ECOLE NATIONALE D'INGÉNIEUR DE TUNIS

Soutenance de mini projet

Moteur de matching de nom

Réalisé par :
Roua Oueslati & Melek jemili

Présentation du Projet

Moteur de Matching de nom

Le moteur de matching constitue le cœur du système de traitement des noms .Il a pour rôle de traiter des listes de noms en appliquant des techniques de recherche, de comparaison et de déduplication.

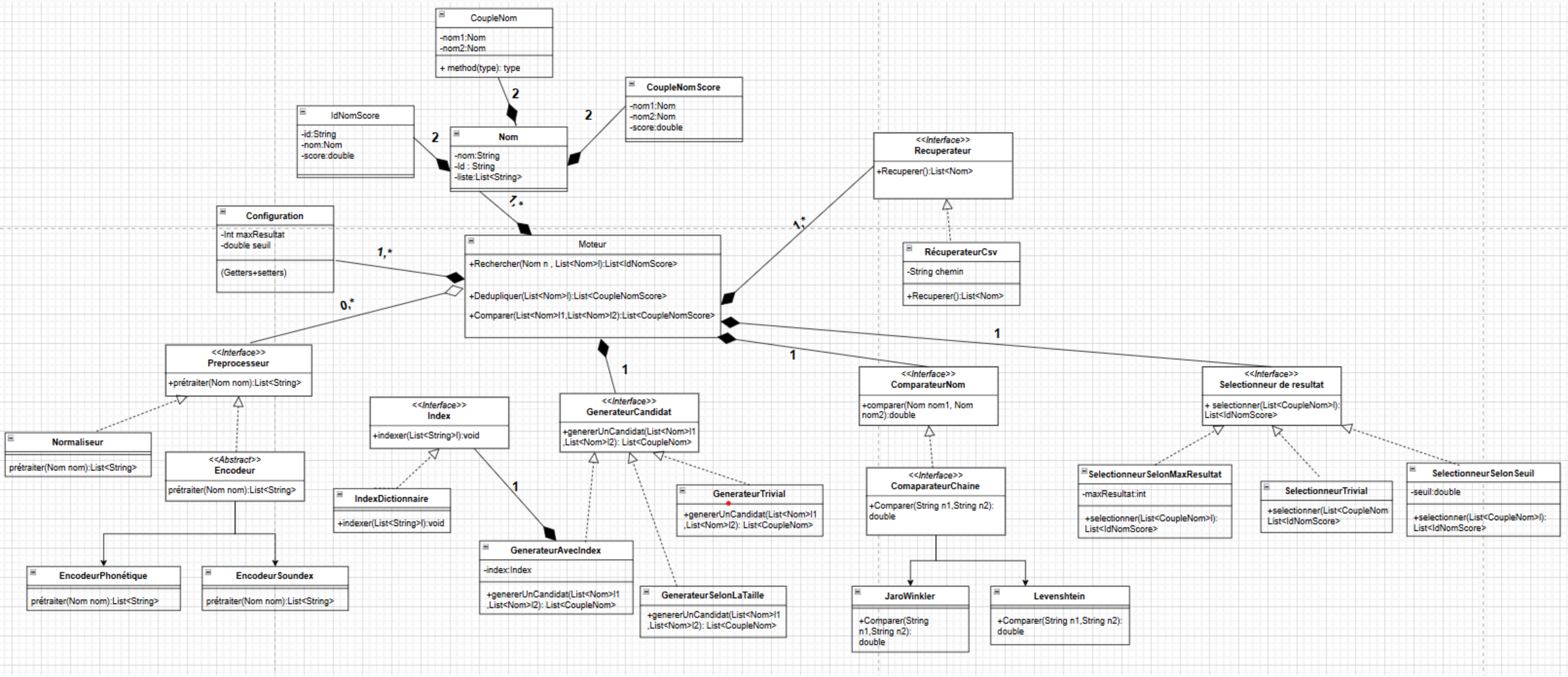
La méthode **rechercher** permet de retrouver un nom dans un fichier CSV en combinant : le prétraitement pour standardiser les données , l'indexation pour optimiser l'accès au informations et une mesure de similarité pour évaluer les correspondances enfin un sélectionneur doit filtrer les résultats fournies .

La méthode **dédupliquer** retourne le liste des noms doublons dans un fichier CSV en appliquant le même traitement que la méthode rechercher tout en conservant qu'une seule occurrence des entrés similaires .

La méthode **comparer** qui a pour rôle de détecter les correspondances entre deux fichiers CSV distinctes et retourne un couple de nom comparé avec leur score calculé à l'aide d'une mesure de similarité choisie.

