**État de l’art II**

**Outils de détection de plagiat**

Travail de Bachelor 2020

**Original Journalistic Information Label and Traceability**

Etudiant : Nicolas Solioz

Professeure : Nicole Glassey Balet

**Table des matières**

[Préambule 3](#_Toc38377403)

[Prepostseo 3](#_Toc38377404)

[Interface web 3](#_Toc38377405)

[API 5](#_Toc38377406)

[Plagiarism Checker by EduBirdie 7](#_Toc38377407)

[Interface web 7](#_Toc38377408)

[API 8](#_Toc38377409)

[Plagiarismsearch 8](#_Toc38377410)

[Interface web 8](#_Toc38377411)

[API 10](#_Toc38377412)

[Copyleaks 13](#_Toc38377413)

[Interface web 13](#_Toc38377414)

[API 15](#_Toc38377415)

[Unicheck 17](#_Toc38377416)

[Interface web 17](#_Toc38377417)

[API 21](#_Toc38377418)

[Comparaison 27](#_Toc38377419)

[Conclusion et choix 29](#_Toc38377420)

# Préambule

Il existe plusieurs critères permettant de déterminer la qualité d’un contenu journalistique, comme indiqué dans l’état de l’art I. Notre projet se concentre sur l’indicateur du plagiat. Cette caractéristique complexe peut être analysé via différents outils existants. Cet état de l’art analysera les solutions suivantes :

• Plagiarism Checker by EduBirdie

• plagiarismsearch

• copyleaks

• unicheck

• Prepostseo (pas étudié par l’institut)

Un état de l’art préliminaire a été déjà été réalisé par l’institut d’informatique de gestion pour ces logiciels, à l’exception de Prepostseo. Notre état de l’art amènera quelques compléments notamment via des démonstrations des solutions API et une comparaison des coûts des différents outils.

# Prepostseo

## Interface web

L’outil web proposé par « Prepostseo » permet d’insérer un texte dans un formulaire. L’outil s’occupe ensuite de déterminer si le texte est original ou non. La version gratuite de Prepostseo limite le nombre de mots à 1'000. Pour notre test, nous avons utilisé le texte d’un article du Temps.

Une image contenant capture d’écran

Description générée automatiquement

Une image contenant alimentation, dessin

Description générée automatiquement

Nous constatons que, puisque le texte a été copié et collé dans le formulaire, Prepostseo détermine que le 90% de l’article est plagié.

Une image contenant capture d’écran

Description générée automatiquement

Afin d’analyser l’efficacité de cet outil, nous avons également inséré un texte paraphrasé. Cela nous permet de déterminer si Prepostseo peut détecter les phrases reformulées.

Une image contenant capture d’écran

Description générée automatiquement

Nous constatons que la page web de source, [www.letemps.ch](http://www.letemps.ch), a été retrouvée malgré la reformulation des phrases.

## API

Cet outil, non testé par l’institut d’informatique de gestion, propose une API bien documentée. En théorie, cet outil devrait nous permettre de vérifier soit une phrase individuelle, soit un article entier.

Une image contenant capture d’écran

Description générée automatiquement

Le résultat retourne l’URL où la phrase a été reproduite.

Une image contenant capture d’écran

Description générée automatiquement

Ici, nous constatons que notre requête nous retourne l’article du Temps. Il nous renvoie également d’autres articles de wikipedia, rts etc…

# Plagiarism Checker by EduBirdie

## Interface web

Entièrement gratuit, le Plagiarism Checker offert par EduBirdie est exclusivement disponible sur leur site web <https://edubirdie.com/logiciel-anti-plagiat>. EduBirdie n’offre pas la possibilité de faire ces recherches avec une API.

Après plusieurs tests, nous constatons que EduBirdie ne détecte pas efficacement le plagiat. En effet, en effectuant l’analyse sur un texte existant provenant du site [www.letemps.ch](http://www.letemps.ch), EduBirdie n’était pas en mesure de détecter la source de l’article.

Une image contenant capture d’écran

Description générée automatiquement

Figure 1 – Analyse sur https://www.letemps.ch/economie/migros-abandonne-course-grandeur

Malgré la présence de l’article sur le site du temps, EduBirdie estime que le texte est unique à 100%. Cet outil nous semble donc pas suffisamment fiable pour être intégré dans notre projet.

