

Traitement automatique du langage naturel

Traitements automatiques du langage naturel

Pré-traitement

Segmentation (*tokenization*)

Lemmatisation (*lemmatization*)

Racinalisation (*stemming*)

Étiquetage morpho-syntaxique (*part-of-speech tagging*)

Notions

Mots vides (*stopwords*)

Entités nommées (*named entities*)

Techniques complémentaires

Analyse de sentiment (*sentiment analysis*)

Modélisation thématique (*topic modeling*)

Reconnaissance optique des caractères (OCR - « océrisation »)

Traitement automatique du langage naturel

Segmentation (*tokenization*)

Découpage d'un texte en **unités lexicales** qui correspondent généralement aux mots.



Tournoi de badminton universitaire : l'UdeM domine, l'UQAM repart avec le bronze

Tristan Champagne-Lessard | 21 février 2020

Traitement automatique du langage naturel

Segmentation (*tokenization*)

```
texte =  
"Tournoi de badminton universitaire : l'UdeM  
domine, l'UQAM repart avec le bronze après six  
heures de compétition. Un texte de Claudine Giroux  
et de Tristan Champagne-Lessard."
```

```
<type 'str'>
```

Traitement automatique du langage naturel

Segmentation (*tokenization*)

```
tokens =  
['Tournoi', 'de', 'badminton', 'universitaire',  
':', "l'", 'UdeM', 'domine', ',', 'l'', 'UQAM',  
'repart', 'avec', 'le', 'bronze', 'après', 'six',  
'heures', 'de', 'compétition', '.', 'Un', 'texte',  
'de', 'Claudine', 'Giroux', 'et', 'de', 'Tristan',  
'Champagne', '-', 'Lessard', '.']  
  
<type 'list'>
```

Traitements automatiques du langage naturel

Lemmatisation

Regroupement sous une **forme canonique** (qui correspond en général à celles qu'on retrouve dans un dictionnaire) des occurrences du texte.

En français, la lemmatisation va généralement consister à effectuer les opérations suivantes :

Transformer tous les **verbes** conjugués à l'**infinitif**.

Ramener tous les **noms** au **singulier**.

Réduire tous les **adjectifs** au **mASCulin singulier**.

Changer les **élisions** en une forme sans élision.

Traitement automatique du langage naturel

Lemmatisation

lemmes =

```
[ 'tournoi', 'de', 'badminton', 'universitaire',  
  ':', 'le', 'UdeM', 'domine', ',', 'l', 'UQAM',  
 'repartir', 'avec', 'le', 'bronze', 'après',  
 'six', 'heure', 'de', 'compétition', '.', 'un',  
 'texte', 'de', 'Claudine', 'Giroux', 'et', 'de',  
 'Tristan', 'champagne', '-', 'Lessard', '..']
```

Traitement automatique du langage naturel

Lemmatisation

lemmes =

```
[ 'tournoi', 'de', 'badminton', 'universitaire',  
  ':', 'le', 'UdeM', 'domine', ',', 'l', 'UQAM',  
  'repartir', 'avec', 'le', 'bronze', 'après',  
  'six', 'heure', 'de', 'compétition', '.', 'un',  
  'texte', 'de', 'Claudine', 'Giroux', 'et', 'de',  
  'Tristan', 'champagne', '-', 'Lessard', '..']
```

Traitement automatique du langage naturel

Lemmatisation

Pratique pour regrouper des mots/idées semblables.

Mais la lemmatisation n'est pas toujours nécessaire ou requise.

Par exemple, si vous vous intéressez aux relations de pouvoir, les verbes conjugués « peux », « pourra » ou « pu » seront lemmatisés à « pouvoir » et il est possible que vous confondiez le verbe avec le nom « pouvoir ».

Peut produire des erreurs : le nom « élue » confondu avec « élire »...

Elle doit donc être justifiée.

Traitement automatique du langage naturel

Racinement (stemming)

racines =

```
[ 'tournoi', 'de', 'badminton', 'universitair',  
':', 'le', 'udem', 'domin', ',', 'l', 'uqam',  
'repart', 'avec', 'le', 'bronz', 'apres', 'six',  
'heur', 'de', 'compétit', '.', 'un', 'text', 'de',  
'claudin', 'giroux', 'et', 'de', 'tristan',  
'champagn', '-', 'lessard', '..']
```

Plus efficace pour certains cas où on veut regrouper mots de même famille : « Palestine » et « Palestiniens », par exemple.

Plus destructif. Bien peser le pour et le contre avant d'y recourir.

Traitement automatique du langage naturel

Étiquettagé morpho-syntaxique (*part-of-speech tagging*)

```
{ 'Tournoi' : 'NOUN__Gender=Masc|Number=Sing' }
{ 'de' : 'ADP__' }
{ 'badminton' : 'NOUN__Gender=Masc|Number=Sing' }
{ 'universitaire' : 'ADJ__' }
{ ':' : 'PUNCT__' }
{"l'" : 'DET__Definite=Def|Number=Sing|
PronType=Art' }
{ 'UdeM' : 'PROPN__Number=Sing' }
{ 'domine' : 'NOUN__Gender=Fem|Number=Sing' }
{ ',' : 'PUNCT__' }
{ '1' : 'NUM__NumType=Card' }
{ 'UQAM' : 'PROPN__' }
```

Traitement automatique du langage naturel

Étiquettagé morpho-syntaxique (*part-of-speech tagging*)

```
{ 'UQAM': 'PROPN____' }
{ 'repart': 'VERB__Gender=Masc | Number=Sing |
Tense=Past | VerbForm=Part' }
{ 'avec': 'ADP____' }
{ 'le': 'DET__Definite=Def | Gender=Masc | Number=Sing |
PronType=Art' }
{ 'bronze': 'NOUN__Gender=Masc | Number=Sing' }
{ 'après': 'ADP____' }
{ 'six': 'NUM__NumType=Card' }
{ 'heures': 'NOUN__Gender=Fem | Number=Plur' }
{ 'de': 'ADP____' }
{ 'compétition': 'NOUN__Gender=Fem | Number=Sing' }
```

Traitement automatique du langage naturel

Étiquettagé morpho-syntaxique (*part-of-speech tagging*)

```
{ 'Claudine' : 'PROPN__Gender=Fem|Number=Sing' }
{ 'Giroux' : 'PROPN__' }
{ 'et' : 'CCONJ__' }
{ 'de' : 'ADP__' }
{ 'Tristan' : 'PROPN__' }
{ 'Champagne' : 'ADJ__Number=Sing' }
{ '-' : 'PUNCT__' }
{ 'Lessard' : 'PROPN__' }
{ '.' : 'PUNCT__' }
```

Traitement automatique du langage naturel

Découpage en phrases

Phrases =

```
[ 'Tournoi de badminton universitaire : ', "l'UdeM  
domine, l'UQAM repart avec le bronze après six  
heures de compétition.", 'Un texte de Claudine  
Giroux et de Tristan Champagne', '-Lessard.' ]
```

