Soutenance TER M1 : Hotel Advisor

Outils d'analyse automatique d'avis d'hôtel

Étudiants

Bastien Carbonnier
Cedric Durand
Johann Golmard

Encadrants

Mathieu Lafourcade
Pierre Pompidor

4 juin 2019

Introduction

Sujet

Réaliser une application qui analyse automatiquement des avis textuels sur des hôtels pour ensuite « colorer » les différents descripteurs des ontologies représentant ces différents hôtels.

Idée générale

Avis

la chambre était très sale

Après polarisation

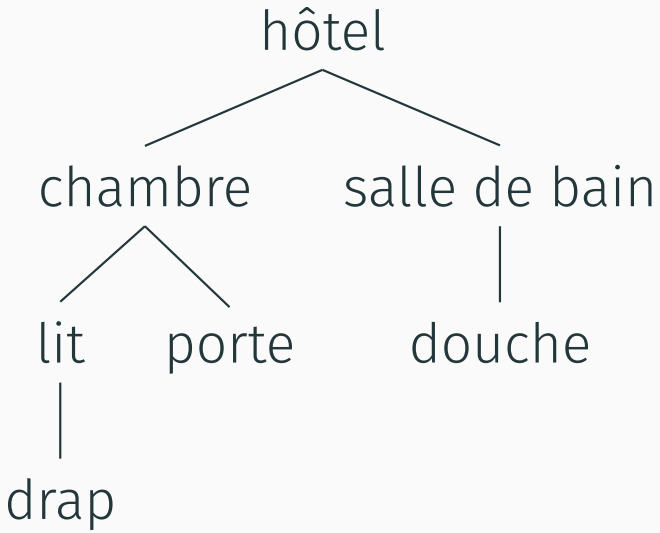
la chambre était très sale

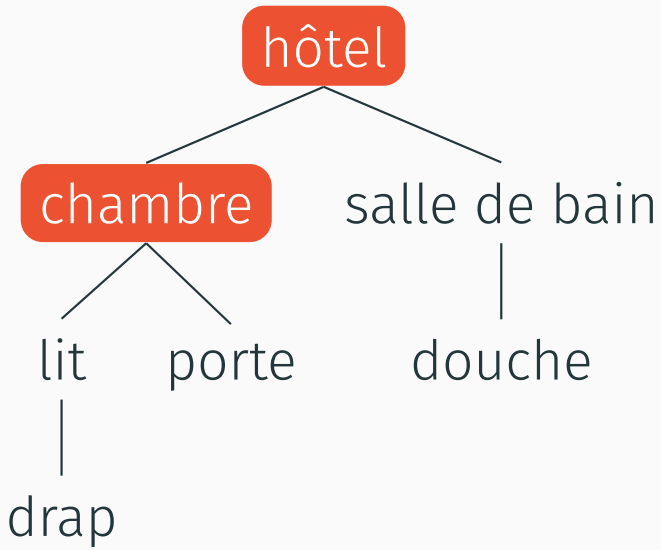
Après propagation

la chambre était très sale

Idée générale

Ontologies :





Sommaire

1. État de l’art

2. Outils utilisés

3. Architecture de notre projet

4. Limites et évolutions

État de l’art

Étiquetage morphosyntaxique

Définition

Consiste à identifier pour chaque mot sa nature grammaticale à partir du contexte de la phrase dans laquelle il se situe et de connaissances lexicales.

- Un mot peut avoir plusieurs natures
- Librairies Python TreeTagger, SpaCy, NLTK
- **Important** : Base de notre propagation de polarisation

