

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
Prepostseo.....	3
Interface web	3
API	5
Plagiarism Checker by EduBirdie	7
Interface web	7
API	8
Plagiarismsearch.....	8
Interface web	8
API	10
Copyleaks.....	13
Interface web	13
API	15
Unicheck.....	17
Interface web	17
API	21
Comparaison	27
Conclusion et choix	29

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 analysée via différents outils existants. Cet état de l'art analysera les solutions suivantes :

- Plagiarism Checker by EduBirdie
- plagiarismsearch
- copyleaks
- unichack
- 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.

https://www.prepostseo.com/plagiarism-checker

OPEN

Migros abandonne la course à la grandeur
Commerce de détail
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
Un logo Migros à Zurich, le 17 février 2018. — © KEYSTONE/Melanie Duchene
Servan Peca
Publié jeudi 27 juin 2019 à 19:48 Modifié jeudi 27 juin 2019 à 21:34

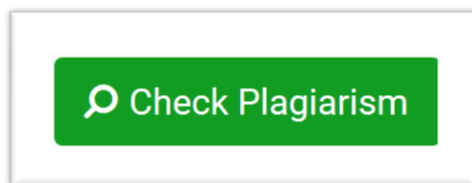
port à bug

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.

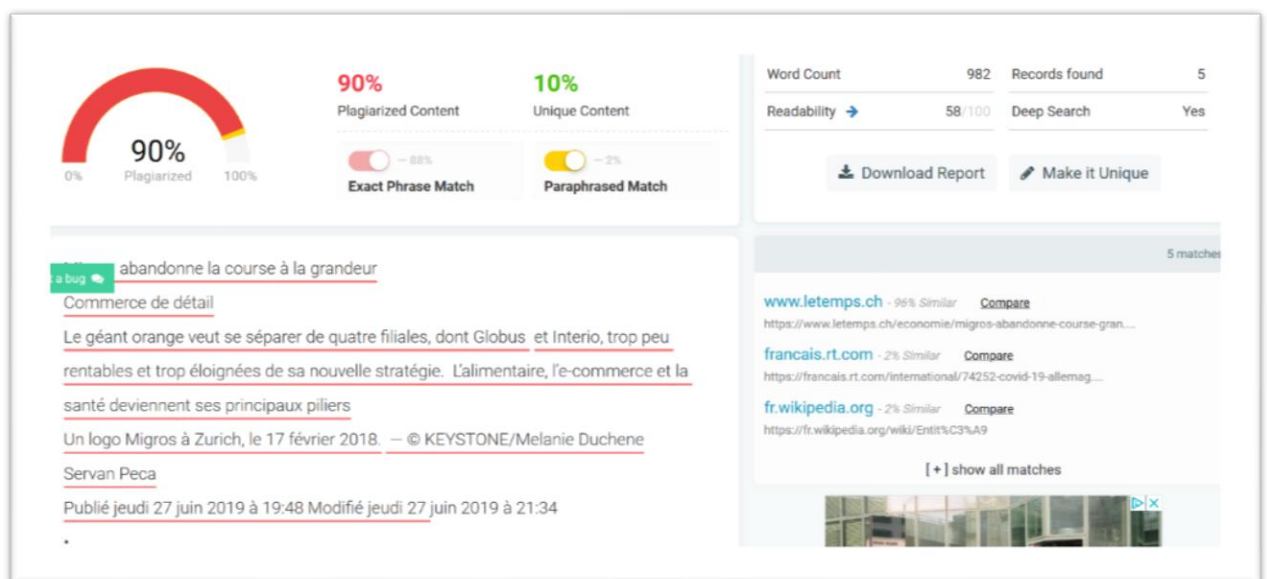
Select File .doc / .docx / .pdf / .txt Words: 1,358

Exclude Url Language: English

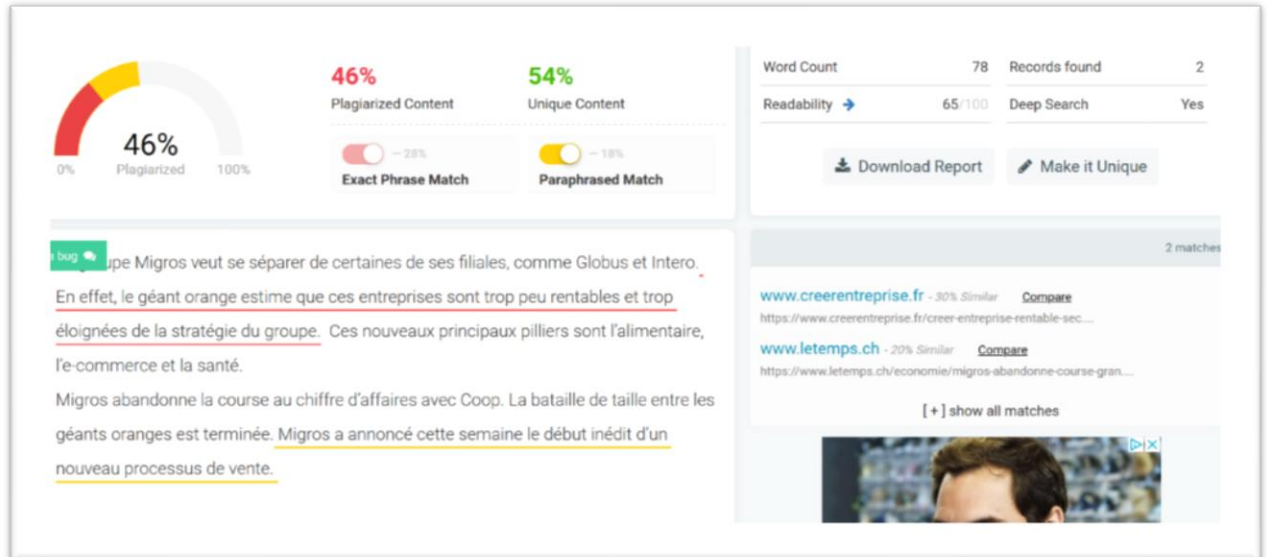
ADVERTISEMENT



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



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.



Nous constatons que la page web de source, www.letemps.ch, a été retrouvée malgré la reformulation des phrases.