Traitement automatique du langage naturel

Mots vides (*stopwords*)

```
tokens =  
['Tournoi', 'de', 'badminton', 'universitaire',  
':', "l'", 'UdeM', 'domine', ',', 'l'', 'UQAM',  
'repart', 'avec', 'le', 'bronze', 'après', 'six',  
'heures', 'de', 'compétition', '.', 'Un', 'texte',  
'de', 'Claudine', 'Giroux', 'et', 'de', 'Tristan',  
'Champagne', '-', 'Lessard', '.']
```

```
len(tokens) = 33
```

Traitement automatique du langage naturel

Mots vides (*stopwords*)

```
tokens =  
[ 'Tournoi',          'badminton',   'universitaire',  
':',                 'UdeM',        'domine',      ',',      'UQAM',  
'repart',            'bronze',       'texte',  
'heures',            'compétition',  ':',      'Tristan',  
'Claudine',          'Giroux',       '-.',  
'Champagne',         'Lessard',     '.']
```

len(tokens) = 21

Traitement automatique du langage naturel

Mots vides (*stopwords*)

```
tokens =  
[ 'Tournoi', 'badminton', 'universitaire', ':',  
'UdeM', 'domine', ',', 'UQAM', 'repart', 'bronze',  
'heures', 'compétition', '.', 'texte', 'Claudine',  
'Giroux', 'Tristan', 'Champagne', '-', 'Lessard',  
. ]
```

```
len(tokens) = 21
```

Traitement automatique du langage naturel

Mots vides (*stopwords*)

```
tokens =  
[ 'Tournoi', 'badminton', 'universitaire', 'UdeM',  
'domine', 'UQAM', 'repart', 'bronze', 'heures',  
'compétition', 'texte', 'Claudine', 'Giroux',  
'Tristan', 'Champagne', 'Lessard' ]
```

On peut aussi retrancher la ponctuation (la ponctuation n'est pas incluse dans les « mots vides »)

```
len(tokens) = 16
```

Traitement automatique du langage naturel

Mots vides (*stopwords*)

'toute', 'toutefois', 'utes', 't', 't',
'trente', 'tes', 'trois', 't', 't',
'trois', 'lement', 'op', 't', 't',
'tsou', 't', 't', 't', 't', 't',
'uni', 't', 't', 't', 't', 't',
'vais', 't', 't', 't', 't', 't',
'ving', 't', 't', 't', 't', 't',
'voici', 't', 't', 't', 't', 't',
'vous-mêmes', 'vu', 'vé', 'vôtre', 'vôtres',
'zut', 'à', 'â', 'ca', 'ès', 'étaient', 'étais',
'était', 'Il est possible d'en ajouter ou d'en enlever.'

Traitement automatique du langage naturel

Entités nommées (*named entities*)

```
entites =  
[ 'Tournoi de badminton universitaire', 'UdeM  
domine', 'l'UQAM', 'Claudine Giroux', 'Tristan  
Champagne', 'Lessard']
```

Traitement automatique du langage naturel

Entités nommées (*named entities*)

{ 'UdeM domine' : 'MISC' }

{ 'l'UQAM' : 'ORG' }

{ 'Claudine Giroux' : 'PER' }

{ 'Tristan Champagne' : 'PER' }

{ 'Québec' : 'LOC' }

{ 'Devoir' : 'LOC' }

{ 'Le Petit Bonheur' : 'MISC' }

{ 'Société québécoise du cannabis' : 'ORG' }

{ 'Boeing' : 'ORG' }

{ 'Abidjan' : 'LOC' }

Fonctionne mieux avec du texte
anglais...

Traitement automatique du langage naturel

Analyse de fréquence

« *Bag of words* »

Calcul des fréquences.

TF/IDF

Term Frequency/
Inverse Document Frequency

Traitement automatique du langage naturel

Analyse de fréquence

TF/IDF

Exemple avec *Globe & Mail*

Analyse de l'utilisation du « terroris* »

1er juin 1970

searched for evidence Mrs Bronfman said she had no idea why the house
was selected by bombers This was political kind of thing with
terrorists.

The Brontmans have lived in their house for about 15 years Mr Bronfman
said just hope the police get them before they kill ame- body