Comparaison TreeTagger et SpaCy

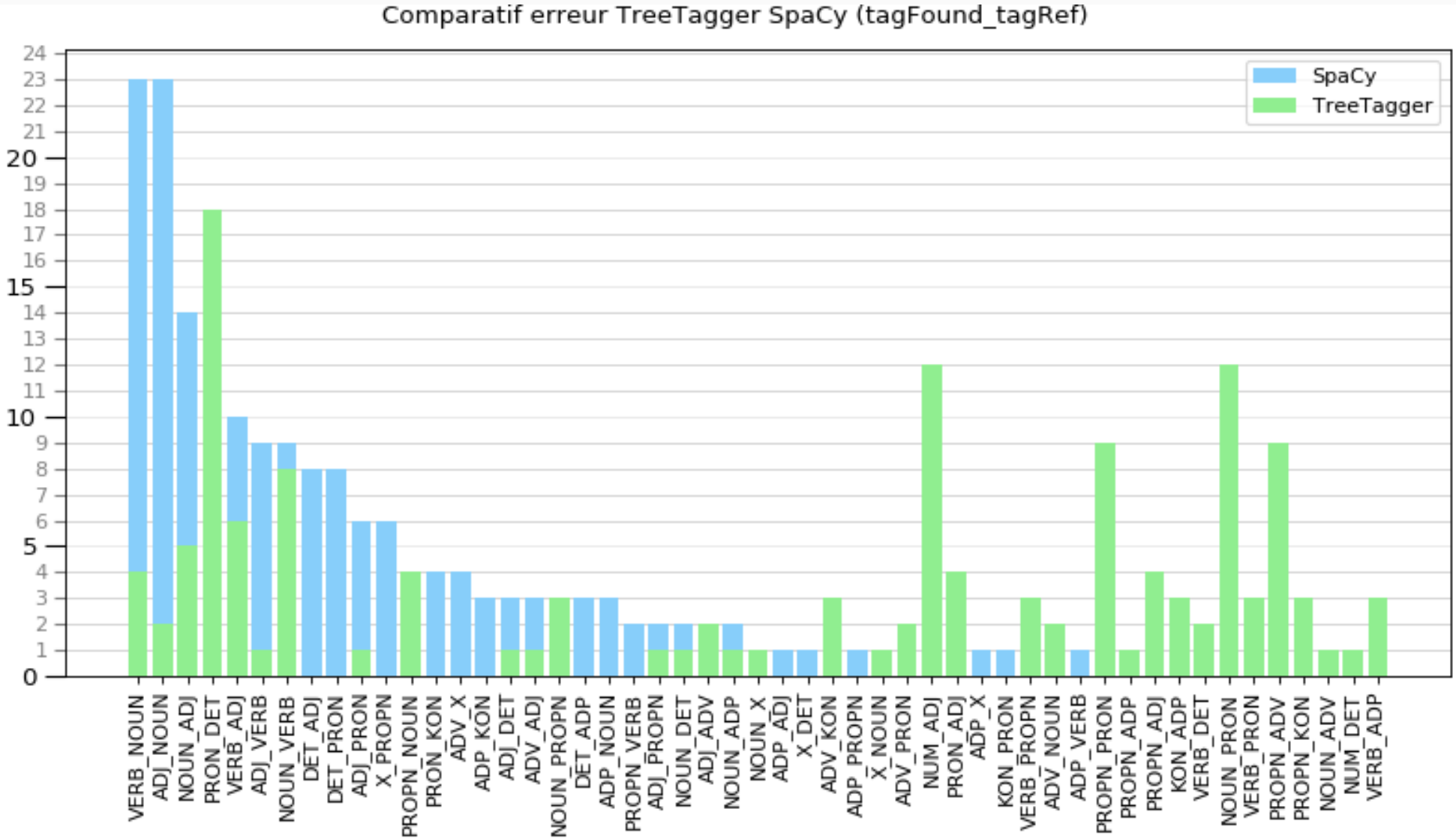
Protocole

- 15 Avis réels (2331 mots)
- Redressement orthographique
- Étiquetage grammatical en utilisant les 2 librairies
- 291 termes étiquetés différemment
- Recherche des bonnes natures pour ces termes
- Comparaison et interprétation des résultats

Précisions :

- TreeTagger = $\simeq$ 53%
- SpaCy = $\simeq$ 46%

Comparaison TreeTagger et SpaCy



Comparaison TreeTagger et SpaCy

SpaCy (pred_ref)			TreeTagger (pred_ref)		
verb_noun	**adj_noun**	**noun_adj**	**pron_det**	**noun_adj**	**noun_pron**
mer	jardins	junior	chaque	futée	Il
déjeuner	safari	grossier	cet	troisième	tout
coucher	buffets	bel	les	solo	J
bar	armoire	sympa	cette	petits	Tous
déjeuner	pente	chaude	le	moderne	Ils
forfait	bars	superbe	les		
fruits	magasins	gratuite			
jacuzzis	restaurants	cool			
vigile	vagues	correct			
serviettes	lits	médiocre			
Suites	bouilloire	abordables			