Diagramme de Classes

Diagramme de Classes



Exécution de la recherche

RECHERCHE DANS 800:

```
Votre choix : 1
Saisir  le nom à rechercher:
To LAM
Fournir le fichier CSV:
c:/Users/EXTRA/OneDrive/Bureau/fichier/peps_names_800.csv
Nombre d'éléments dans liste1: 800
Résultat de recherche de To LAM:
id: NK-4BFW3CnQSzEkYeYyi3M7YU nom: To LAM score: 1.0
id: NK-4Fmc5dEYLhNtQM4yzt3Upb nom: To LAM score: 1.0
id: NK-Lk3jsaKE9r5KNq5dwXoVUw nom: "Boois score: 0.0
id: Q100138576 nom: Person score: 0.0
id: Q100138616 nom: Person score: 0.0
id: Q100138677 nom: Person score: 0.0
id: Q100138928 nom: Person score: 0.0
id: Q100146568 nom: Tsoede score: 0.0
id: Q100151842 nom: Person score: 0.0
id: Q100151864 nom: Person score: 0.0
id: Q100155186 nom: "Уткин score: 0.0
id: Q1001652 nom: "Firtl score: 0.0
```

```
Nombre d'éléments dans liste1: 800
Résultat de recherche de Kim Sung-du:
id: NK-3YjDY5aMyodvzSapTXwBap nom: Kim Sung-du score: 1.0
id: NK-48GBUX5D8rRe4sEzSAVVaA nom: Wayne Scott score: 0.0
id: NK-4Y628EQF3zdkudnaxQjKAn nom: Nato Taiwia score: 0.0
id: NK-62dyQCoW6zPMHoaV9W6H4D nom: Kang Su-rin score: 0.0
id: NK-6PGnZUqZn6cNddiWuuVccb nom: Ali al-ABID score: 0.0
id: NK-6g6bhJYQAAYmFKR7mqckRE nom: [illegible] score: 0.0
id: NK-8K7MAfEPSxP9ezEGuZHQxR nom: Tan Kok Yam score: 0.0
id: NK-8wPYaJSwftbAJnv8EBccic nom: Kim Jin-kuk score: 0.0
id: NK-9BkQCLWH94JXyueuKZSssq nom: Dileep Nair score: 0.0
id: NK-AWyMEdQGDxEEyJ8TrMT7G4 nom: Kim Jong-ok score: 0.0
id: NK-B4iN25Yc7vXzMRfyEr9PGH nom: Ismail ESAU score: 0.0
id: NK-BFDhuGBocu5yjiWXVbtidG nom: Pak Kum-hui score: 0.0
id: NK-Br6YBcimMN6a2ZSTvNaciG nom: [illegible] score: 0.0
id: NK-CMqpboBSPSQW5vbKiYf4UZ nom: Cecil McKIE score: 0.0
id: NK-Ek4t7MWubYLoEB5sS2azVP nom: Mr. Wang Yi score: 0.0
id: NK-FqSYqiJh3xZDqDvao26yYH nom: Alpha SESAY score: 0.0
id: NK-JBm3Bet4MdTeWcC3XchvDN nom: Cho Yong-su score: 0.0
id: NK-JBsPp2pvCNu4ovcPkLYNif nom: Andrew KWAN score: 0.0
id: NK-LKgfCDKxTgxNZqfNtRrJjj nom: DANATA Paul score: 0.0
id: NK-LUNmjRgsbawoDEDF9kVWKU nom: Arben Gashi score: 0.0
id: NK-NVmQpHnXRctWqid3HhyoXu nom: Karen CHONG score: 0.0
id: NK-NWimo9bKX4KBY9v4gKj6Hg nom: Kim Yong-ho score: 0.0
```