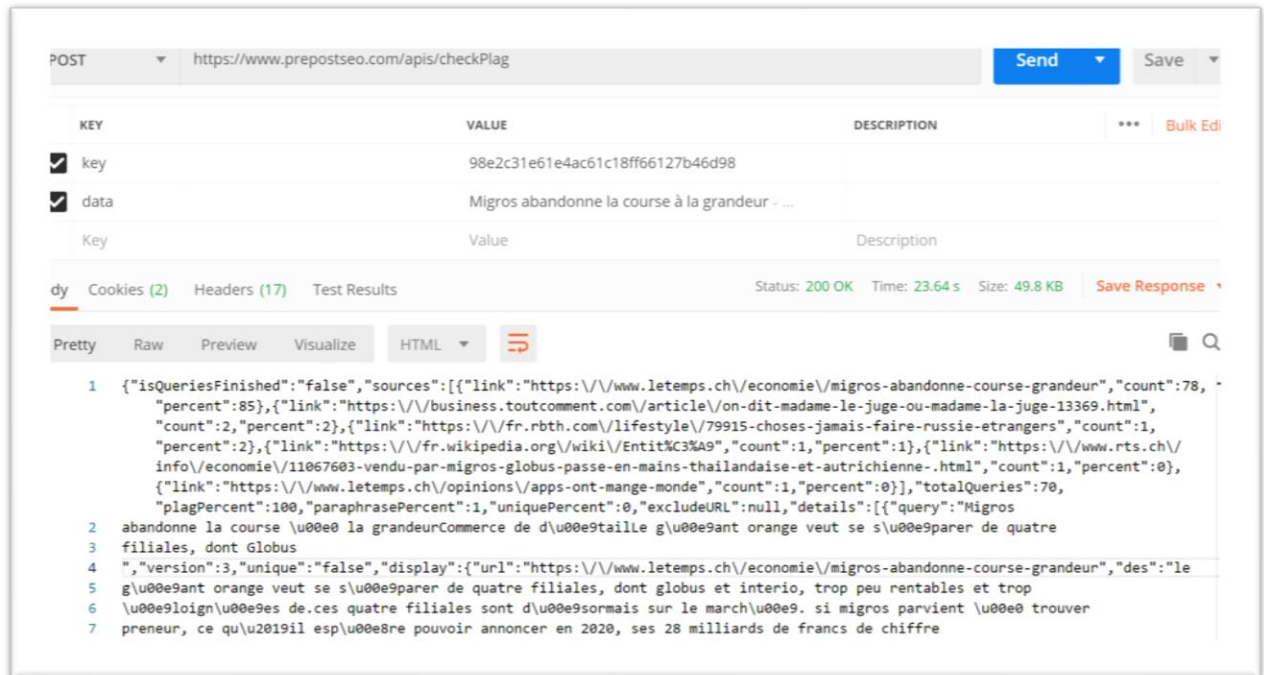
API

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.

```
curl -X POST https://www.prepostseo.com/apis/checkSentence \  
-d "key=YOUR_KEY"  
-d "query=Inside that cage there was a green teddy bear"
```

```
{  
  "query": "Inside that cage there was a green teddy bear",  
  "totalMatches": "4",  
  "unique": "false",  
  "webs": [  
    {  
      "title": "Best university and college",  
      "des": "On the red table, there was a purple curtain. Underneath that was a silve",  
      "url": "http://bestuniversityandcollage.blogspot.com/"  
    },  
    {  
      "title": "clcservicesblog | Best Debt Management Company",  
      "des": "clcservicesblog Best Debt Management Company. Menu Skip to content. ... Unc",  
      "url": "https://clcservicesblog.wordpress.com/"  
    },  
    {  
      "title": "CLC Services : Debt Management | clcservicesblog",  
      "des": "CLC Services : Debt Management. ... Inside that cage there was a green te",  
      "url": "https://clcservicesblog.wordpress.com/2014/11/01/clc-services-debt-management/"  
    },  
    {  
      "title": "ONLINE EDUCATION & USA INSURANCE UPDATES",  
      "des": "Inside that cage there was a green teddy bear, with the number 43 written",  
      "url": "http://eduusainsu.blogspot.com/"  
    }  
  ]  
}
```

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



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, EduBirdie n'était pas en mesure de détecter la source de l'article.



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.

☒ **letemps.ch**

... jeudi 27 juin 2019 à 21:34 Si tant est qu'elle existait réellement, Migros renonce à gagner la course au chiffre d'affaires avec Coop. Finie la lutte de taille ...

<https://www.letemps.ch/economie/migros-abandonne-course-grandeur>

Plagiarism: 100.00 %

☒ **monchange.ch**

... Le Temps Pas de commentaire 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 ... 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 ... , le géant orange a acquis en 2015 la majorité de l'une des entreprises les plus ... 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 ... A l'image d'un Nestlé, d'un ABB ou, de manière plus discrète, d'une SGS, le groupe ...

<https://monchange.ch/migros-abandonne-la-course-a-la-grandeur/>

Plagiarism: 100.00 %

☒ **pressreader.com**

Jun 28, 2019 - 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 ...

<https://www.pressreader.com/switzerland/le-temps/20190628/281513637693928>

Plagiarism: 100.00 %

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
<p>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.</p> <p>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.</p> <p>Globus, Interio, Gries Deco (Depot) et M-way. Ces quatre filiales sont désormais sur le marché. Si Migros parvient à trouver preneur, ce</p>	<p>Le groupe Migros veut se séparer de certaines de ses filiales, comme Globus et Interio. En effet, le géant orange estime que ces entreprises sont trop peu rentables et trop éloignées de la stratégie du groupe. Ces nouveaux principaux piliers sont l'alimentaire, l'e-commerce et la santé.</p> <p>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.</p>

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.	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.

letemps.ch

Plagiarism: 80.00 %

... -clés Fil d'Ariane Accueil Economie **Migros abandonne la course** à la grandeur Publicité **Migros abandonne la course** à la grandeur ... 'elle existait réellement, **Migros** renonce à gagner **la course au chiffre d'affaires avec Coop**. Finie **la lutte de taille** ...
<https://www.letemps.ch/economie/migros-abandonne-course-grandeur>

monchange.ch

Plagiarism: 80.00 %

... Bitcoin Taux de Change Contact **Migros abandonne la course** à la grandeur Daniel Seo 27 ... 'elle existait réellement, **Migros** renonce à gagner **la course au chiffre d'affaires avec Coop**. Finie **la lutte de taille** ...
<https://monchange.ch/migros-abandonne-la-course-a-la-grandeur/>

pressreader.com

Plagiarism: 66.67 %

Jun 28, 2019 - 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** ...
<https://www.pressreader.com/switzerland/le-temps/20190628/281513637693928>

agefi.com

Plagiarism: 42.86 %

... Suppléments et Magazines Découvrir Retrouvez **les suppléments et magazines** ... **en vente des filiales dont Globus, Interio, m-way et Gries Deco**. MH **Migros** cherche pour ses **filiales Globus, Interio, m-way et le groupe Gries Deco** ... avec **Migros** et ne se sont pas ...
<https://www.agefi.com/home/entreprises/detail/edition/online/article/le-groupe-migros-focalise-son-orientation-strategique-et-lance-un-processus-dadaptation-de-son-portefeuille-le-geant-orange-met-en-vente-des-filiales-dont-globus-interio-m-way-et-gries-deco-488774.html>

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.