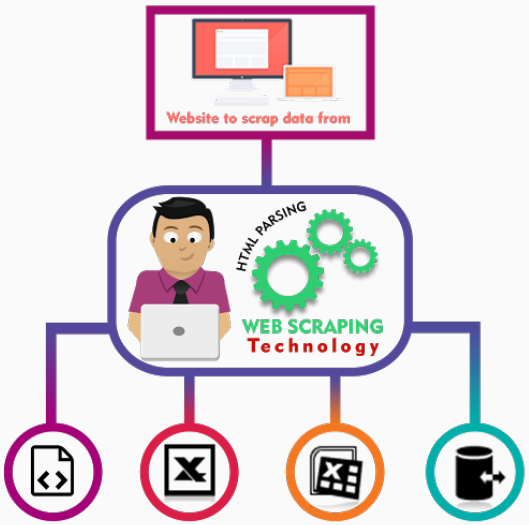
Table 1 – Mots mal étiquetés

Comparaison TreeTagger et SpaCy

Conclusion

- Erreurs d'étiquetages de TreeTagger plus facilement réparables
- Erreurs d'étiquetages de SpaCy impactent la sémantique de la phrase
- Étude relative et non complète

Scraping




```
135  4.0|Ce que l'on a aimé : Le all inclusive avec un grand choix de plats et des alcools internationaux. La proximité du
     centre commercial Yumbo, Très grande chambre studio avec belle terrasse et vue mer, mais un peu surannée, un solarium
     naturiste sur le toit de l'hôtel.
136  4.0|Bien dans l'ensemble, nous avons passé une semaine agréable, la piscine est assez petite quand aux transats il faut se
     lever de bonheur pour pouvoir en bénéficier car il sont prix d'assaut.A revoir l'insonorisation des chambres qui fait
     défaut
137  4.0|hôtel impeccable, jardins magnifiques et entretenus tous les jours, nourriture variée et excellente(notamment les jus
     de fruits presses le matin)literie très confortable.petit bémol, peu de Français, donc, peu de personnel le parlant...a
     conseiller, la formule demi pension suffit largement a mon avis
138  4.0|Avons passé une semaineBuffet varié Personnel agréable Animation présente , de beaux spectacles le soirDommage que le
     programme n'est pas diffusé sur papier dans les chambres Beaucoup de monde mais 2 piscines dont 1 au calme
139  4.0| bon rapport qualité prix très peux de chose à dire sur l'ensemble des prestations,nous sommes satisfaits aucune
     remarque, peut être l'animation à revoir qui pour moi est leur point faible,nous étions 3 adultes et 3 enfants dont 2
     bébés tout était parfait pour ce séjour d'une semaine en famille
140  4.0|J'ai voyagé à travers toute la planète et les îles canaries sont parmi les plus belles destinations. Le personnel de
     occidental margaritas est formidable. Un professionnalisme hors pair, une incroyable gentillesse. Parmi les meilleurs
     services au monde. Malheureusement la clientèle n'est pas au même niveau. Je me suis fais agresser par un espagnol dont la
     mère a prétendu qu'il était policier pour m'impressionner. J'ai réclamé à la réception et j'espère qu'ils ne vont pas
     prendre le parti de leur compatriote. Ce serait vraiment dommage. En tout cas ces messieurs dames devraient savoir qu'on
     ne peut être impressionné par la police espagnole que quand on a quelque chose à se reprocher. Dans leur cas ce devrait
     être la mauvaise éducation. Le polo que ce messieur à jeté à la piscine vaut beaucoup plus que ses manières de bas fond
141  4.0|hotel joli un peu en travaux en ce moment pisine bien restaurant pas mal mais ca fais trop usine! vu que cest all
     inclus personnel sympa tres bien situé pour les gays cest proche de yumbo 5 min et 10 min de la plage donc cest
     excellent!l'accueil sympa des chats à l hotel trop mignon hihih
142  5.0|Excellent établissement dépaysant, cadre idyllique, personnel dévoué, et à l'écoute de nos souhaits, aimables et
     souriants, agréables sympathique. Établissement propre, et toujours bien entretenu.Très bon souvenirs, y reviendrais bien
     volontiers.
```