RECHERCHE DANS 1000:

```
Nombre d'éléments dans liste1: 1000
Résultat de recherche de Mustapha Abubakar:
id: NK-2AaUhJdZeBxnxVKVF7fTVv nom: Marica Montemaggi score: 0.0
id: NK-2wf3y8pBpVqPuU4neGTVbY nom: Mustapha Abubakar score: 1.0
id: NK-45hRm9KFehNsMxt8QYSUYN nom: NIBIGIRA Ezéchiel score: 0.0
id: NK-5coggoDn3ein6xorxZeYmW nom: Ogundimu Oluyinka score: 0.0
id: NK-5hJDngM8Evt9sUN2dqyLyb nom: Husayn al-QATRANI score: 0.0
id: NK-79kK8saK3fHEUE3V5rqM8e nom: Giorgi Mukbaniani score: 0.0
id: NK-8gS9PcxB2L6cCuQ7JJeyuK nom: Dwayne S. Seymour score: 0.0
id: NK-AxcnWtuu4JKBrZFgR7NtjP nom: วิทวัส บุญญสถิตย์ score: 0.0
id: NK-BTqdvNKmAmyYE2n74sK4uS nom: GAFURERO Léocadie score: 0.0
id: NK-CNqPJdh6fn9QCuUWHE7YPJ nom: Chukwuma Nwazunku score: 0.0
id: NK-CS5tgzSxgDoC5jFLvYFuxT nom: Sarjit Singh GILL score: 0.0
id: NK-Fa9cztB5iftXe7NknXnDst nom: Wilmer Leal Pérez score: 0.0
id: NK-FszBgxS59nUsi3iox6SUfj nom: NDUWIMANA Edouard score: 0.0
id: NK-GchNZeNGEqRnSZSd3WBeBF nom: Yogida Sawmynaden score: 0.0
id: NK-HS2bW47n75pDzZCqUjW33K nom: Giorgi Gelashvili score: 0.0
id: NK-JTm97UHayZHbfu83yUL7ZU nom: HABONIMANA Odette score: 0.0
id: NK-K857svDCCXSF6zmzUKWPSW nom: สิงห์สึก สิงห์ไพร score: 0.0
id: NK-LfRwjzNHCtyTz7zVaGZd26 nom: Phout Simmalavong score: 0.0
id: NK-Nhbi3pGsJ7zFABcxx2yTbX nom: Renato FLORENTINO score: 0.0
id: NK-P47DSVBMhkjXq8Yz8J2xSt nom: Ketevan Jachvadze score: 0.0
id: NK-PHXA238b37u5GM86yUxnXA nom: Ketleen FLORESTAL score: 0.0
id: NK-O5czaG7YUZBt3NWiWAkpSL nom: Karupu Sebastiaan score: 0.0
```

RECHERCHE DANS 658K:

```
Menu [Java Application] C:\Program Files\Java\jdk-21\bin\javaw.exe (12 mai 20
id: Q111457967 nom: Person score: 1.0
id: Q111593161 nom: GaaSyy score: 0.0
id: Q111600205 nom: Person score: 1.0
id: Q111792961 nom: Person score: 1.0
id: Q111909058 nom: Dai Le score: 0.0
id: Q111911956 nom: B. Ram score: 0.0
id: Q111912406 nom: Artaha score: 0.0
id: Q111912568 nom: Tagtal score: 0.0
id: Q111912621 nom: Akheqa score: 0.0
id: Q111934064 nom: Person score: 1.0
id: Q112084996 nom: Person score: 1.0
id: Q112447502 nom: Person score: 1.0
id: Q112659775 nom: Tsieté score: 0.0
id: Q112710881 nom: Tuy Ry score: 0.0
id: Q112965986 nom: Mo Hua score: 0.0
id: Q112991655 nom: "Жулин score: 0.0
id: Q112992783 nom: "Вагин score: 0.0
id: Q112992794 nom: "Дааев score: 0.0
id: Q112992858 nom: Person score: 1.0
id: Q112993042 nom: "Хорев score: 0.0
id: Q112993211 nom: "Ёлкин score: 0.0
id: Q112993344 nom: Person score: 1.0
id: Q112993579 nom: "Осина score: 0.0
id: Q112993589 nom: "Басик score: 0.0
```

Exécution de la comparaison

COMPARAISON 100 ET 200:

```
Votre choix : 2
Fournir le premier fichier CSV:
c:/Users/EXTRA/OneDrive/Bureau/fichier/peps_names_200.csv
Fournir le deuxième fichier CSV:
c:/Users/EXTRA/OneDrive/Bureau/fichier/peps_names_100.csv
Résultat de comparaison:
Nombre de couples générés : 949
nom1: DEMIDOVICH VASILIJ nom2: DEMIDOVICH VASILIJ score: 1.0
nom1: DEMIDOVICH VASILIJ nom2: Mr. Trevor Prescod score: 0.0
nom1: DEMIDOVICH VASILIJ nom2: Samonyane Ntsekele score: 0.0
nom1: DEMIDOVICH VASILIJ nom2: Ibrahim Bawa Kamba score: 0.0
nom1: DEMIDOVICH VASILIJ nom2: Joseph Obinna Ogba score: 0.0
nom1: DEMIDOVICH VASILIJ nom2: Dina Nath Dhungyel score: 0.0
nom1: Marica Montemaggi nom2: Marica Montemaggi score: 1.0
nom1: Marica Montemaggi nom2: Mustapha Abubakar score: 0.0
nom1: Marica Montemaggi nom2: NIBIGIRA Ezéchiel score: 0.0
nom1: Marica Montemaggi nom2: Ogundimu Oluyinka score: 0.0
nom1: Marica Montemaggi nom2: Husayn al-QATRANI score: 0.0
nom1: Marica Montemaggi nom2: Giorgi Mukbaniani score: 0.0
nom1: Marica Montemaggi nom2: Dwayne S. Seymour score: 0.0
nom1: PHIRAPHAN Saliratthawiphak nom2: PHIRAPHAN Saliratthawiphak score: 1.0
nom1: PHIRAPHAN Saliratthawiphak nom2: Héctor Ireneo Mares Cossío score: 0.0
nom1: PHIRAPHAN Saliratthawiphak nom2: Saúl Antonio Ortega Campos score: 0.0
```