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.

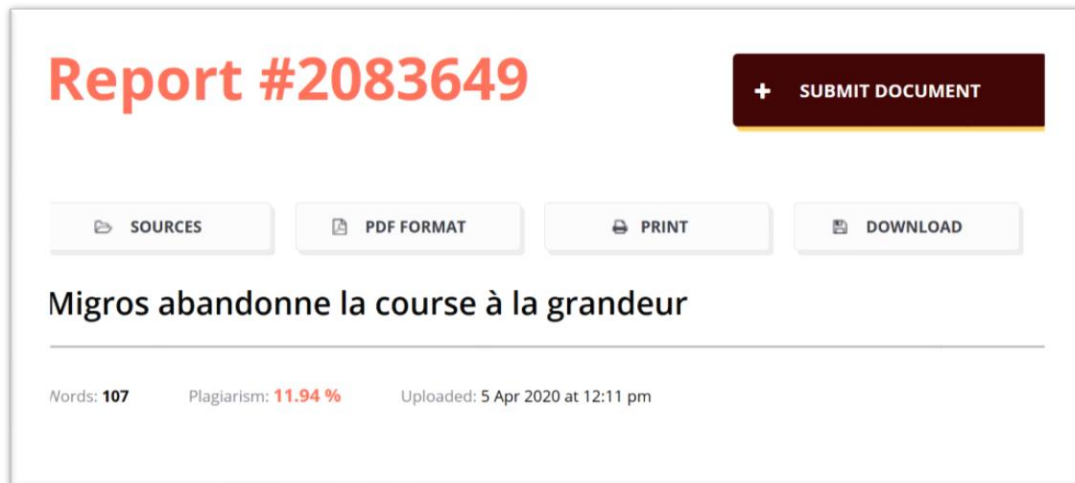
Ensuite pour exécuter une requête, nous devons insérer nos identifiants de connexions dans le fichier init.api.php. Les identifiants sont disponibles sur le profil plagiarismsearch, directement sur le site.

A screenshot of a web form for configuring the Plagiarismsearch API. It has three input fields: 'url' with the value 'https://plagiarismsearch.com/api/v3', 'Jser' with the value 'nicolas.solioz@students.hevs.ch', and 'key' with the value 'q1184819pjzkrnvaaccabab-71324088'.

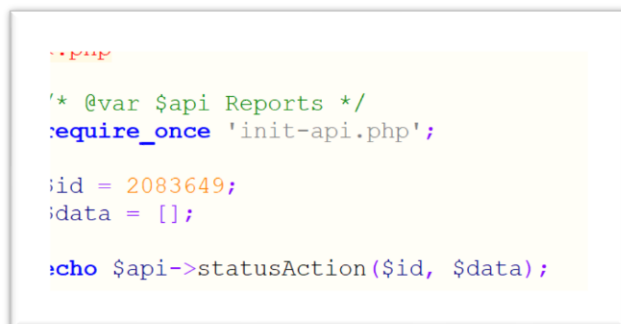
Ces informations sont mis de cette manière dans le fichier init-api.php :

```
<code><pre><code>
```

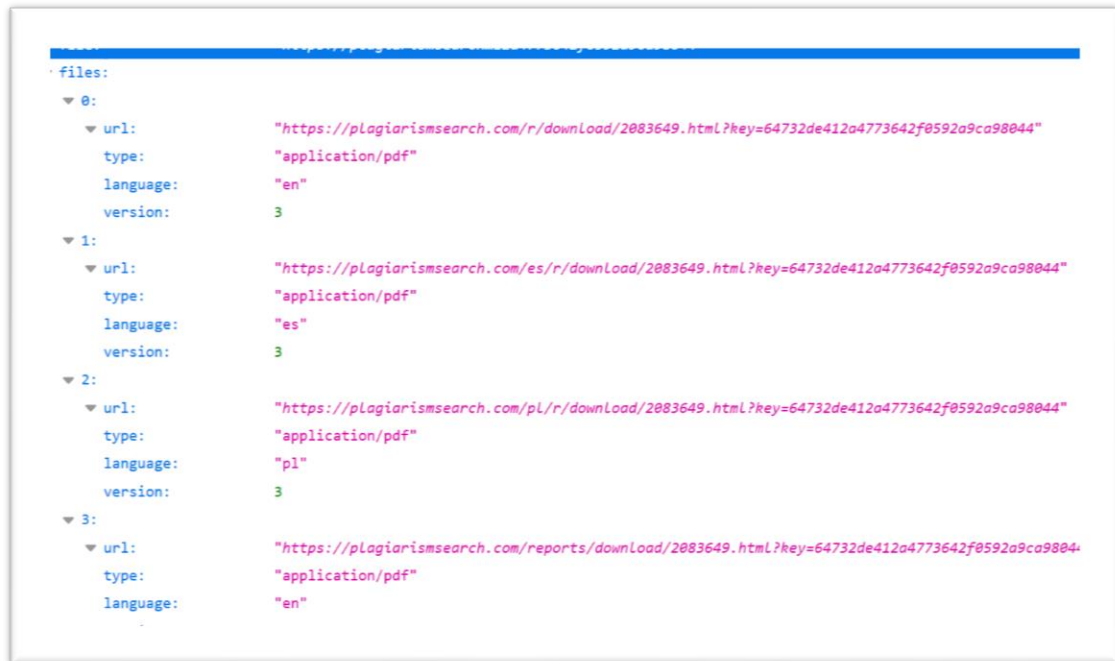
Finalement, pour tester l'API nous souhaitons récupérer les informations de notre recherche déjà effectuée. Pour se faire, nous remplaçant l'identifiant du « report » dans le fichier status.php. L'identifiant du « report » est indiqué au sommet de la page de résultat de notre recherche, directement sur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



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



Une fois toutes 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>.