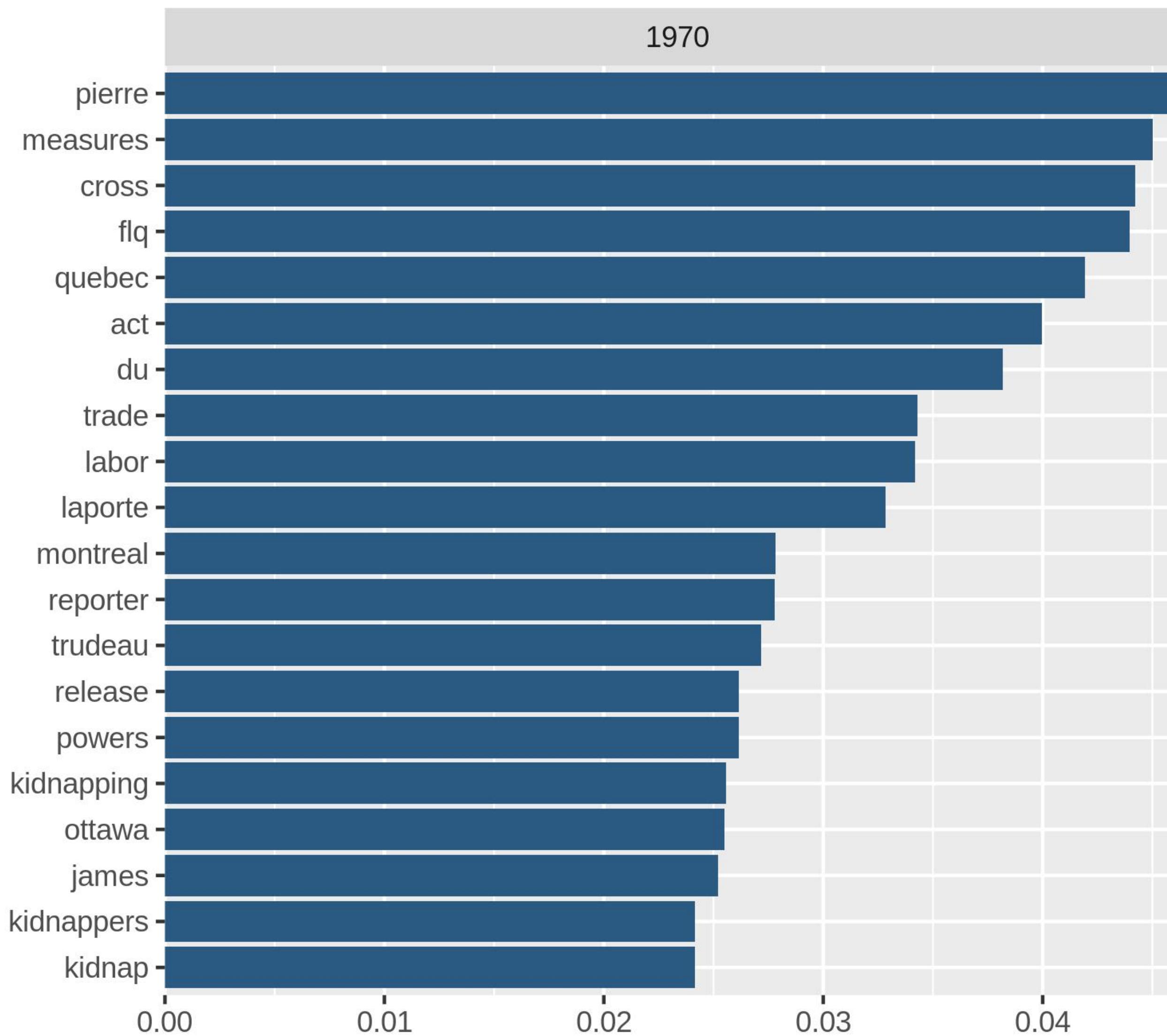
}

25 mots avant

}

25 mots après

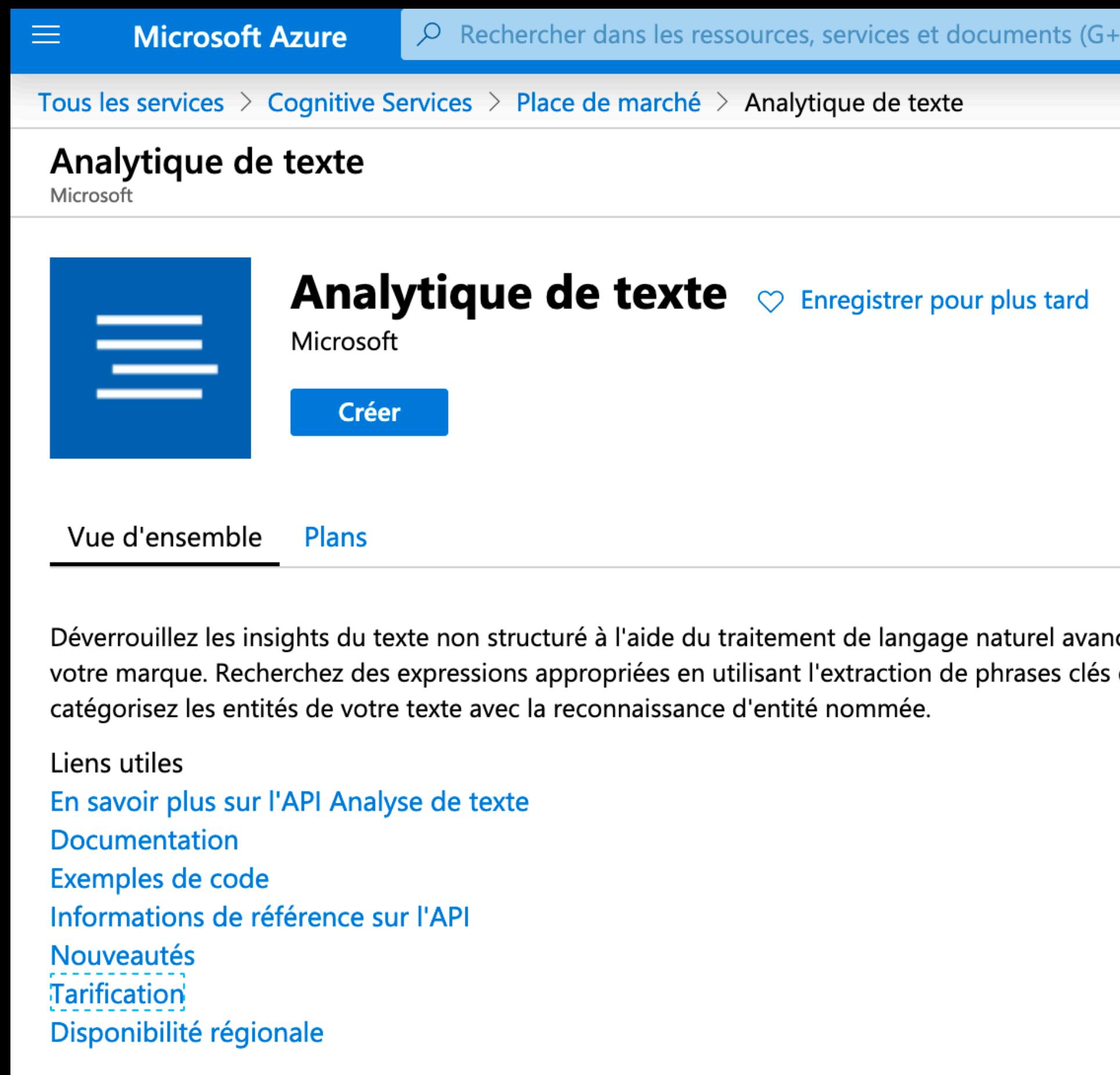
Traitements automatiques du langage naturel



Fréquence de
chaque mot dans
toute l'année 1970
divisé par
l'inverse de sa
fréquence dans les
autres années

Traitement automatique du langage naturel

Analyse de sentiments (*sentiment analysis*)



The screenshot shows the Microsoft Azure Cognitive Services marketplace. The top navigation bar includes the Microsoft Azure logo, a search bar, and a menu icon. Below the navigation, the breadcrumb trail reads: Tous les services > Cognitive Services > Place de marché > Analytique de texte. The main content area features a large blue card for the "Analytique de texte" service by Microsoft. The card includes a blue icon with three horizontal lines, a title, a "Créer" button, and a "Vue d'ensemble" tab which is currently selected. Below the card, a descriptive text block explains that the service helps unlock insights from unstructured text using advanced natural language processing. A "Liens utiles" section follows, listing links to the API documentation, examples of code, and information about the API.

Prolongement de votre compte o365.uqam.ca

Recherchez « **Cognitive Services** »