COMPARAISON 4K ET 16K:

```
nom1: Ri Myong-chol nom2: Ri Myong-chol score: 1.0
nom1: Ri Myong-chol nom2: SNG Chern Wei score: 0.0
nom1: Ri Myong-chol nom2: LIM Der Shing score: 0.0
nom1: Ri Myong-chol nom2: Mari Alkatiri score: 0.0
nom1: Ri Myong-chol nom2: Mahen JHUGROO score: 0.0
nom1: Ri Myong-chol nom2: Kirk HUMPHREY score: 0.0
nom1: Kim Kum-chol nom2: Kim Kum-chol score: 1.0
nom1: Kim Kum-chol nom2: Aida Dërguti score: 0.0
nom1: Kim Kum-chol nom2: ELIAS GONDJI score: 0.0
nom1: Kim Kum-chol nom2: Kang Yong-su score: 0.0
nom1: Kim Kum-chol nom2: Choe Yong-ho score: 0.0
nom1: Kim Kum-chol nom2: Pak Jong-nam score: 0.0
nom1: Kim Kum-chol nom2: Pak Chun-nam score: 0.0
nom1: ဦးဇော်သာဥ: nom2: ဦးဇော်သာဥ: score: 1.0
nom1: Héctor Ireneo Mares Cossío nom2: PHIRAPHAN Saliratthawiphak score: 0.0
nom1: Héctor Ireneo Mares Cossío nom2: Héctor Ireneo Mares Cossío score: 1.0
nom1: Héctor Ireneo Mares Cossío nom2: Saúl Antonio Ortega Campos score: 0.0
nom1: Héctor Ireneo Mares Cossío nom2: GALLO CANTERA LUIS ENRIQUE score: 0.0
nom1: Mustapha Abubakar nom2: Marica Montemaggi score: 0.0
nom1: Mustapha Abubakar nom2: Mustapha Abubakar score: 1.0
nom1: Mustapha Abubakar nom2: NIBIGIRA Ezéchiel score: 0.0
nom1: Mustapha Abubakar nom2: Ogundimu Oluyinka score: 0.0
nom1: Mustapha Abubakar nom2: Husayn al-QATRANI score: 0.0
```

COMPARAISON 16K ET 32K:

```
nom1: Selguet Née Achta Aguidi nom2: Dorel-Gheorghe Acatrinei score: 0.0
nom1: Selguet Née Achta Aguidi nom2: Valentin-Ilie Făgărășian score: 0.0
nom1: Selguet Née Achta Aguidi nom2: Muhamad Yusoff Mohd Noor score: 0.0
nom1: Selguet Née Achta Aguidi nom2: Carolyn Trench-Sandiford score: 0.0
nom1: Selguet Née Achta Aguidi nom2: Valentin Rică Cioromelea score: 0.0
nom1: Selguet Née Achta Aguidi nom2: Aditya Halindra Faridzky score: 0.0
nom1: Selguet Née Achta Aguidi nom2: Prabhakar Balwant Vaidya score: 0.0
nom1: Selguet Née Achta Aguidi nom2: Ganesh Shankar Vidyarthi score: 0.0
nom1: Selguet Née Achta Aguidi nom2: Sardar Tara Singh Ghaiba score: 0.0
nom1: Selguet Née Achta Aguidi nom2: George-Cristian Mitricof score: 0.0
nom1: Selguet Née Achta Aguidi nom2: Willis E. Blackshear Jr. score: 0.0
nom1: Selguet Née Achta Aguidi nom2: Jordi Hernández Martínez score: 0.0
nom1: Selguet Née Achta Aguidi nom2: Sylvanus Adiewere Nsofor score: 0.0
nom1: Selguet Née Achta Aguidi nom2: Mohammed Boakye Agyemang score: 0.0
nom1: Selguet Née Achta Aguidi nom2: Darko Asomaning Nicholas score: 0.0
nom1: Selguet Née Achta Aguidi nom2: "Jonathan Robert Owiredu score: 0.0
nom1: Selguet Née Achta Aguidi nom2: Adam Baako Nortey Yeboah score: 0.0
nom1: Selguet Née Achta Aguidi nom2: Emmanuel Oscar Ameyedowo score: 0.0
nom1: Selguet Née Achta Aguidi nom2: "Frederic W. Schlosstein score: 0.0
nom1: Selguet Née Achta Aguidi nom2: Edward D. Harrington Jr. score: 0.0
nom1: Selguet Née Achta Aguidi nom2: Maria Anatolyevna Livzan score: 0.0
nom1: Selguet Née Achta Aguidi nom2: Charlotte Bach Thomassen score: 0.0
nom1: Selguet Née Achta Aguidi nom2: Antonio Martorell Lacave score: 0.0
nom1: Selguet Née Achta Aguidi nom2: Dukakinzade Mehmed Pasha score: 0.0
```