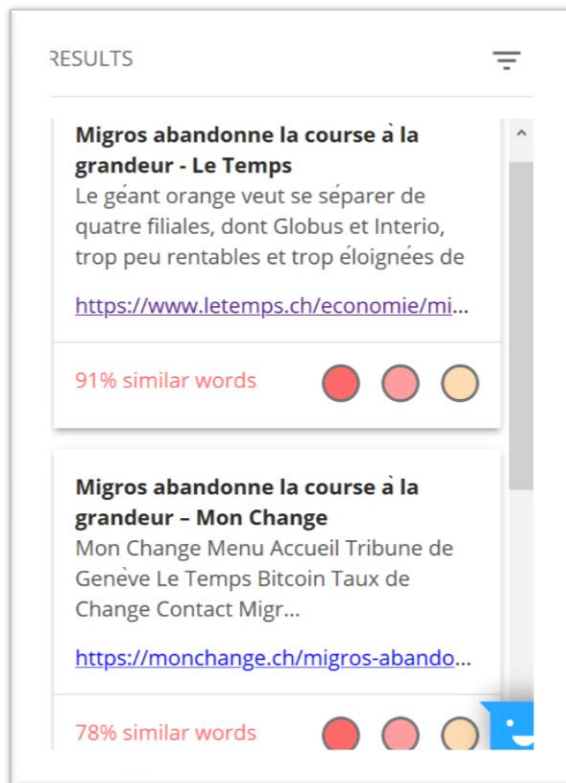
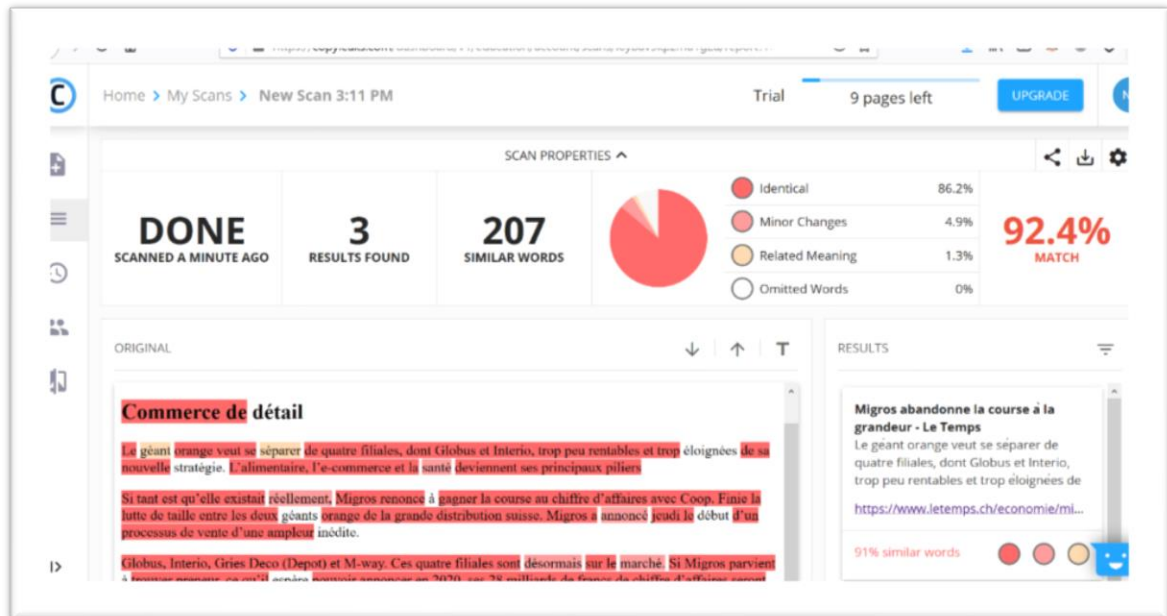
L'API est fonctionnelle. Nous pouvons accéder au PDF de notre « report » en plusieurs langues.

Plagiarismsearch offre 100 requêtes. Toutes requêtes supplémentaires sont payantes.

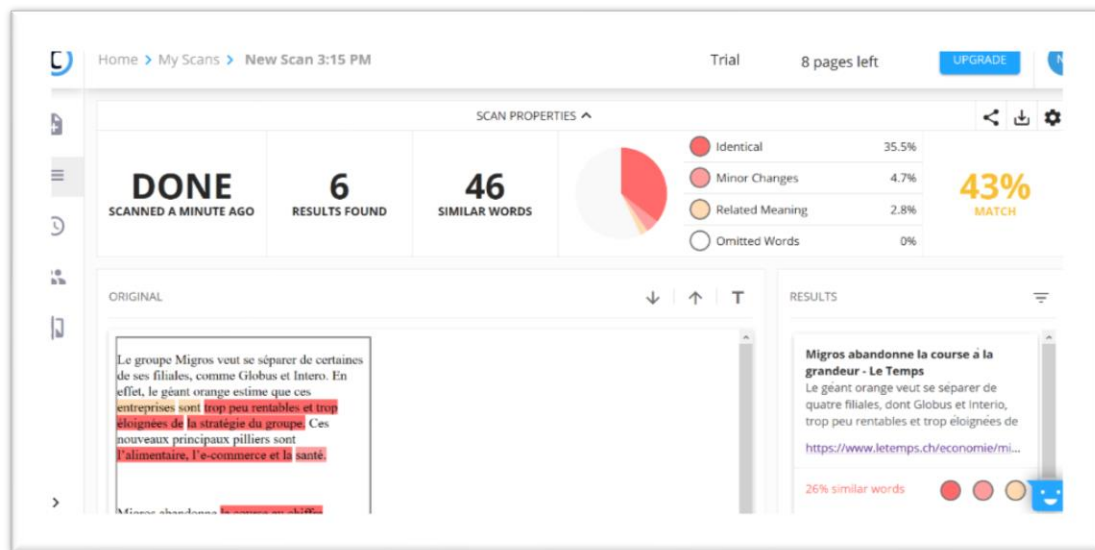
Copyleaks

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.