Ontologie

Définition

- Outils de représentation d'un corpus
- Ensemble structuré de concepts sous forme de graphe
- Relations sémantiques et/ou de composition et héritage

Ontologie

Termes de l'ontologie

- Domaine restreint à l'hôtellerie
- Chaque terme colorié selon sa polarité
- Ontologie fournie par notre encadrant après formatage

Format JSON

- Lisibilité aisée
- Représentation arborescente
- Compatible avec MongoDB

Ontologie

Représentation avec D3.js

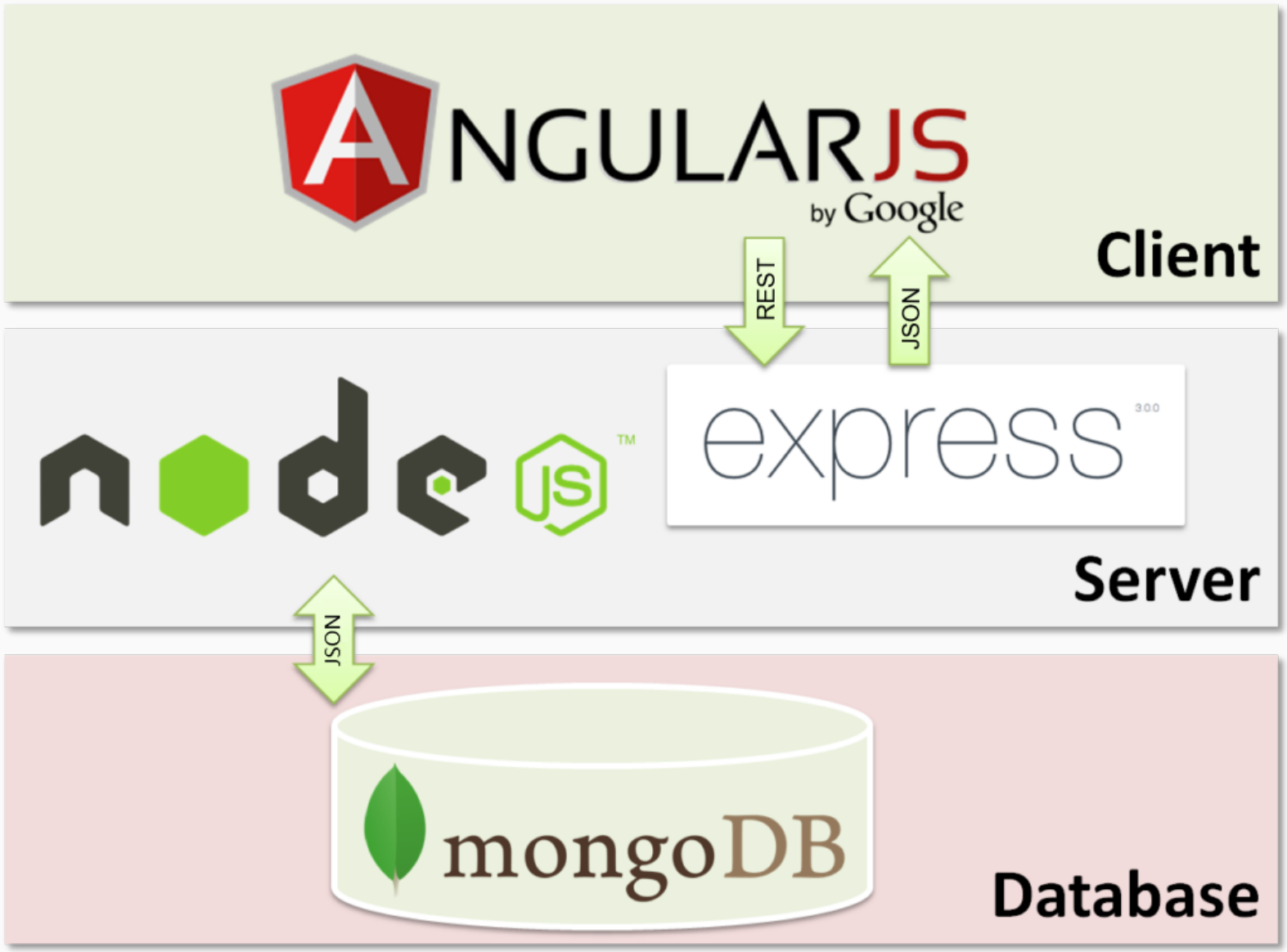
- Permet de lier des données à un objet DOM
- Offre de nombreuses interactions personalisables et fluides
- Librairie présente sur Angular, qui peut être découplée en sous-modules
- Documentation complète et conséquente