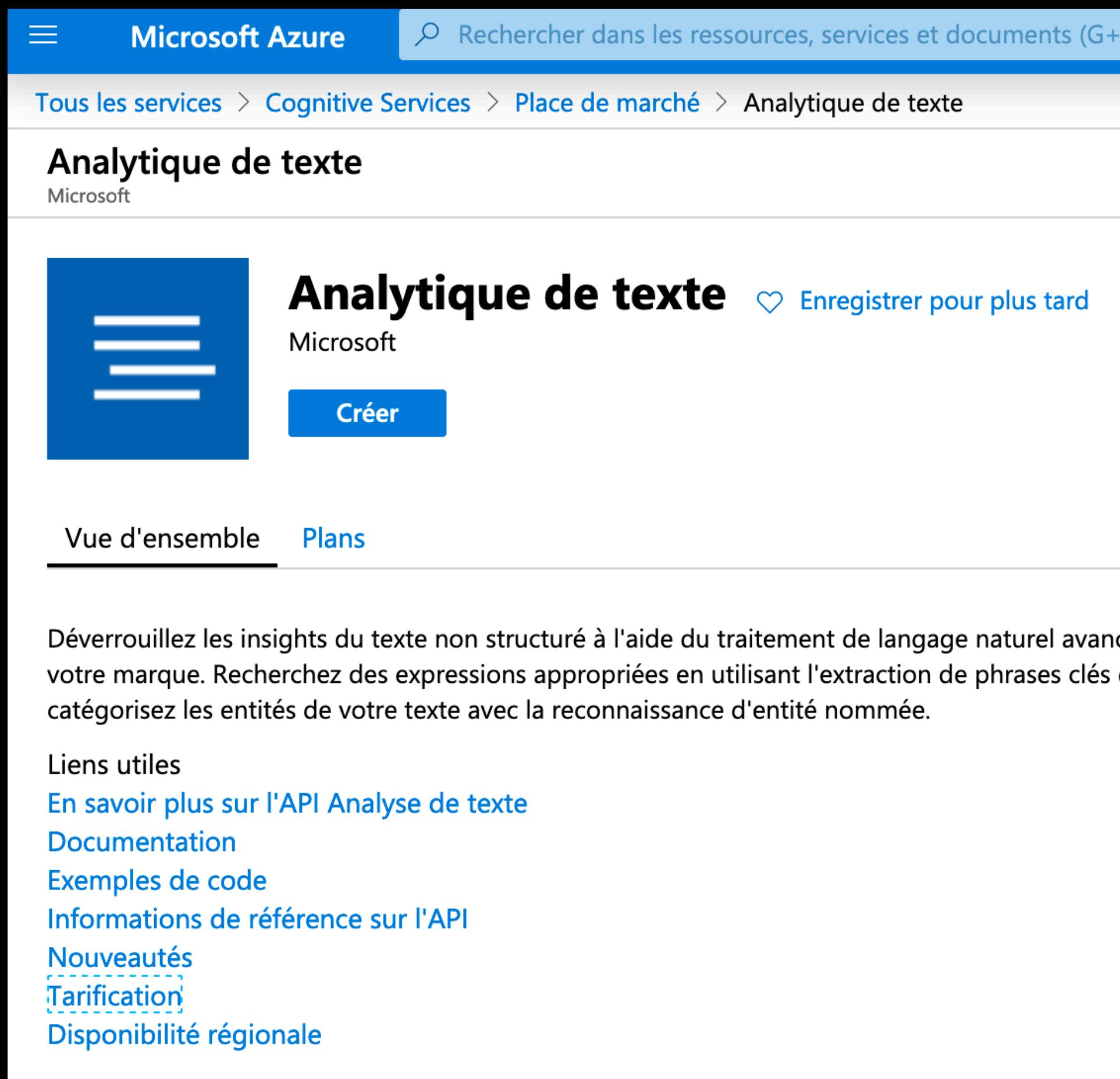
Puis « **Analytique de texte** »

- Traduction
- Détection de langue
- Entités nommées
- Analyse de sentiments

Limite: 5000 appels / mois gratuits

Traitement automatique du langage naturel

Analyse de sentiments (*sentiment analysis*)



The screenshot shows the Microsoft Azure Cognitive Services Text Analytics page. At the top, there's a navigation bar with 'Microsoft Azure' and a search bar. Below it, the path 'Tous les services > Cognitive Services > Place de marché > Analytique de texte' is visible. The main title 'Analytique de texte' is displayed with a Microsoft logo. A large blue button labeled 'Créer' is prominent. Below the button, there are two tabs: 'Vue d'ensemble' (selected) and 'Plans'. A descriptive paragraph explains the service's purpose: 'Déverrouillez les insights du texte non structuré à l'aide du traitement de langage naturel avancé. votre marque. Recherchez des expressions appropriées en utilisant l'extraction de phrases clés et catégorisez les entités de votre texte avec la reconnaissance d'entité nommée.' A sidebar on the left lists useful links: 'Liens utiles', 'En savoir plus sur l'API Analyse de texte', 'Documentation', 'Exemples de code', 'Informations de référence sur l'API', 'Nouveautés', 'Tarification' (highlighted with a dashed border), and 'Disponibilité régionale'.

{"language": "fr", "text": "Tournoi de badminton universitaire : l'UdeM domine, l'UQAM repart avec le bronze après six heures de compétition. Un texte de Claudine Giroux et de Tristan Champagne-Lessard."}

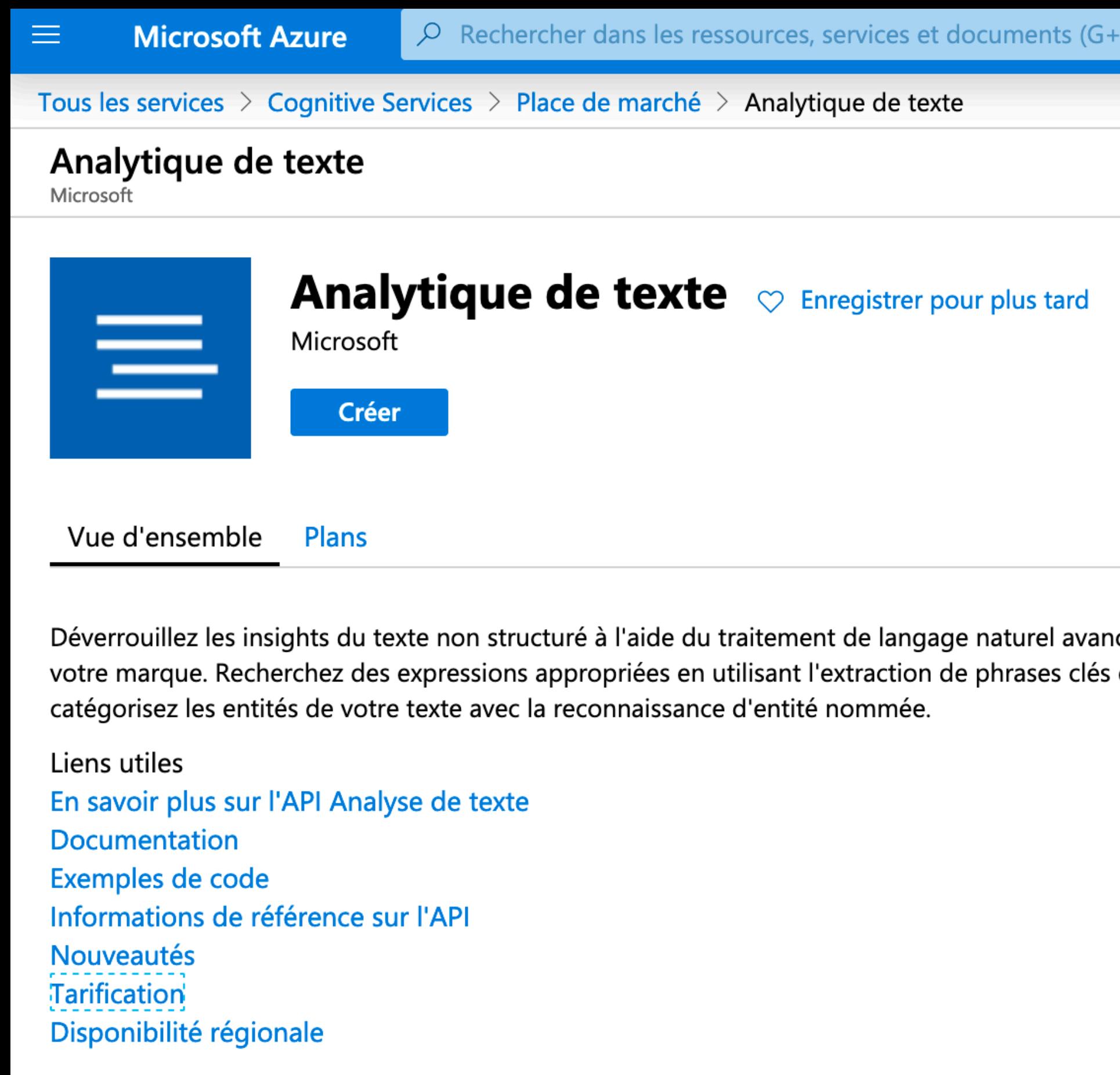
{ 'score': 0.6334215998649597 }

1 = positif

0 = négatif

Traitement automatique du langage naturel

Analyse de sentiments (*sentiment analysis*)