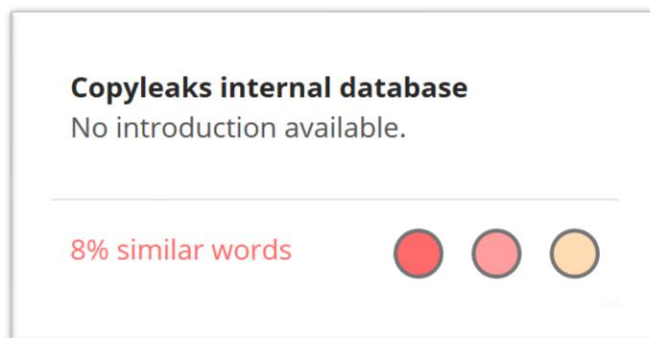


Nous constatons que les résultats sont similaires à ceux de plagiarismsearch. L'interface de Copleaks est pourtant plus développée. Un système de code couleur permet de voir si les termes sont copiés exactement ou paraphrasés. Si nous faisons le même essai avec le texte modifié (voir tableau 1), voici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



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 déceler 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.



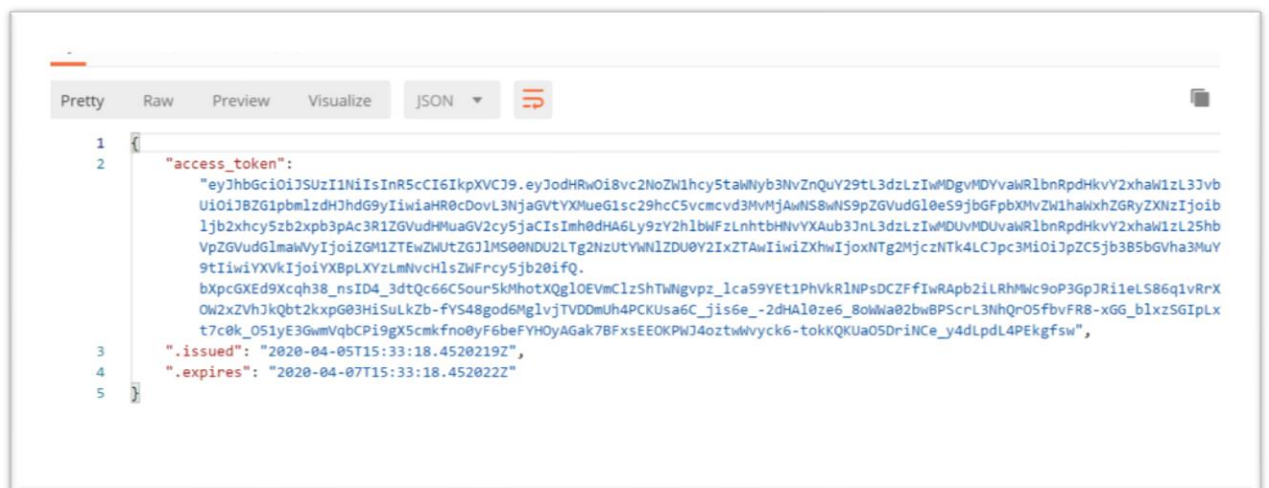
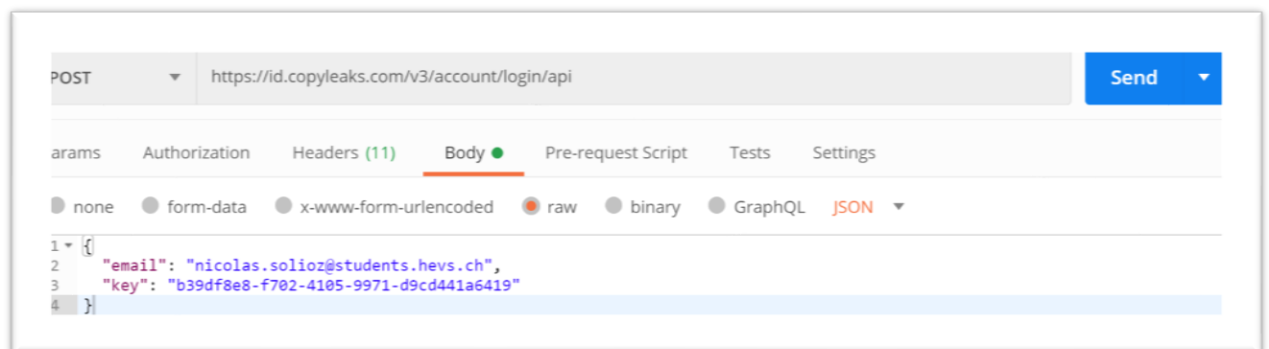
API

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.

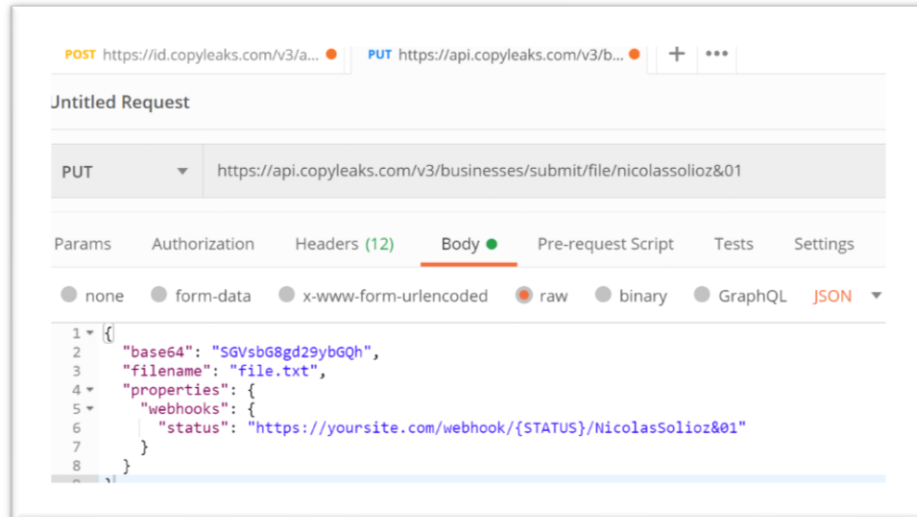


Dans un premier temps, nous devons obtenir un token d'identification. Ce token s'obtient via une méthode POST comprenant notre adresse e-mail et notre clé de profil.



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 ».