## API

L’outil proposé par EduBirdie n’offre, à notre connaissance, aucune solution API.

# Plagiarismsearch

## Interface web

L’outil plagiarismsearch fonctionne de façon similaire à EduBirdie. C’est une plateforme web permettant de coller son texte dans un champ et de lancer une recherche de plagiat.

Afin de comparer les outils, nous avons utilisé le même article du Temps.

Rapidement, nous constatons que plagiarismsearch est un outil plus développé. En effet, ce dernier détecte le plagiat de l’article et offre même des liens menant vers les sites concernés.

Une image contenant capture d’écran

Description générée automatiquement

Afin de déterminer l’étendue de la détection de plagiat, nous avons réécrit le texte du Temps en remplaçant certains mots par des synonymes. En effet, il est intéressant de savoir si l’outil permet de détecter des phrases reformulées.

Voici un tableau représentant à gauche le texte original et à droite le texte modifié.

Tableau 1 - Modification du texte de l'article du Temps

|  |  |
| --- | --- |
| **Texte original** | **Texte modifié** |
| Le géant orange veut se séparer de quatre filiales, dont Globus et Interio, trop peu rentables et trop éloignées de sa nouvelle stratégie. L’alimentaire, l’e-commerce et la santé deviennent ses principaux piliers.  Si tant est qu’elle existait réellement, Migros renonce à gagner la course au chiffre d’affaires avec Coop. Finie la lutte de taille entre les deux géants orange de la grande distribution suisse. Migros a annoncé jeudi le début d’un processus de vente d’une ampleur inédite.  Globus, Interio, Gries Deco (Depot) et M-way. Ces quatre filiales sont désormais sur le marché. Si Migros parvient à trouver preneur, ce qu’il espère pouvoir annoncer en 2020, ses 28 milliards de francs de chiffre d’affaires seront amputés d’environ 1,5 milliard, soit 5% du total. | Le groupe Migros veut se séparer de certaines de ses filiales, comme Globus et Intero. En effet, le géant orange estime que ces entreprises sont trop peu rentables et trop éloignées de la stratégie du groupe. Ces nouveaux principaux pilliers sont l’alimentaire, l’e-commerce et la santé.  Migros abandonne la course au chiffre d’affaires avec Coop. La bataille de taille entre les géants oranges est terminée. Migros a annoncé cette semaine le début inédit d’un nouveau processus de vente.  Les quatre filiales, Globus, Interio, M-Way et Gries Deco sont désormais en vente. En trouvant acheteur, le chiffre d’affaires de Migros sera réduit de 5%, soit environ 1.5 milliard. |

Malgré la modification du texte, plagiarismsearch a réussi à retrouver les articles originaux. Nous constatons que le taux de plagiat est réduit, puisque les mots ne correspondent pas exactement.

Une image contenant capture d’écran

Description générée automatiquement

Ainsi l’outil permet de détecter le plagiat de manière efficace.

## API

Plagiarismsearch offre également l’accès à son outil de détection de plagiat via une API bien documentée. Afin de tester son fonctionnement, il nous est possible d’effectuer quelques requêtes gratuites. Cependant, afin de débloquer des requêtes supplémentaires il faut payer.

Pour tester les fonctionnalités de l’API, nous avons installé XAMPP qui nous permet de faire tourner en local un serveur Apache.

Une image contenant table

Description générée automatiquement

Plagiarismsearch nous permet de télécharger des scripts php d’exemple directement sur le site <https://plagiarismsearch.com/files/sample-api.zip>. Ces derniers ont été téléchargés et sauvegardés dans le répertoire de XAMPP C:\xampp\htdocs\plagiarismsearch.

Ensuite pour exécuter une requête, nous devons insérer nos identifiants de connexions dans le fichier init.api.php. Les identifiants sont disponibles sur le profil plagiarismsearch, directement sur le site.