The screenshot shows the Microsoft Azure Cognitive Services Text Analytics interface. At the top, there's a navigation bar with 'Microsoft Azure' and a search bar. Below it, a breadcrumb trail shows 'Tous les services > Cognitive Services > Place de marché > Analytique de texte'. The main title 'Analytique de texte' is displayed with a Microsoft logo. A large blue button labeled 'Créer' is prominent. Below the button, there are two tabs: 'Vue d'ensemble' (selected) and 'Plans'. A descriptive text block explains the service's purpose: 'Déverrouillez les insights du texte non structuré à l'aide du traitement de langage naturel avancé. Votre marque. Recherchez des expressions appropriées en utilisant l'extraction de phrases clés et catégorisez les entités de votre texte avec la reconnaissance d'entité nommée.' A sidebar on the left lists useful links: 'Liens utiles', 'En savoir plus sur l'API Analyse de texte', 'Documentation', 'Exemples de code', 'Informations de référence sur l'API', 'Nouveautés', 'Tarification' (highlighted with a dashed border), and 'Disponibilité régionale'.

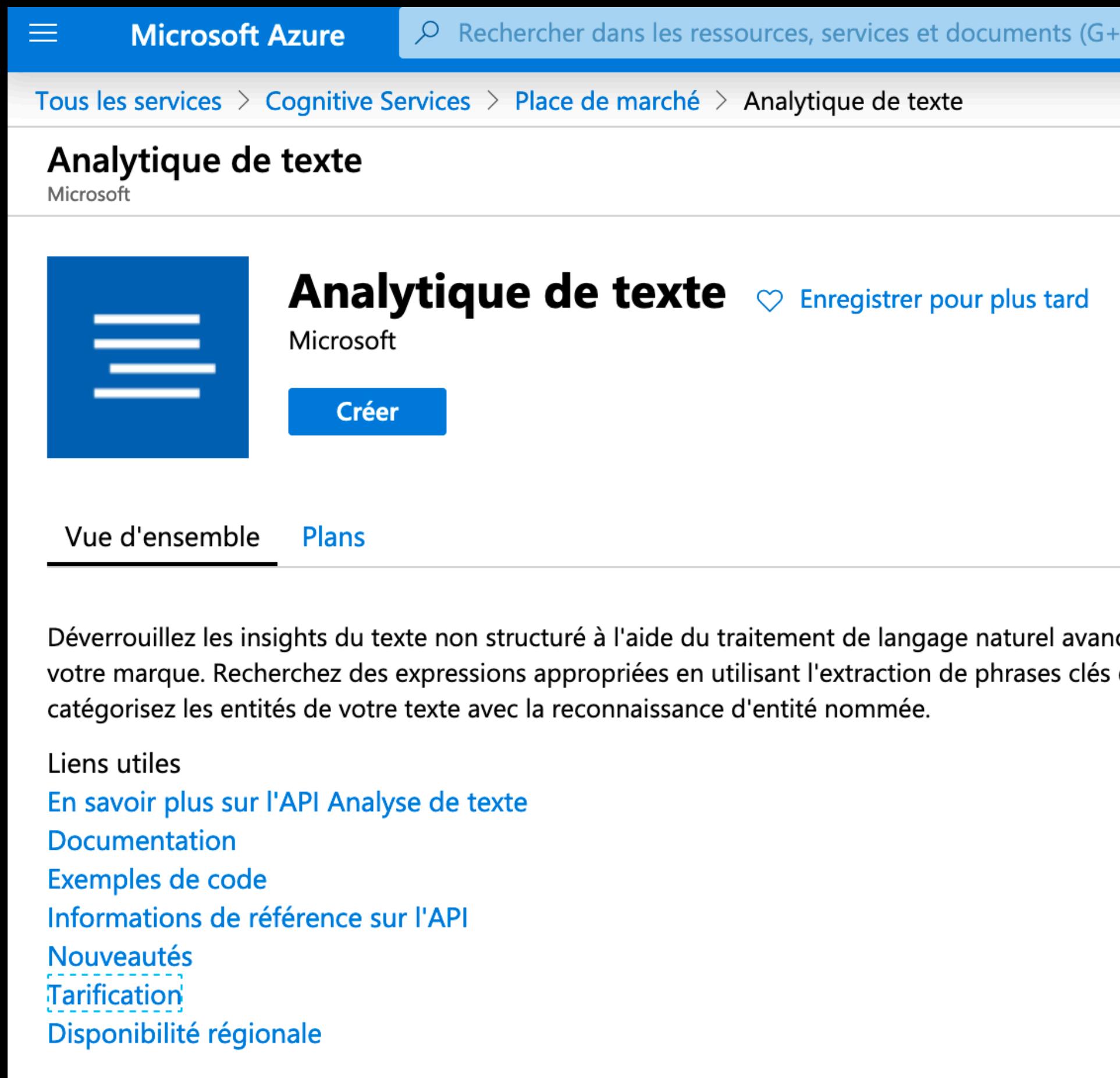
{"language": "fr", "text": "Radio-Canada est un repaire de communistes payer par Justin Trudeau pour nous mentire. Ses l'usine à Fake News par excellence. Tout sa, payer avec nos taxe. Fermer Radio-Canada au pc, sa presse!!!"}
{"score": 0.5465889573097229}

1 = positif

0 = négatif

Traitement automatique du langage naturel

Analyse de sentiments (*sentiment analysis*)



The screenshot shows the Microsoft Azure portal interface for the Text Analytics service. At the top, there's a navigation bar with 'Microsoft Azure' and a search bar. Below it, the breadcrumb navigation shows 'Tous les services > Cognitive Services > Place de marché > Analytique de texte'. The main title 'Analytique de texte' is displayed with a Microsoft logo. A large blue button labeled 'Créer' is prominent. Below the title, there are two tabs: 'Vue d'ensemble' (selected) and 'Plans'. A descriptive text block explains the service's purpose: 'Déverrouillez les insights du texte non structuré à l'aide du traitement de langage naturel avancé. votre marque. Recherchez des expressions appropriées en utilisant l'extraction de phrases clés et catégorisez les entités de votre texte avec la reconnaissance d'entité nommée.' A section titled 'Liens utiles' lists various links: 'En savoir plus sur l'API Analyse de texte', 'Documentation', 'Exemples de code', 'Informations de référence sur l'API', 'Nouveautés', 'Tarification' (which is dashed), and 'Disponibilité régionale'.

