaque mot sont identiques.
- Remplis, une à une, les cases avec la règle des minimums.
- La distance de Levenshtein est la valeur dans la case en bas à droite.

Reprenons l'exemple : de **PAS** à **PLAT**

		P	L	A	T
	0	1	2	3	4
P	1				
A	2				
S	3				

(a) Initialisation

		P	L	A	T
	0	1	2	3	4
P	1	**0**			
A	2				
S	3				

(b) Première case

		P	L	A	T
	0	1	2	3	4
P	1	**0**	**1**	**2**	**3**
A	2				
S	3				

(c) Première ligne

		P	L	A	T
	0	1	2	3	4
P	1	**0**	**1**	**2**	**3**
A	2	**1**	**1**	**1**	**2**
S	3	**2**	**2**	**2**	**2**

(d) Tableau et distance

(a) Le tableau est initialisé comme à la question précédente ; (b) on commence à compléter le tableau en suivant la règle du minimum ; (c) on remplit une à une les cases de la première ligne avant de passer à la suivante ; (d) le tableau est complet ; la distance de Levenshtein est la valeur en bas à droite.

Voici un exemple plus compliqué, qui calcule la distance de Levenshtein entre **VOILE** et **CERISE** qui vaut 4.

		C	E	R	I	S	E
	0	1	2	3	4	5	6
V	1	**1**	**2**	**3**	**4**	**5**	**6**
O	2	**2**	**2**	**3**	**4**	**5**	**6**
I	3	**3**	**3**	**3**	**3**	**4**	**5**
L	4	**4**	**4**	**4**	**4**	**4**	**5**
E	5	**5**	**4**	**5**	**5**	**5**	**4**

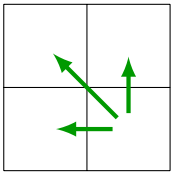
Calcule la distance de Levenshtein entre les mots suivants :

BUS	et	**BRUT**
FRUIT	et	**CRIS**
PETITE	et	**LETTRE**
AVION	et	**BATEAU**

4. **Retrouver les opérations.**

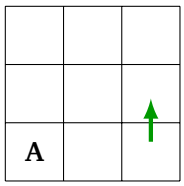
Avec un peu plus d'efforts, on retrouve les opérations nécessaires pour passer d'un mot à l'autre.

- Il faut d'abord trouver un chemin décroissant de la dernière case calculée (qui contient la distance de Levenshtein), vers la première case (qui contient le 0 en haut à gauche). Pour une case, on va vers l'une des trois cases qui a permis de réaliser la règle du minimum (donc vers une valeur la plus petite possible).

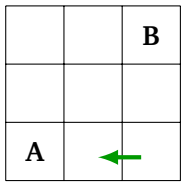


Ensuite, on passe du mot vertical au mot horizontal, en associant à certaines flèches une opération :

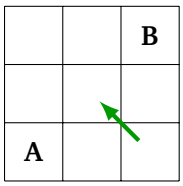
- si la valeur reste la même, on ne réalise aucune opération ;
- si la flèche est ↑, on supprime une lettre du mot vertical ;
- si la flèche est ←, on insère une lettre du mot horizontal dans le mot vertical ;
- si la flèche est ↖, on remplace une lettre du mot vertical par une lettre du mot horizontal.



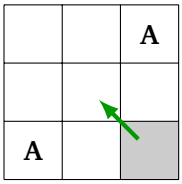
Supprimer **A**



Insérer **B** après **A**



Remplacer **A** par **B**



Ne rien faire.

Comment passer de **PAS** à **PLAT** en 2 opérations ? Tout d'abord, on trouve un chemin décroissant en suivant la règle des minimums.

		P	L	A	T
	0	1	2	3	4
P	1	0	1	2	3
A	2	1	1	1	2
S	3	2	2	2	2

Ainsi, en partant du bas à droite :

- on part du mot vertical **PAS** ;
- première flèche ↖ : on remplace le **S** par un **T** : **PAT** ;
- seconde flèche ↖ : comme on ne change pas de valeur, on ne fait rien ;
- troisième flèche ← : on insère le **L** après le **P** : **PLAT** ;
- dernière flèche ↖ : comme on ne change pas de valeur, on ne fait rien.

On a donc :

PAS → **PAT** → **PLAT**

Pour passer de **VOILE** à **CERISE**, voici un chemin (d'autres sont possibles).

		C	E	R	I	S	E
	0	1	2	3	4	5	6
V	1	1	2	3	4	5	6
O	2	2	2	3	4	5	6
I	3	3	3	3	3	4	5
L	4	4	4	4	4	4	5
E	5	5	4	5	5	5	4

En partant du bas à droite et en ne considérant que les flèches où la valeur change :

- la deuxième flèche est ↖ : **VOILE** devient **VOISE** ;
- la quatrième flèche est ← : **VOISE** devient **VORISE** ;
- la cinquième flèche est ↖ : **VORISE** devient **VERISE** ;
- la sixième flèche est ↖ : **VERISE** devient **CERISE**.

Trouve les opérations qui permettent de passer d'un mot à l'autre en un minimum d'étapes :

BUS	et	**BRUT**
FRUIT	et	**CRIS**
PETITE	et	**LETTRE**
AVION	et	**BATEAU**