Une image contenant capture d’écran

Description générée automatiquement

Ces informations sont mis de cette manière dans le fichier init-api.php :



Finalement, pour tester l’API nous souhaitons récupérer les informations de notre recherche déjà effectuée. Pour se faire, nous remplaçant l’identifiant du « report » dans le fichier status.php. L’identifiant du « report » est indiqué au sommet de la page de résultat de notre recherche, directement sur la page internet.

Une image contenant capture d’écran

Description générée automatiquement

Ce numéro doit être mis dans le fichier status.php de cette façon :

Une image contenant capture d’écran, oiseau

Description générée automatiquement

Une fois toute ces modifications faites, nous pouvons exécuter le script du fichier « status-php » directement dans notre navigateur web. Pour cela, Apache doit être démarré via XAMPP. Voici l’URL utilisé : <http://localhost:8080/plagiarismsearch/status.php>.

Une image contenant capture d’écran

Description générée automatiquement

L’API est fonctionnelle. Nous pouvons accéder au PDF de notre « report » en plusieurs langues.

Plagiarismsearch offre 100 requêtes. Toutes requêtes supplémentaires sont payantes.

# Copyleaks

## Interface web

Copyleaks, tout comme EduBirdie et plagiarismcheckers, possède une interface web nous permettant de vérifier du texte brut. Nous procédons avec le même texte de test provenant du Temps afin de déterminer l’efficacité de la détection du plagiat.

Une image contenant capture d’écran

Description générée automatiquement

Une image contenant capture d’écran

Description générée automatiquement

Nous constatons que les résultats sont similaires à ceux de plagiarismsearch. L’interface de Copyleaks est pourtant plus développée. Un système de code couleur permet de voir si les termes sont copiés exactement où paraphrasés. Si nous faisons le même essai avec le texte modifié (voir tableau 1), voici le résultat :

Une image contenant capture d’écran

Description générée automatiquement

L’URL original est retrouvée. Pourtant, le taux de plagiat est de 43% seulement alors que Plagiarismchecker nous donnait un taux de 80%. Il semblerait donc que Plagiarismchecker permet de mieux décéler des textes modifiés. Des tests supplémentaires doivent être effectués afin de valider cette hypothèse.

Nous notons également que Copyleaks effectue également des analyses dans la base de données interne de l’outil, ce qui peut permettre d’optimiser la recherche si l’outil est utilisé à large échelle.

Une image contenant capture d’écran

Description générée automatiquement

## API

Copyleaks offre également une API. Celle-ci est documentée en ligne : <https://api.copyleaks.com/>.

Le nombre de requête est limité à 10 pages. Les requêtes supplémentaires sont payantes. Nous avons testé l’API sur Postman.

Une image contenant texte

Description générée automatiquement

Dans un premier temps, nous devons obtenir un token d’identification. Ce token s’obtient via une méthode POST comprenant notre adresse e-mail et notre clé de profil.

Une image contenant capture d’écran

Description générée automatiquement

Une image contenant capture d’écran

Description générée automatiquement

Avec ce token, nous pouvons désormais effectuer, en théorie, du détection de plagiat avec du texte brut. Cependant, malgré la documentation disponible sur le site, nous n’avons pas réussi à effectuer une requête. Un courriel a été transmis à l’équipe de développement de Copyleaks, mais ces derniers nous n’ont pas transmis de précision concernant l’utilisation de leur outil. A ce jour, aucune réponse n’a été apporté par Copyleaks.

Après plus de recherches de notre côté, nous constatons que le problème vient certainement de la mise en place des « webhooks ».

Une image contenant capture d’écran

Description générée automatiquement

Cette propriété est nécessaire pour que la requête fonctionne. Nous devons ainsi configurer notre serveur afin qu’il puisse recevoir le statut de notre requête API. Cela nous semble être une étape supplémentaire encombrante, les autres outils sont plus simples.

# Unicheck

## Interface web

Unicheck ne permet pas de faire des recherches sur sa base de données de manière gratuite. Ainsi, nous avons obtenu des crédits auprès de l’institut afin d’effectuer quelques tests.

Les recherches peuvent se faire depuis l’interface proposée par Unicheck à l’URL suivant : <https://my.unicheck.com/>. Afin de tester les fonctionnalités, nous avons créé une nouvelle recherche avec le même article du temps.