Exécution de la déduplication

DÉDUPLICATION DE 1K:

```
------------------------------------
Votre choix : 3
Fournir le fichier CSV à traiter:
c:/Users/EXTRA/OneDrive/Bureau/fichier/peps_names_4k.csv
Résultat de la déduplication:
ID: NK-4Fmc5dEYLhNtQM4yzt3Upb Nom: To LAM
ID: Q100403535 Nom: Shafiqul Islam
ID: Q101438914 Nom: Aftab Hussain
ID: Q10335644 Nom: Li Yang
ID: NK-LaufmSgR3SwndSfX8rZeYx Nom: THIEN Kwee Eng
ID: Q101094857 Nom: "Медведев
ID: Q100138616 Nom: Person
ID: NK-YnpuTdyJus5aivTB8Cidg9 Nom: Ri Yong-chol
```

DÉDUPLICATION DE 4K:

```
------------------------------------
Votre choix : 3
Fournir le fichier CSV à traiter:
c:/Users/EXTRA/OneDrive/Bureau/fichier/peps_names_4k.csv
Résultat de la déduplication:
ID: NK-4Fmc5dEYLhNtQM4yzt3Upb Nom: To LAM
ID: Q100403535 Nom: Shafiqul Islam
ID: Q101438914 Nom: Aftab Hussain
ID: Q10335644 Nom: Li Yang
ID: NK-LaufmSgR3SwndSfX8rZeYx Nom: THIEN Kwee Eng
ID: Q101094857 Nom: "Медведев
ID: Q100138616 Nom: Person
ID: NK-YnpuTdyJus5aivTB8Cidg9 Nom: Ri Yong-chol
```

DÉDUPLICATION DE 64K:

```
Menu [Java Application] C:\Program Files\Java\jdk-21\bin\javaw.exe  (12 mai 2025, 20:45:56 el
c:/Users/EXTRA/OneDrive/Bureau/fichier/peps_names_64k.c
Résultat de la déduplication:
ID: Q112800328 Nom: Mary Miller
ID: Q125747292 Nom: Mike Brown
ID: Q110655570 Nom: Oumou Coulibaly
ID: Q116199650 Nom: Malachie MANAOUDA
ID: Q11066687 Nom: Zhang Guochu
ID: Q111281518 Nom: Abigail Damasane
ID: Q126369965 Nom: Rajesh Kumar Singh
ID: Q114243575 Nom: Dragoslav Jovanović
ID: Q110095791 Nom: Sunil Kumar
ID: Q109489128 Nom: "Волков
ID: Q115801176 Nom: Rajesh Kumar Mishra
ID: Q108593769 Nom: Balbir Singh
ID: Q12578390 Nom: "Ахметов
ID: Q118808 Nom: Christiane Brunner
ID: Q110993772 Nom: Leah Scott
ID: Q111906812 Nom: Vicente Sarmiento
ID: Q104290542 Nom: "Кузнецов
ID: Q116056503 Nom: Yordan Yordanov
ID: Q12499475 Nom: Mulyadi
ID: Q101438914 Nom: Aftab Hussain
ID: Q125867206 Nom: Virendra Singh
ID: Q111938020 Nom: Michael Powell
ID: Q11551854 Nom: Yoshitaka Ikeda
ID: Q109772096 Nom: Franz I. Manderson
ID: Q111831069 Nom: Nirmal Singh
ID: Q116056175 Nom: Galina Georgieva
ID: Q116056183 Nom: Nikolay Kostadinov
ID: Q12593548 Nom: Tong Jong-ho
```

DÉDUPLICATION DE 512K:

```
ID: Q131174543 Nom: William John Hill
ID: Q133302824 Nom: "Гречин
ID: Q115268452 Nom: Philippe Lacombe
ID: Q133301824 Nom: "Юнусов
ID: Q130526162 Nom: John Evans
ID: Q133299854 Nom: "Постников
ID: Q111141531 Nom: Ibrahim Umar Potiskum
ID: Q133304911 Nom: "Силаев
ID: Q133302144 Nom: "Яшкин
ID: Q133300665 Nom: "Тимофеев
ID: Q133300199 Nom: "Субботин
ID: Q16204904 Nom: Robert Garcia
ID: Q112996329 Nom: "Волынский
ID: Q133300368 Nom: "Шаталов
ID: Q124350768 Nom: Jodie Haydon
ID: Q133302996 Nom: "Королькова
ID: Q133298341 Nom: "Шамхалов
ID: Q124769785 Nom: "Арсентьев
ID: Q133299851 Nom: "Осокин
ID: Q109455544 Nom: "Ходжаева
ID: Q107345606 Nom: Mzikayifane Elias Khumalo
ID: Q16730345 Nom: David Johnson
ID: Q23565789 Nom: Martin Schmidt
ID: Q133304439 Nom: "Клюшникова
ID: Q133304822 Nom: "Приходько
ID: Q113827327 Nom: Mary Jeanette Murray
ID: Q124289099 Nom: David Smith
ID: Q16181667 Nom: Ri Ryong-nam
ID: Q16194700 Nom: Gregory A. Miller
ID: Q124040032 Nom: Richard Henderson
```