{"language": "fr", "text": "Je vais t'occire, espèce de Bachi-Bouzouk de moule à gaufres! Je vais t'écraser, sinistre scolopendre d'ectoplasme de triple crétin des Alpes de mille milliards de tonnerre de Brest!"}

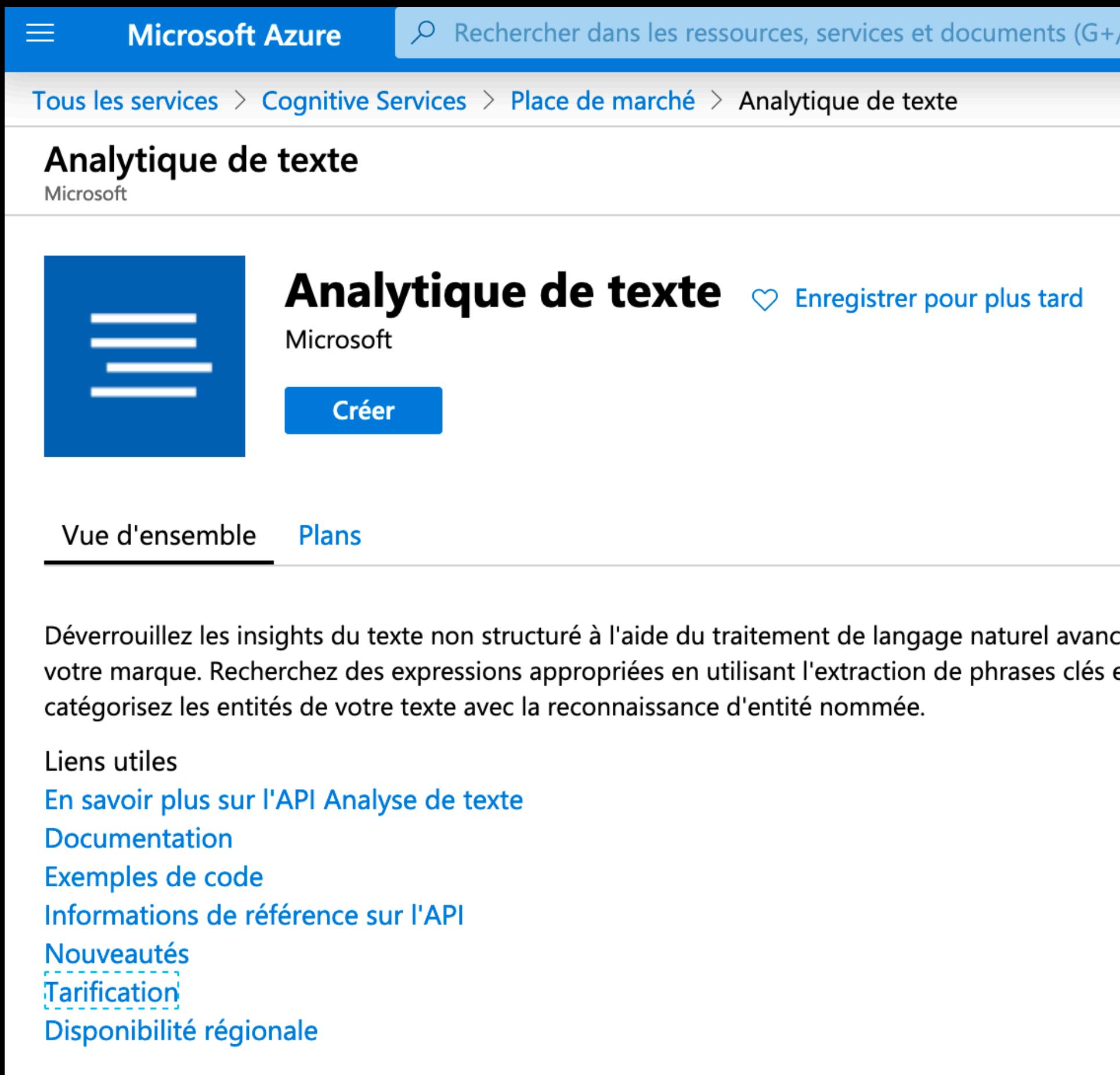
{ 'score': 0.4743112325668335}

1 = positif

0 = négatif

Traitement automatique du langage naturel

Analyse de sentiments (*sentiment analysis*)

A screenshot of the Microsoft Azure portal showing the Text Analytics service. The top navigation bar includes 'Microsoft Azure' and a search bar. Below it, the breadcrumb navigation shows 'Tous les services > Cognitive Services > Place de marché > Analytique de texte'. The main title 'Analytique de texte' is displayed with a Microsoft logo. A large blue button labeled 'Créer' is visible. The page content area starts with a heading 'Analytique de texte' and a Microsoft logo, followed by a 'Créer' button. Below this, there are tabs for 'Vue d'ensemble' (selected) and 'Plans'. A descriptive text block explains the service's purpose: 'Déverrouillez les insights du texte non structuré à l'aide du traitement de langage naturel avancé. votre marque. Recherchez des expressions appropriées en utilisant l'extraction de phrases clés et catégorisez les entités de votre texte avec la reconnaissance d'entité nommée.' A sidebar on the left lists useful links: 'Liens utiles', 'En savoir plus sur l'API Analyse de texte', 'Documentation', 'Exemples de code', 'Informations de référence sur l'API', 'Nouveautés', 'Tarification' (highlighted with a dashed border), and 'Disponibilité régionale'.

```
{"language": "fr", "text": "Comme je  
ne m'y attendais pas, mon cœur  
s'est mis à me parler de toi! Je  
t'aime tant!"}
```