Une image contenant capture d’écran

Description générée automatiquement

Une image contenant capture d’écran

Description générée automatiquement

L’intégralité de l’article du Temps a été copié et collé dans un nouveau fichier, créé directement via l’interface web de Unicheck.

Une image contenant capture d’écran

Description générée automatiquement

Une image contenant capture d’écran

Description générée automatiquement

Une image contenant capture d’écran

Description générée automatiquement

Après contrôle, nous constatons que Unicheck détecte plus de 96% de similarité. Les 3 sources détectées sont les mêmes qui ont été observées avec les autres outils. Une image contenant capture d’écran

Description générée automatiquement

Unicheck propose également des couleurs en fonction du degré de plagiat du texte. Cela peut être utile pour offrir au journaliste un moyen d’améliorer plus efficacement leurs articles.

Une image contenant capture d’écran

Description générée automatiquement

Afin de tester la précision de l’outil, nous allons également porter l’analyse sur le texte modifié, présent dans le tableau 1.

Une image contenant capture d’écran

Description générée automatiquement

Une image contenant capture d’écran

Description générée automatiquement

Nous constatons que, contrairement à d’autres outils, Unicheck n’est pas en mesure de détecter du texte similaire.

## API

Unicheck offre également un outil API pour faire les recherches. La documentation est disponible ici : <https://unicheck.com/plagiarism-api-documentation>.

Pour faire fonctionner l’API, il faut générer un token via certains identifiants, disponible sous notre profil unicheck.

Une image contenant capture d’écran

Description générée automatiquement

La commande de génération de token est la suivante :

|  |
| --- |
| curl -X POST \  https://api.unicheck.com/oauth/access-token \  -H 'Content-Type: application/x-www-form-urlencoded' \  -d 'grant\_type=<grant\_type>&client\_id=<client\_id>&client\_secret=<client\_secret>' |

Celle-ci est exécutée dans Postman afin d’obtenir le token.

Une image contenant capture d’écran

Description générée automatiquement

Une image contenant capture d’écran

Description générée automatiquement

Nous pouvons utiliser ce token pour uploader un fichier sur Unicheck pour, par la suite, le vérifier.

Une image contenant capture d’écran

Description générée automatiquement

Une image contenant capture d’écran

Description générée automatiquement

Une image contenant capture d’écran, carte

Description générée automatiquement

Ensuite, sur le fichier que nous devons d’uploader sur Unicheck, nous pouvons démarrer le contrôle de plagiat. Pour cela, il est important de préciser dans le « Body » l’identifiant du fichier qui a été uploadé. Dans notre cas, l’ID est « 11be643…f14 ».

Une image contenant capture d’écran

Description générée automatiquement

Une image contenant capture d’écran

Description générée automatiquement

Finalement, pour obtenir le résultat de cette analyse, nous devons exécuter une dernière requête en précisant l’identifiant de la recherche. Dans notre cas, c’est « a4eab…ca3b ».

Une image contenant capture d’écran

Description générée automatiquement

Une image contenant texte, carte

Description générée automatiquementUne image contenant capture d’écran

Description générée automatiquement

Malheureusement, l’API ne nous renvoie pas les URL similaires mais uniquement leur « count ».

# Comparaison

## Taille de l’échantillon

Afin de comparer ces outils, nous avons effectué un test de plagiat avec 63 articles existants. La répartition des articles est comme suit :

* 15 articles payants du Nouvelliste. Langue : français.
* 15 articles gratuits du Nouvelliste. Langue : français.
* 5 articles gratuits du Blick. Langue : allemand.
* 5 articles gratuits de SRF. Langue : allemand.
* 2 articles payants de TheLocal. Langue : anglais.
* 3 articles gratuits de LeNews. Langue : anglais.
* 5 articles gratuits de SwissInfo. Langue : anglais.
* 3 statuts Facebook. Langue : français, allemand, anglais.
* 2 tweets Twitter. Langue : français, anglais.
* 1 article gratuit de SwissInfo. Langue : traduit de l’anglais au français.
* 1 article gratuit de LeNews. Langue : traduit de l’anglais au français.
* 1 article gratuit de SRF. Langue : traduit de l’allemand au français.
* 2 articles gratuits du Blick. Langue : traduits de l’allemand au français.
* 3 articles gratuits du Nouvelliste dont le contenu a été reformulé avec nos propres mots. Langue : français.