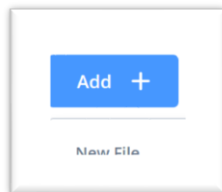
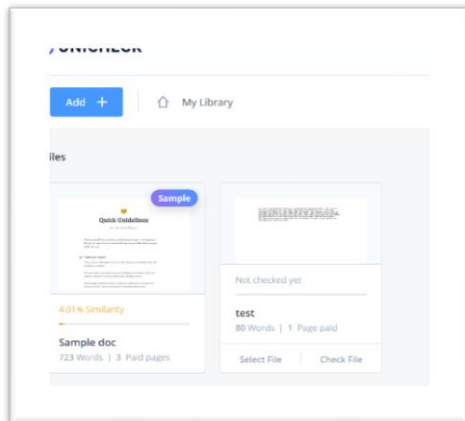
Cette propriété est nécessaire pour que la requête fonctionne. Nous devons ainsi configurer notre serveur afin qu'il puisse recevoir le statut de notre requête API. Cela nous semble être une étape supplémentaire encombrante, les autres outils sont plus simples.

Unicheck

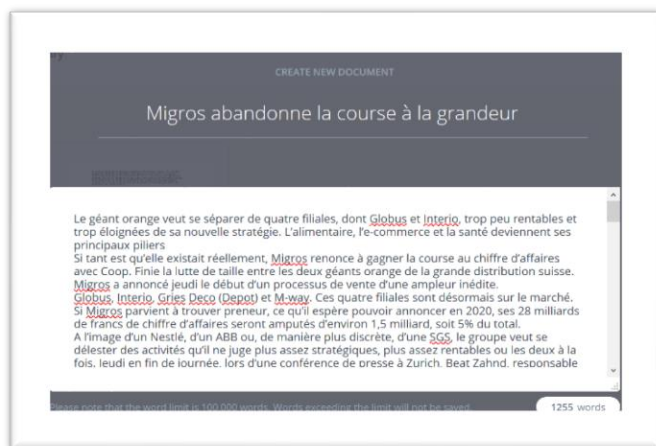
Interface web


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.



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





Not checked yet


Migros abandonne la co...
1 260 Words | 5 Paid pages

Select File | Check File

Confirm the check

You have chosen 1 document

SORTED BY PAID PAGES

01	Migros abandonne la course à la grandeur 1 260 Words 5 paid pages	
----	---	---

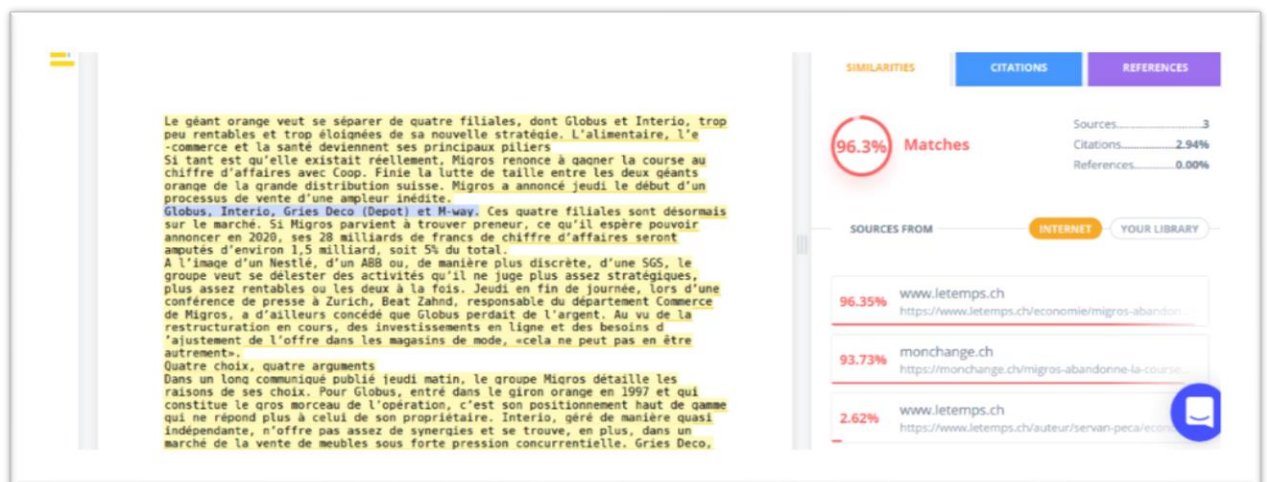
1 Document | 5 Paid pages | 95 Pages remain

Start check

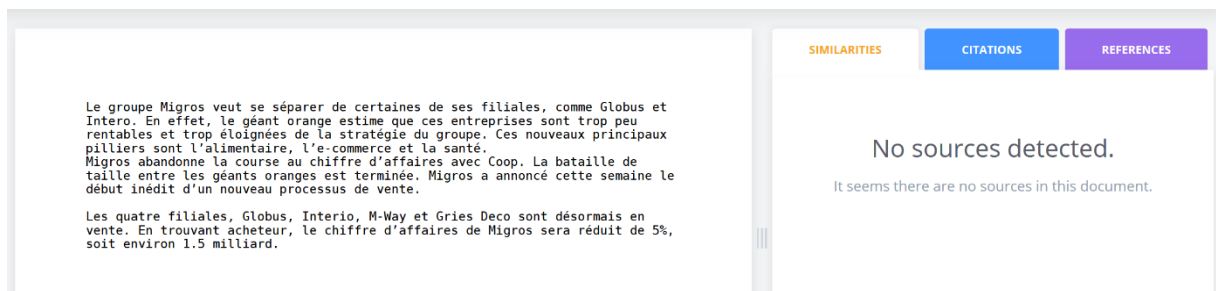
Après contrôle, nous constatons que Unicheck détecte plus de 96% de similarité. Les 3 sources détectées sont les mêmes qui ont été observées avec les autres outils.



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.