Outils utilisés

JeuxDeMots

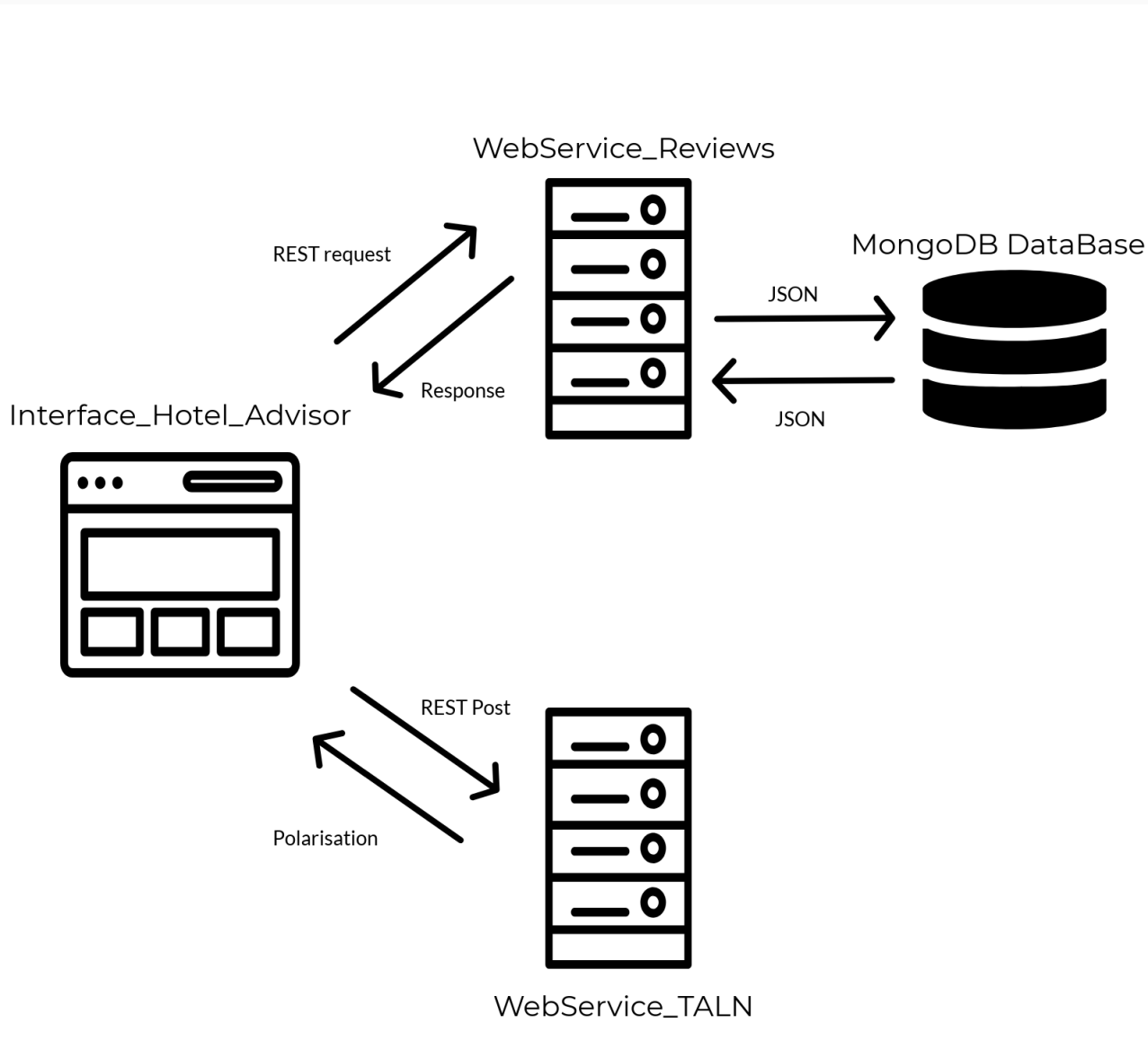
- Réseau lexical
- GWAP (Game with a purpose)
- Point d'accès rezo_dump.php
- Utilisation d'un cache