Nous excluons l’outil « Plagiarism Checker by EduBirdie » de notre étude. En effet, ce dernier ne possède aucun accès via API, ainsi nous ne pouvons l’intégrer à notre projet.

La taille de l’échantillon s’est limitée à 63 pour des raisons technologiques et financières. La mise à disposition d’un budget pour payer des crédits pour tous les outils n’a pas été jugée nécessaire. Nous estimons que cette taille d’échantillon nous amènera suffisamment de clarté afin de prendre une décision.

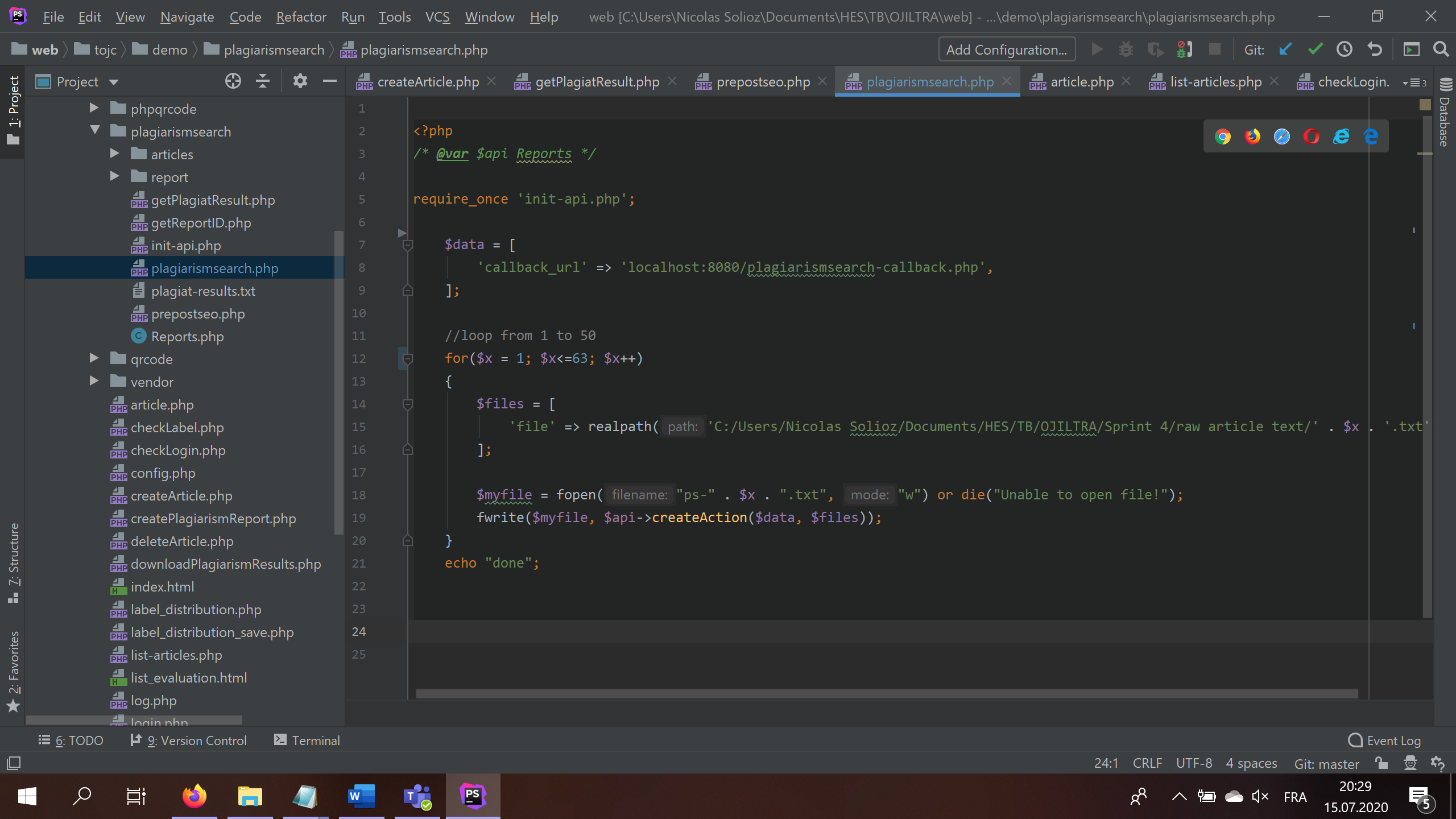
## Méthodologie de test

Afin de comparer l’efficacité de ces outils, nous sommes partis de l’hypothèse suivante : Puisque ces articles existent et sont à disposition en ligne, les outils de détection de plagiat devrait retourner un résultat de « 100% de contenu plagié ». En effet, puisque le texte est reproduit exactement dans la requête, il n’y a pas de raison que notre API ne récupère pas la totalité de l’article. Cette méthode nous permet de déterminer quels outils sont plus efficaces, notamment lorsqu’il s’agit d’accéder à des informations masquées derrière un « paywall » ou contenues dans des réseaux sociaux, potentiellement plus difficiles d’accès.

Nous avons exécuté nos requêtes de différentes manières :

* Pour Copyleaks, Unicheck et Prepostseo, nous avons téléchargé les fichiers « .txt » contenant le texte de nos articles directement depuis l’interface de leurs sites web.
* Pour PlagiarismSearch, nous avons développé une application PHP qui appelle l’API au sein d’une boucle.

Dès que nous avons obtenu nos résultats, nous les avons relevé et stocké dans un fichier Excel afin de les comparer.



Nous avons décidé de ne pas développer d’application PHP pour les applications Prepostseo, Copyleaks et Unicheck car ces derniers possédaient des interfaces web faciles d’utilisation. Nous avons estimé que cette démarche nécessitait moins de temps.