DÉDUPLICATION DE 658K:

```
ID: Q111291907 Nom: Biggie J. Matiza
ID: Q112991649 Nom: "Решетников
ID: Q109487551 Nom: "Харченко
ID: Q114139749 Nom: José Nunes
ID: Q114397632 Nom: Chris Brown
ID: Q107298361 Nom: Phyllis Chemutai
ID: Q12281363 Nom: Ivo Atanasov
ID: Q12596635 Nom: Pak Myong-chol
ID: Q113409643 Nom: "Голохвастов
ID: Q124707136 Nom: John Fuller
ID: Q109772063 Nom: Samuel Bulgin
ID: Q103838581 Nom: "Болховский
ID: Q124771416 Nom: "Арефьев
ID: Q125178837 Nom: "Гагарин
ID: Q121234258 Nom: Abdul Wahid
ID: Q126616505 Nom: Ajay Kumar Singh
ID: Q111967756 Nom: Tyrone Garner
ID: Q124245993 Nom: Abul Kalam Azad
ID: Q12044268 Nom: Pavel Svoboda
ID: Q111804345 Nom: Bijay Kumar Singh
ID: Q113696219 Nom: Mary Jane McKenna
ID: Q125167244 Nom: Carol Bush
ID: Q124770453 Nom: "Макаров
ID: Q109567463 Nom: Jacqueline Lewis
ID: Q11817678 Nom: Piotr Kozłowski
ID: Q110707260 Nom: Michael Murphy
ID: Q12288374 Nom: Nikola Dimitrov
ID: Q12023148 Nom: Jan Svoboda
ID: Q110816136 Nom: "Лазарев
ID: Q120798039 Nom: Min Thein Zan
```