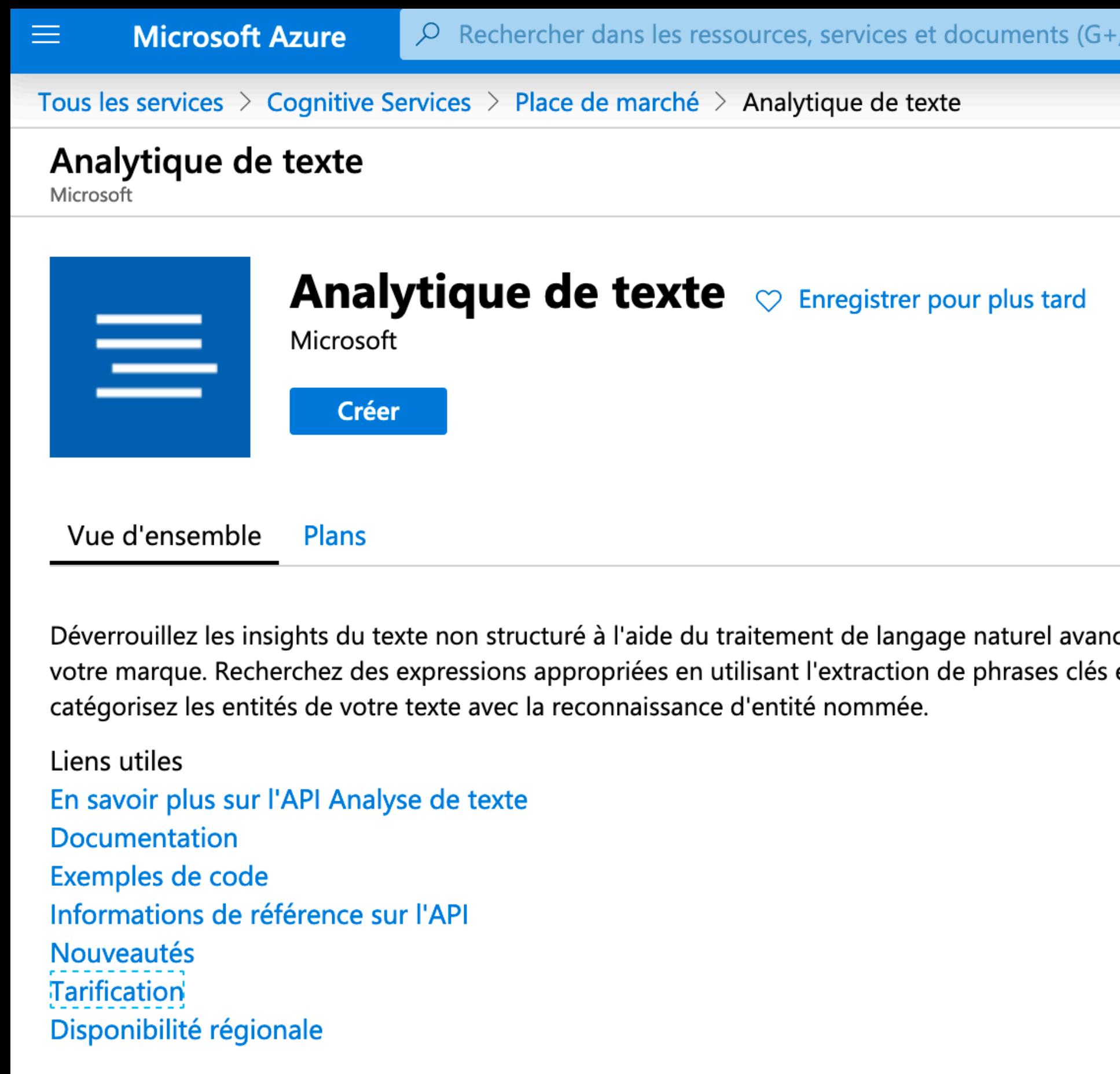
```
{ 'score': 0.6525964736938477 }
```

1 = positif

0 = négatif

Traitement automatique du langage naturel

Analyse de sentiments (*sentiment analysis*)



The screenshot shows the Microsoft Azure portal interface for the Text Analytics service. At the top, there's a navigation bar with the Microsoft Azure logo and a search bar. Below it, a breadcrumb trail shows 'Tous les services > Cognitive Services > Place de marché > Analytique de texte'. The main title 'Analytique de texte' is displayed with a Microsoft logo. To the left is a blue icon with three horizontal lines. In the center, there's a large button labeled 'Analytique de texte' with a Microsoft logo, and a smaller 'Créer' button below it. At the bottom, there are two tabs: 'Vue d'ensemble' (selected) and 'Plans'. A descriptive text block follows, followed by a section titled 'Liens utiles' with several links.

Déverrouillez les insights du texte non structuré à l'aide du traitement de langage naturel avancé. votre marque. Recherchez des expressions appropriées en utilisant l'extraction de phrases clés et catégorisez les entités de votre texte avec la reconnaissance d'entité nommée.

Liens utiles

- [En savoir plus sur l'API Analyse de texte](#)
- [Documentation](#)
- [Exemples de code](#)
- [Informations de référence sur l'API](#)
- [Nouveautés](#)
- [Tarification](#)
- [Disponibilité régionale](#)

{"language": "fr", "text": "Comme je ne m'y attendais pas, mon cœur s'est mis à me parler de toi! Tu es si magnifique! Je t'aime tant!"}

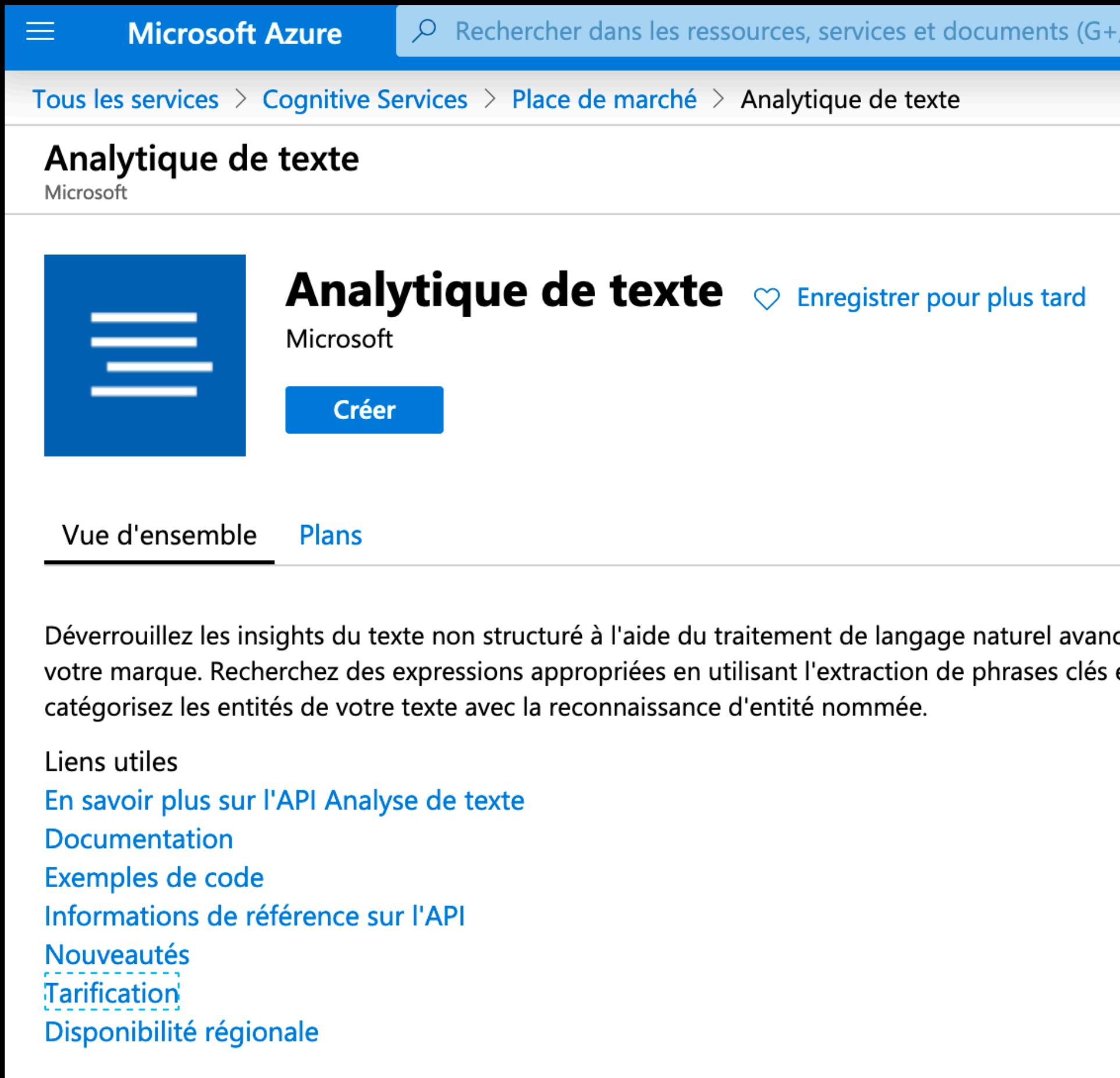
{ 'score': 0.899999761581421 }

1 = positif

0 = négatif

Traitement automatique du langage naturel

Analyse de sentiments (*sentiment analysis*)