## Résultats de notre étude

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Copyleaks | Unicheck | PlagiarismSearch | Prepostseo |
| Moyenne avec Paywall | 26.91% | 26.49% | 25.51% | 21.47% |
| Moyenne sans Paywall | 97.61% | 97.97% | 98.83% | 96.79% |
| Moyenne anglais | 81.63% | 79.96% | 79.46% | 84.50% |
| Moyenne français | 62.84% | 62.89% | 63.74% | 58.67% |
| Moyenne allemand | 97.71% | 99.72% | 98.84% | 95.40% |
| Moyenne anglais avec Paywall | 8.75% | 1.49% | 0.00% | 25.50% |
| Moyenne anglais sans Paywall | 99.85% | 99.58% | 99.32% | 99.25% |
| Moyenne français avec Paywall | 29.33% | 29.83% | 28.91% | 20.93% |
| Moyenne français sans Paywall | 96.35% | 95.95% | 98.56% | 96.40% |
| Moyenne allemand avec Paywall | - | - | - | - |
| Moyenne allemand sans Paywall | 97.71% | 99.72% | 98.84% | 95.40% |
| Moyenne réseaux sociaux | 75.80% | 46% | 80% | 54% |
| Moyenne texte traduit | 7.32% | 0.00% | 2.66% | 4.20% |
| Moyenne texte reformulé | 14% | 0% | 10% | 12% |
| Prix / 100 requêtes | 10.99 | 15 | 9.983333333 | 0.2 |
| Facilité mise en place | complexe | moyenne | moyenne | moyenne |
| Moyenne totale | 65.63% | 62.15% | 65.70% | 61.70% |
| Score final | 4 | 4 | **5** | 4 |

# Conclusion et choix

Notre choix se porte sur l’outil « PlagiarismSearch». Le prix des requêtes proposé par cet outil est considérablement plus bas que les autres. De plus, la mise en place de l’API a été bien plus simple. PlagiarismSearch offre un résultat plus précis, surtout quant il s’agit de rechercher un texte reformulé. C’est pourquoi nous estimons que cet outil est le plus à même de répondre aux besoins de nos clients, les journalistes étant dans l’habitude de reformuler des informations obtenues par des tiers (ex : communiqué de presse).