Architecture MEAN



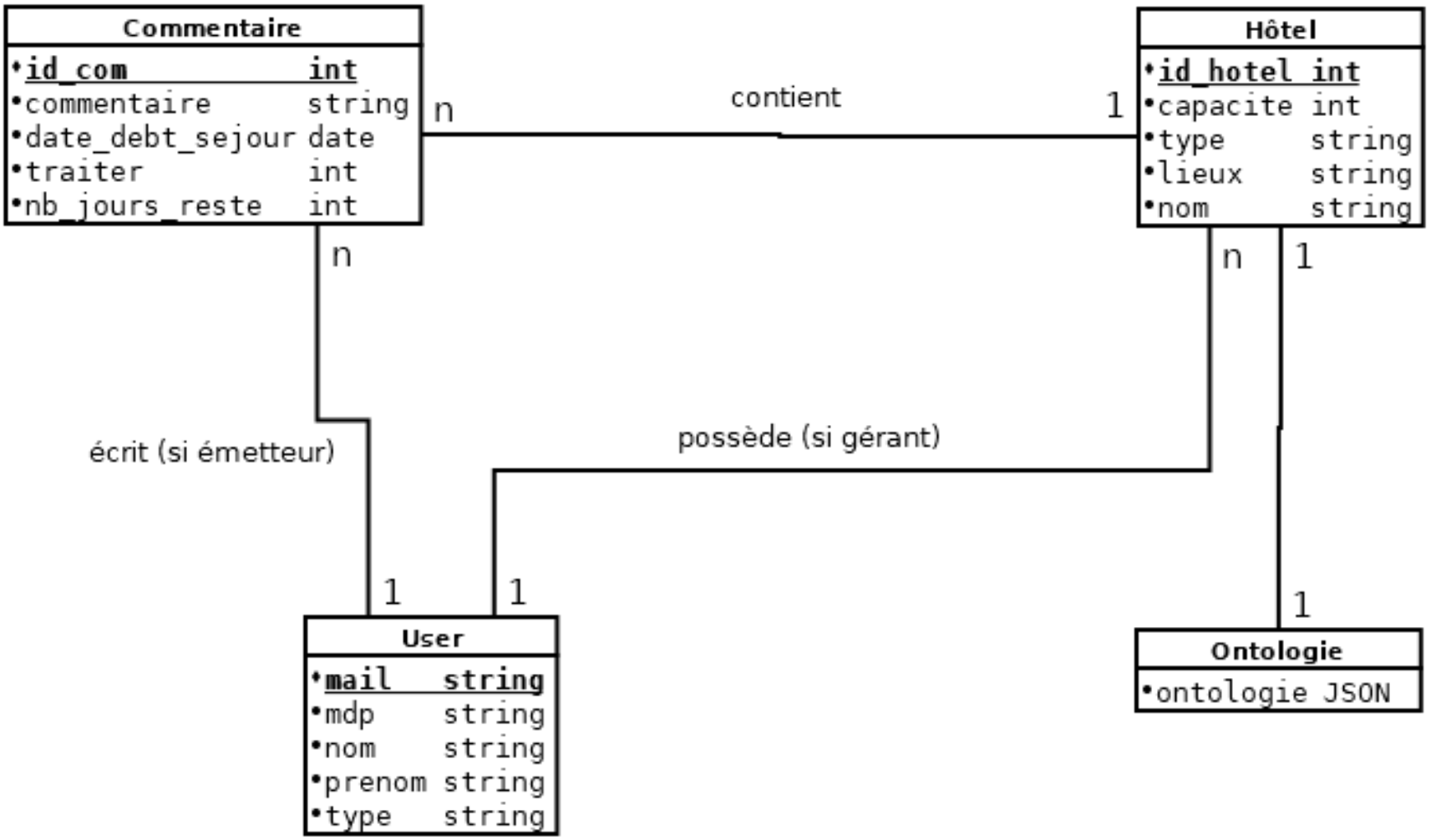
Architecture de notre projet

Schéma global



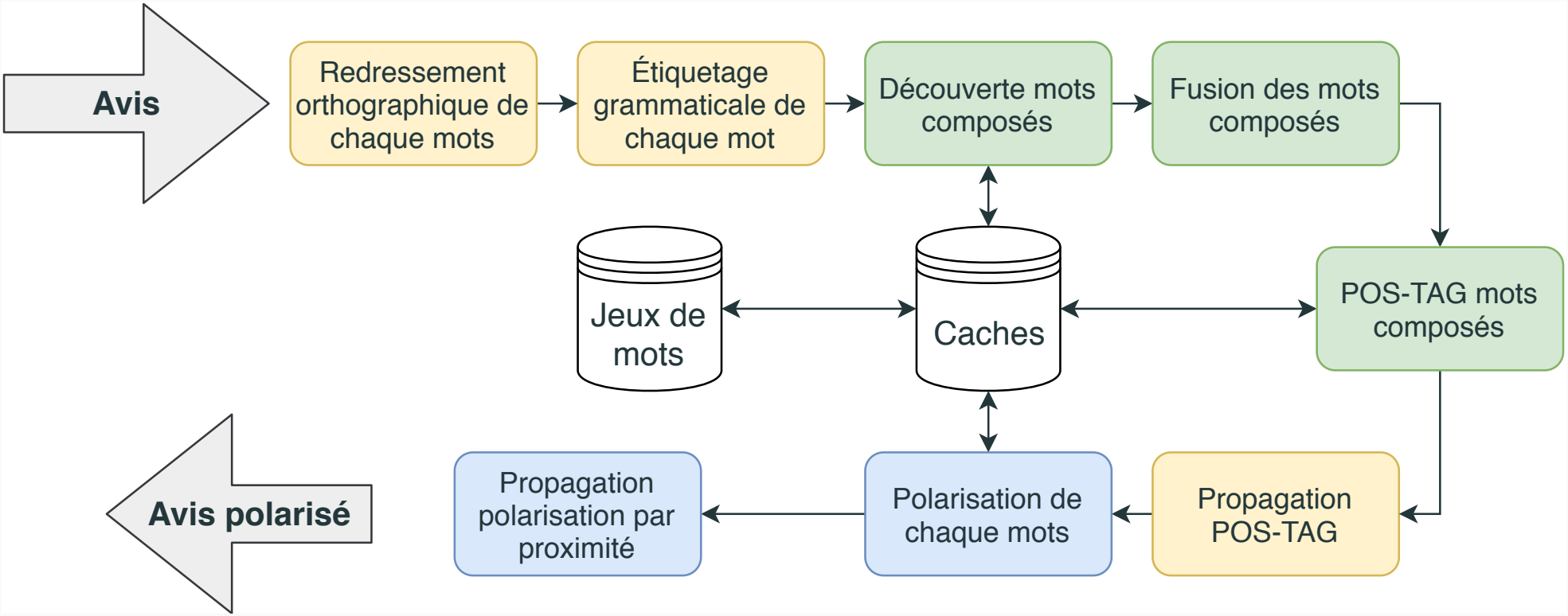
Webservice BDD

Webservice BDD



Webservice TALN

Démarche adoptée



Pré-traitements

Correcteur orthographique

- Module Python spellchecker
- Basé sur la distance de Levenshtein
- Recherche dans une liste contenant la fréquence probable d'un terme

Lemmatisation

- Effectuée par la librairie d'étiquetage grammaticale
- Utilisation du lemme si le mot n'existe pas dans JeuxDeMots