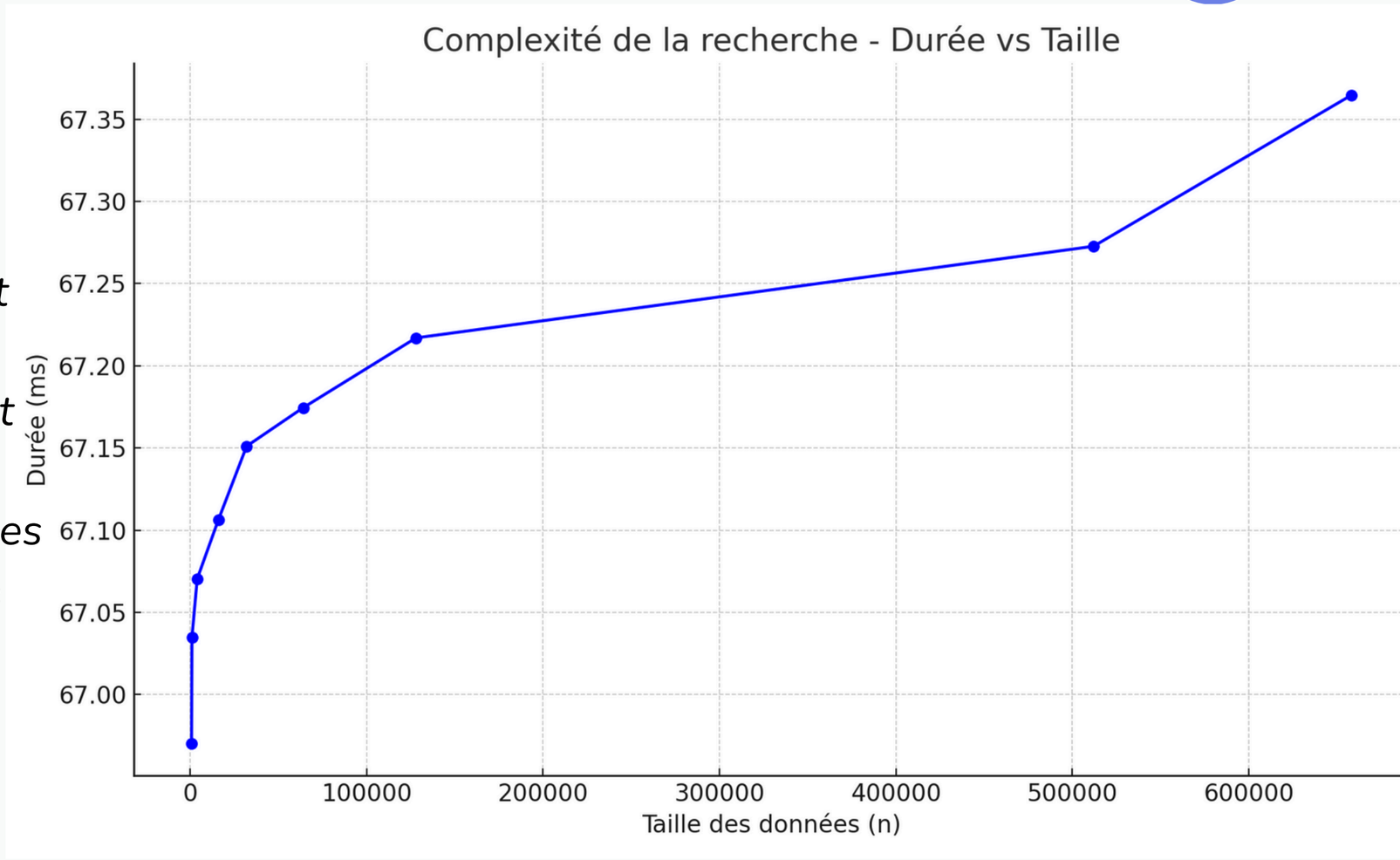
Complexité

Complexité de la recherche:

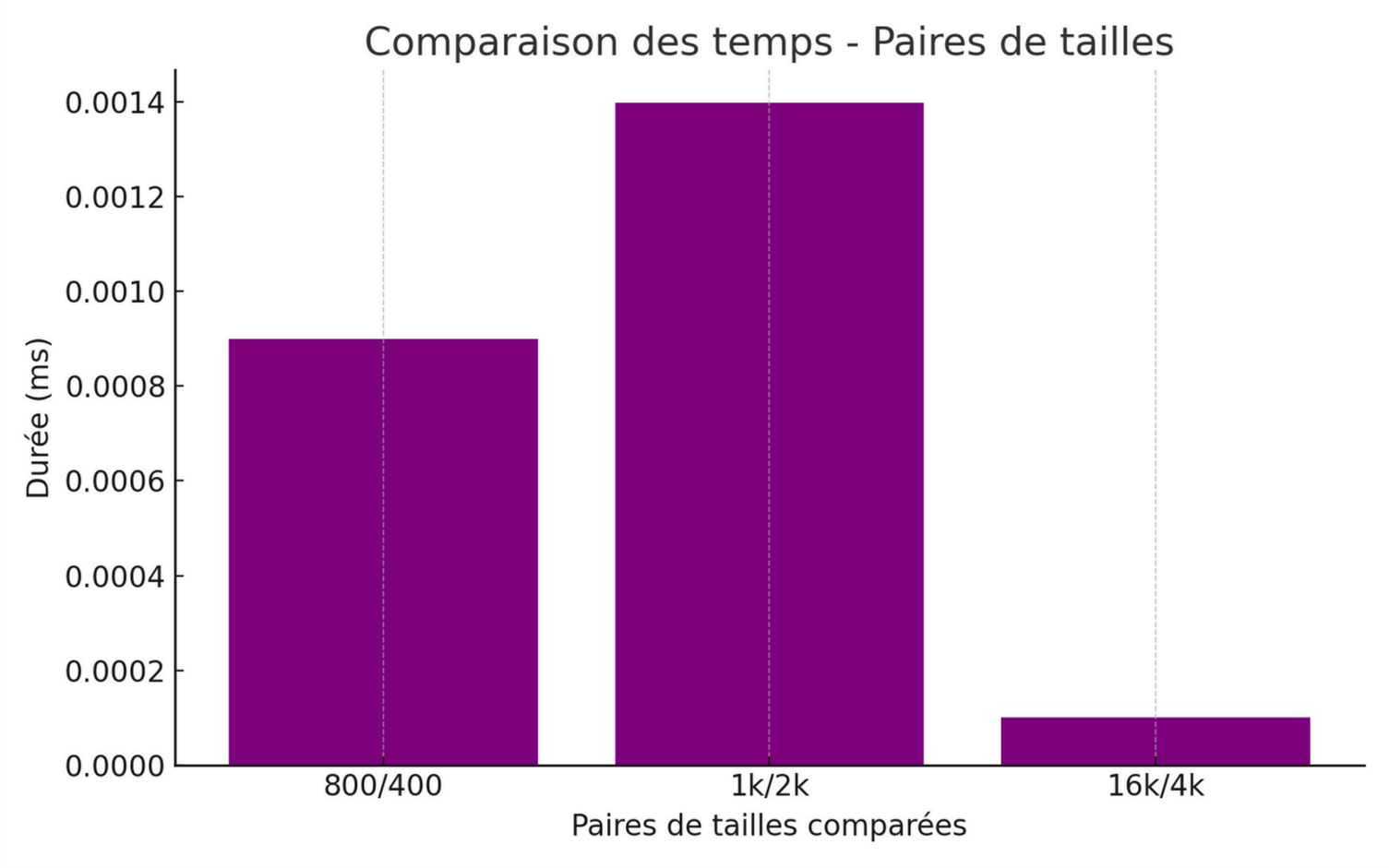
Les caractéristiques principales:

- *La durée augmente rapidement au début (entre 0 et ~50 000 éléments)*
- *Puis la courbe s'aplatit considérablement*
- *Une légère augmentation continue est visible pour les grandes tailles de données*

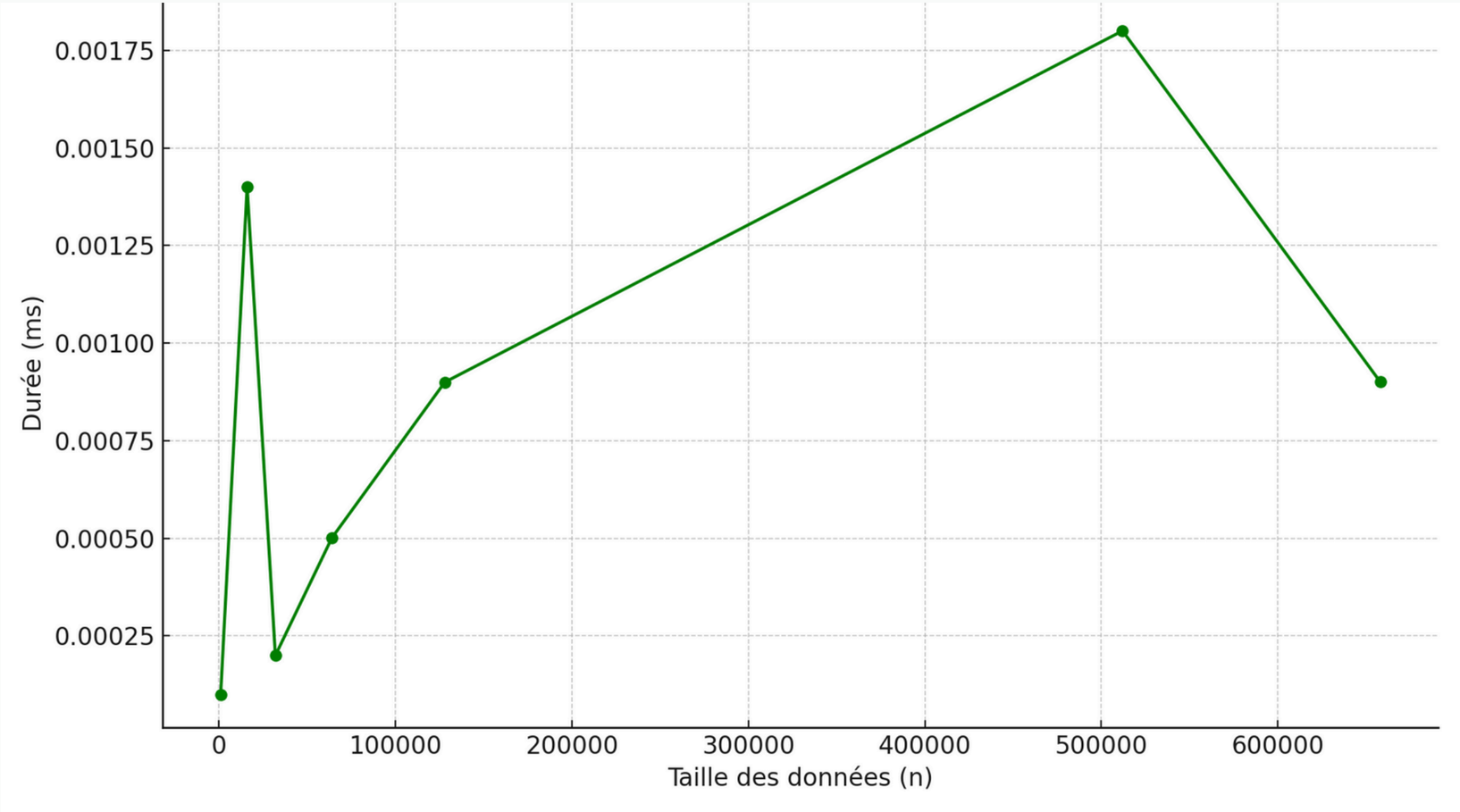
Cette forme logarithmique avec une croissance qui ralentit est typique d'algorithmes de complexité O(log n)



Complexité de la comparaison:



Complexité de la déduplication:



Synthèse

ce moteur est extrêmement optimisé, peut-être utilisé dans des systèmes à haute performance

L'algorithme semble très robuste face à l'augmentation des données.

Ce projet est idéal pour des applications traitant de grandes quantités de données rapidement et efficacement

Merci!