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

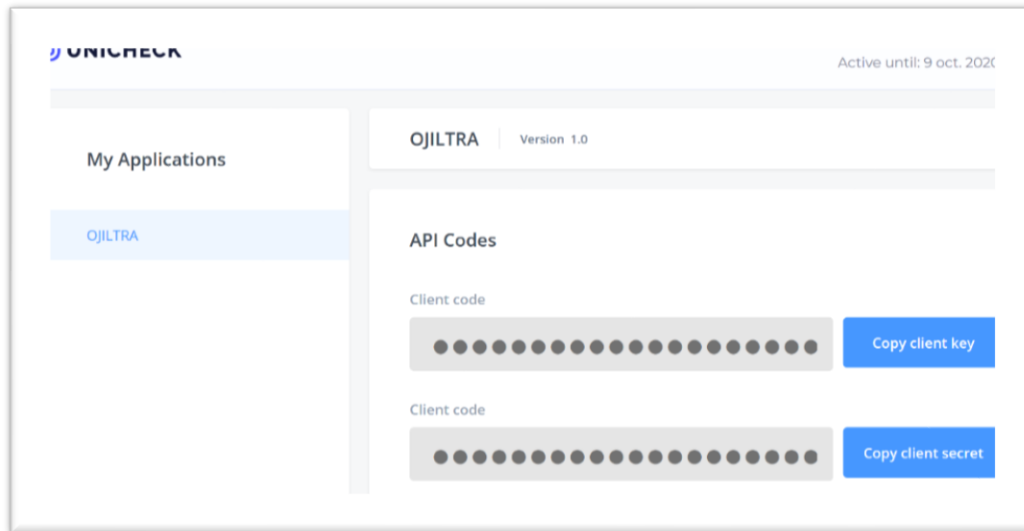


Nous constatons que, contrairement à d'autres outils, Unicheck n'est pas en mesure de détecter du texte similaire.

API

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

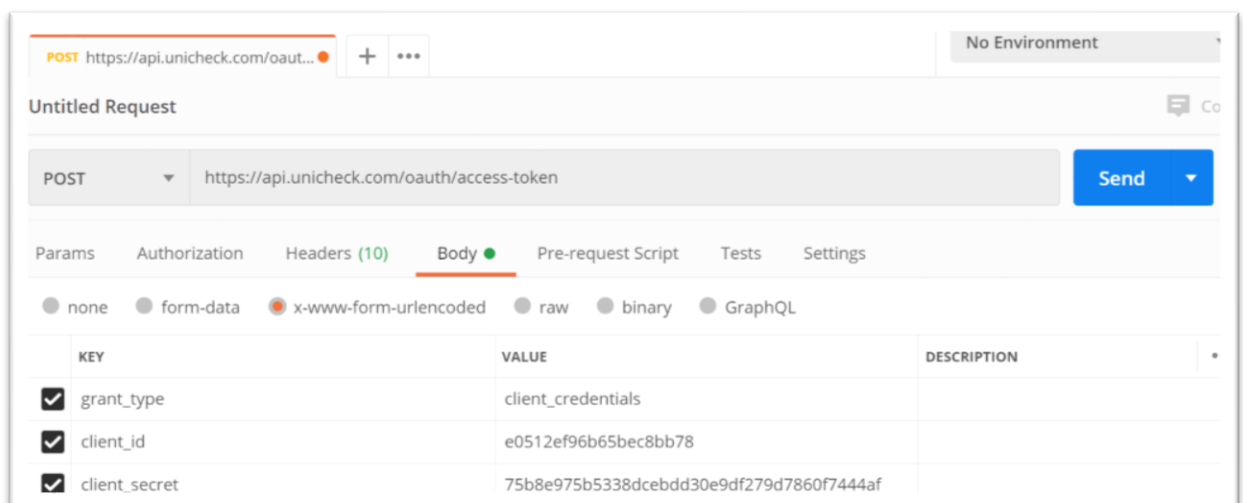
Pour faire fonctionner l'API, il faut générer un token via certains identifiants, disponible sous notre profil unicheck.



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.



```
{
  "token_type": "Bearer",
  "expires_in": 2592000,
  "access_token":
    "eyJ0eXAiOiJKV1QiLCJhbGciOiJSUzI1NiIsImp0aSI6IjEjYnZmM2MTciLCJpYXQiOiE1ODcwNDAwNjQsIm5iZiI6MTU4NzA0MDA2NCwiZXhwIjozNTg5NjMyMDY0LCJzdWIiOiIzMjY2NTkiLCJzY29wZXMiOiJtdFQuu2HH3DGBuWnHOyvazszMdBf3x8YaTfbilbewYrwtH7bGAvlMIbKK2WhldEejVguXVqc7KR371VgfxGYMVObjBC_mjPpK9tDyBjNfwkmpaEH930ztGubA_kgsW4DiXbUyqka19TaXBtXzrpTfs0JWSoloJztjwg_SMGLJrzYLods"
}
```

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

Untitled Request

POST https://api.unicheck.com/files Send

Params Authorization Headers (12) Body Pre-request Script Tests Settings

Headers 9 hidden

KEY	VALUE	DESCRIPTION
<input checked="" type="checkbox"/> Accept	application/vnd.api+json	
<input checked="" type="checkbox"/> Authorization	Bearer eyJ0eXAiOiJKV1QiLCJhbGciOiJSUzI1NiIsImp0aSI6IjEjYnZmM2MTciLCJpYXQiOiE1ODcwNDAwNjQsIm5iZiI6MTU4NzA0MDA2NCwiZXhwIjozNTg5NjMyMDY0LCJzdWIiOiIzMjY2NTkiLCJzY29wZXMiOiJtdFQuu2HH3DGBuWnHOyvazszMdBf3x8YaTfbilbewYrwtH7bGAvlMIbKK2WhldEejVguXVqc7KR371VgfxGYMVObjBC_mjPpK9tDyBjNfwkmpaEH930ztGubA_kgsW4DiXbUyqka19TaXBtXzrpTfs0JWSoloJztjwg_SMGLJrzYLods	
<input checked="" type="checkbox"/> Content-Type	multipart/form-data	
Key	Value	Description

Response

Untitled Request

POST https://api.unicheck.com/files Send

Params Authorization Headers (12) Body Pre-request Script Tests Settings

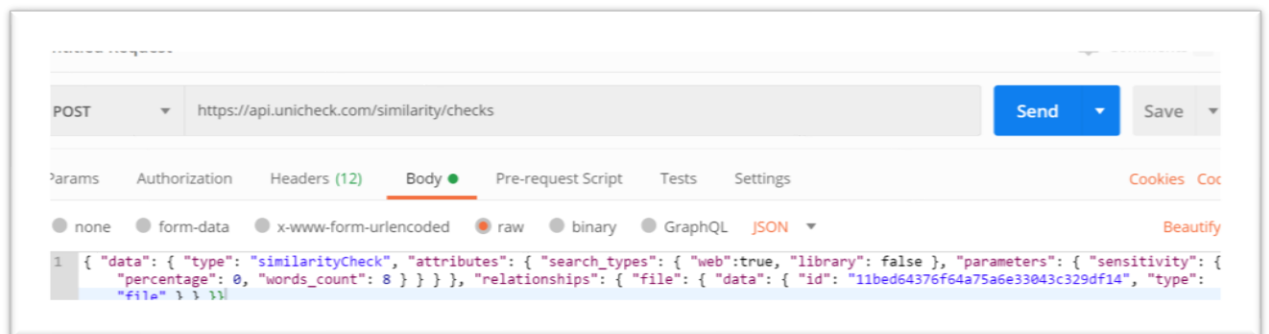
☐ none
 ☒ form-data
 ☐ x-www-form-urlencoded
 ☐ raw
 ☐ binary
 ☐ GraphQL