The screenshot shows the Microsoft Azure portal interface for the Text Analytics service. At the top, there's a navigation bar with 'Microsoft Azure' and a search bar. Below it, a breadcrumb trail shows 'Tous les services > Cognitive Services > Place de marché > Analytique de texte'. The main title 'Analytique de texte' is displayed with a Microsoft logo. To the left is a blue icon with three horizontal lines. Below the title, there's a 'Créer' button. The main content area has two tabs: 'Vue d'ensemble' (selected) and 'Plans'. A descriptive paragraph explains the service's purpose: 'Déverrouillez les insights du texte non structuré à l'aide du traitement de langage naturel avancé. votre marque. Recherchez des expressions appropriées en utilisant l'extraction de phrases clés et catégorisez les entités de votre texte avec la reconnaissance d'entité nommée.' Below this are several utility links: 'Liens utiles', 'En savoir plus sur l'API Analyse de texte', 'Documentation', 'Exemples de code', 'Informations de référence sur l'API', 'Nouveautés', 'Tarification' (highlighted with a dashed border), and 'Disponibilité régionale'.

```
{"language": "fr", "text": "Comme je  
ne m'y attendais pas, mon cœur  
s'est mis à me parler de toi! Tu es  
si magnifique! Je t'aime tant!"}
```

```
{ 'score': 0.8999999761581421}
```

1 = positif

0 = négatif

L'ajout de 4 petits mots change
radicalement le score
À utiliser avec précaution, donc...

Traitements automatiques du langage naturel

Modélisation thématique (*topic modeling*)

OPINIONS RICHARD MARTINEAU

RICHARD
MARTINEAU

4947 chroniques
du 1er janvier 2010
au 1er mai 2019

9,2M de caractères
1,8M de mots

Quels mots ont le plus tendance à se retrouver ensemble dans les mêmes chroniques?



Traitements automatiques du langage naturel

Modélisation thématique (*topic modeling*)

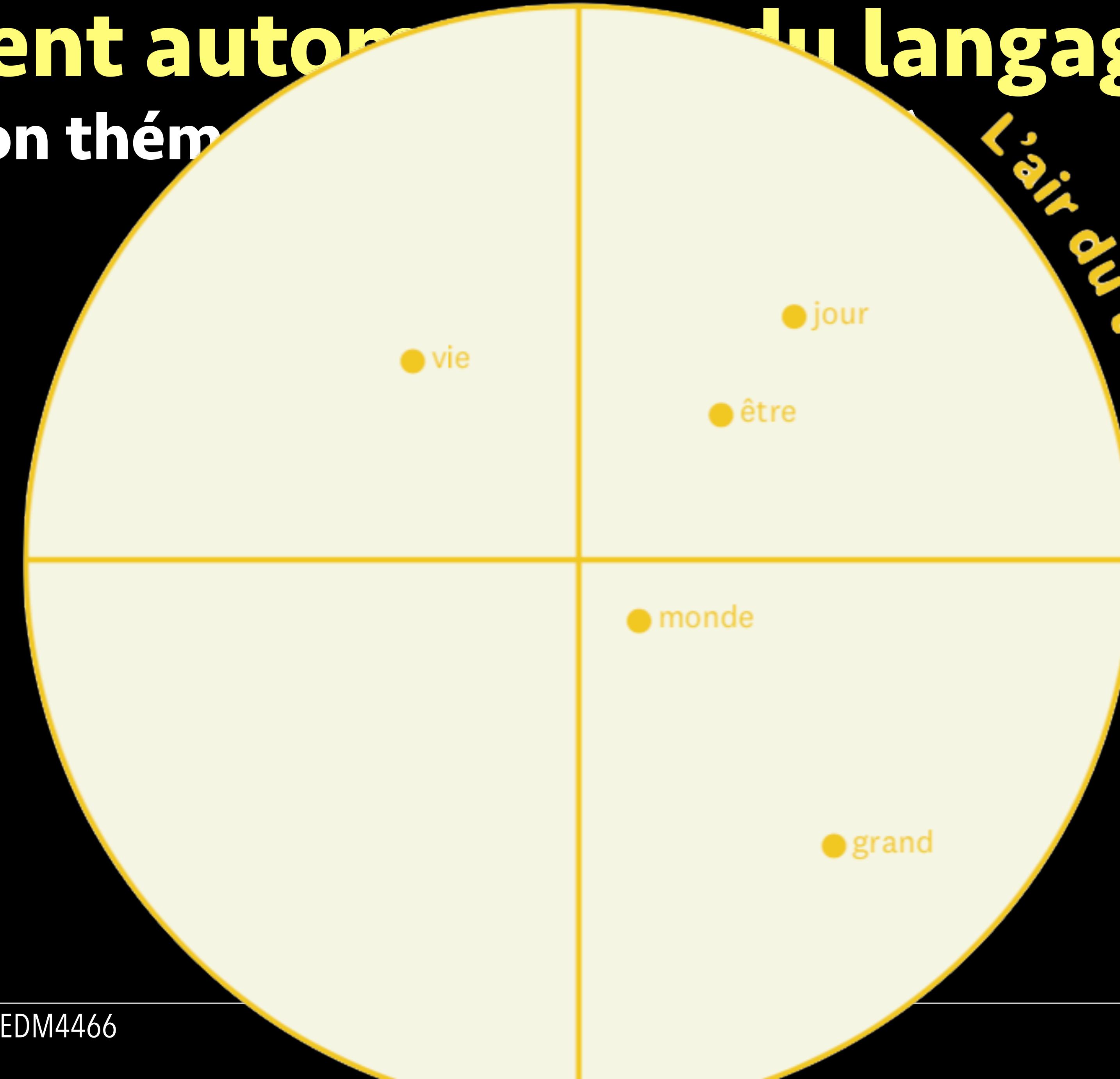
Quels mots ont le plus tendance à se retrouver ensemble dans les mêmes chroniques?

La modélisation thématique ne va pas nommer les thèmes. C'est à vous de trouver qu'est-ce qui lie les mots qui ont été identifiés comme appartenant probablement à un même thème.

Traitement automatisé Modélisation thématique

Le filtre du langage naturel

L'air du temps



Traitement automatique du langage naturel

Modélisation de thèmes (*topic modeling*)

M

Journa

vie
être
jour
monde
grand

air du temps

Le PQ

libéral
ministre
parti
PQ
politique

Traitement automatique du langage naturel *(topic modeling)*

M

Le PQ

libéral

ministre

parti

PQ

politique

L'islam

laïcité

charte

musulman

religieux

voile

Traitement automatique du langage naturel

Reconnaissance optique des caractères (OCR)

Comment lire:

- des PDF « natifs »
- des PDF « image »

Des tableaux dans des
PDF avec Tabula