KEY	VALUE	DESCRIPTION
<input checked="" type="checkbox"/> file	abc.pdf ×	
Key	Value	Description

Response

```
1 {  
2   "jsonapi": {  
3     "version": "1.0"  
4   },  
5   "data": {  
6     "type": "file",  
7     "id": "11bed64376f64a75a6e33043c329df14",  
8     "attributes": {  
9       "name": "abc (1)",  
10      "extension": "pdf",  
11      "state": "prepared",  
12      "words_count": 0,  
13      "pages_count": 0,  
14      "size": 0,  
15      "preview": null,  
16      "created_at": "2020-04-21T13:05:43+00:00",  
17      "updated_at": "2020-04-21T13:05:43+00:00"  
18    },  
19    "links": {  
20      "self": "https://api.unicheck.com/files/11bed64376f64a75a6e33043c329df14"  
21    }  
2  }  
3 }
```

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 ».

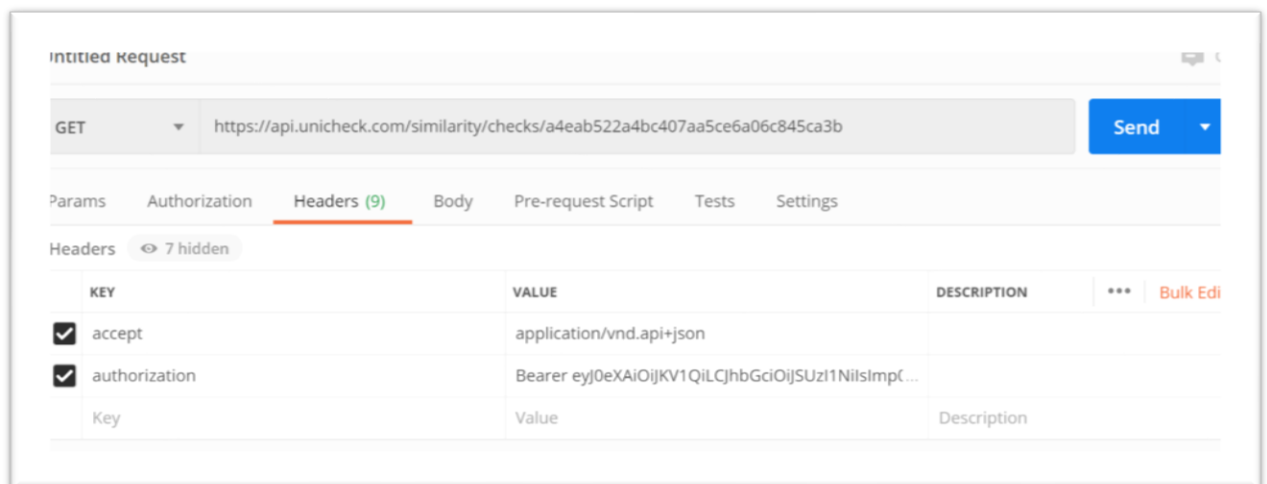



```

1  {
2    "jsonapi": {
3      "version": "1.0"
4    },
5    "data": {
6      "type": "similarity_check",
7      "id": "a4eab522a4bc407aa5ce6a06c845ca3b",
8      "attributes": {
9        "state": "created",
10       "similarity": 0,
11       "search_types": [
12         "web"
13       ],
14       "parameters": {
15         "sensitivity": {
16           "percentage": 0,
17           "words_count": 8
18         }
19       },
20       "checked_at": null,
21       "created_at": "2020-04-21T13:08:36+00:00",
22       "updated_at": "2020-04-21T13:08:36+00:00"
23     }
24   }

```

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 ».



```
1  {
2    "jsonapi": {
3      "version": "1.0"
4    },
5    "data": {
6      "type": "similarity_check",
7      "id": "a4eab522a4bc407aa5ce6a06c845ca3b",
8      "attributes": {
9        "state": "built",
10       "similarity": 96.95000000000003,
11       "search_types": [
12         "web"
13       ],
14       "parameters": {
15         "sensitivity": {
16           "percentage": 0,
17           "words_count": 8
18         }
19       },
20       "checked_at": "2020-04-21T13:09:04+00:00",
21       "created_at": "2020-04-21T13:08:36+00:00",
22       "updated_at": "2020-04-21T13:08:36+00:00"
23     },
24   },
25   "links": {
26     "self": "https://api.unicheck.com/similarity/checks/a4eab522a4bc407aa5ce6a06c845ca3b"
27   },
28   "meta": {
29     "exclude": {
30       "citations": true,
31       "references": true,
32       "citations_percentage": 2.6899999999999999,
33       "references_percentage": 0,
34       "source_count": 0,
35       "citations_count": 0,
36       "references_count": 0,
37       "total_percentage": 2.6899999999999999
38     },
39     "originality": {
40       "sources_count": 7
41     },
42     "cost": {
43       "value": 6,
44       "resource_type": "page"
45     }
46   }
47 }
```

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.

```
<?php
/* @var $api Reports */
require_once 'init-api.php';

$data = [
    'callback_url' => 'localhost:8080/plagiarismsearch-callback.php',
];

//loop from 1 to 50
for($x = 1; $x<=50; $x++)
{
    $files = [
        'file' => realpath( 'path: 'C:/Users/Nicolas Solioz/Documents/HES/TB/03ILTRA/Sprint 4/raw article text/' . $x . '.txt'
    );
    $myfile = fopen( 'filename: "ps-" . $x . ".txt", 'mode: "w" ) or die("Unable to open file!");
    fwrite($myfile, $api->createAction($data, $files));
}
echo "done";
```

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).