Étiquetage morphosyntaxique

Librairie choisie : TreeTagger

- Meilleur étiquetage grammaticale
- Plus rapide : SpaCy ($\simeq$ 10s par avis), TreeTagger ($\simeq$ 4s par avis)

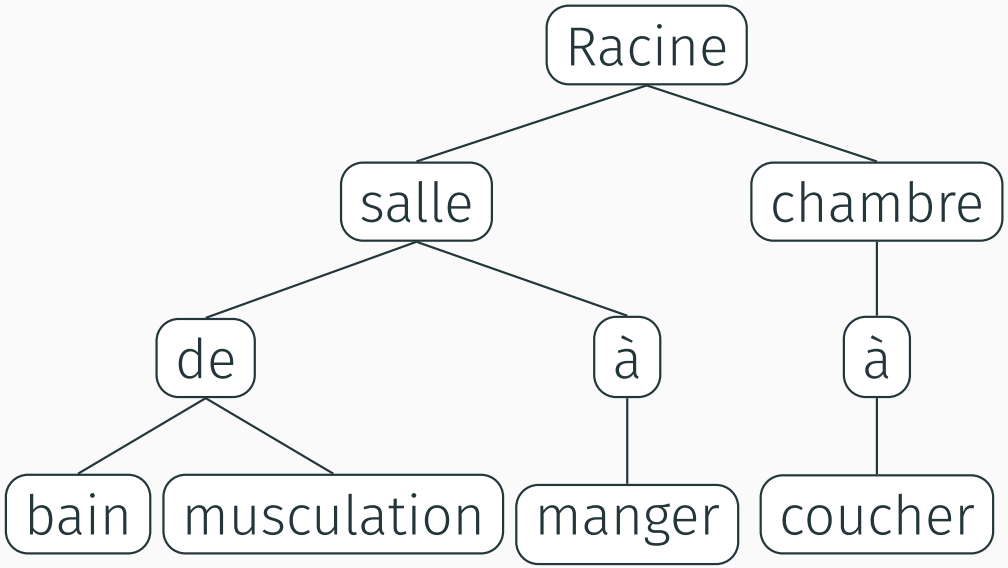
Structure de données utilisée

```
[
  {index : 0, mot : "la", nature : "DET", lem : "le"},
  {index : 1, mot : "chambre", nature : "NOUN", lem : "chambre"
    },
  {index : 2, mot : "était", nature : "VERB", lem : "être"},
  {index : 3, mot : "sale", nature : "ADJ", lem : "sale"}
]
```

Mots composés

Protocole

- Création d'un arbre des préfixes des mots composés
- Recherche pour chaque position i dans la phrase le mot composé le plus grand ayant pour premier mot phrase[i]
- Fusion des mots composés
- Recherche des étiquettes grammaticales de ces nouveaux mots sur JDM



Propagation des étiquettes grammaticales

Protocole

- Pour chaque NOM recherche des ADJs et du VERB
- Pour chaque ADJ recherche des ADVs
- Pour chaque VERB recherche des ADVs

Exemple

```
[
    {index : 0, mot : "la", nature : "DET"},
    {index : 1, mot : "chambre", nature : "NOUN",
      index_verbe : [2],
      index_adj : [4]},
    {index : 2, mot : "était", nature : "VERB"},
    {index : 3, mot : "très", nature : "ADV"},
    {index : 4, mot : "sale", nature : "ADJ",
      index_adv : [3]}
]
```

Polarisation des termes

Idées

1. Gérer la polarité comme un entier (-1, 0 ou +1)
2. Gérer la polarité comme un vecteur

Exemple

```
[   {index : 0, mot : "la", nature : "DET",
        pol: { neg: 0, pos: 0, neutre: 0 }},
    {index : 1, mot : "chambre", nature : "NOUN",
        index_verbe : [2], index_adj : [4],
        pol: { pos: 0.16, neutre: 0.84, neg: 0 }},
    {index : 2, mot : "était", nature : "VERB",
        pol: { pos: 0.75, neutre: 0.16, neg: 0.09 }},
    {index : 3, mot : "très", nature : "ADV"},
    {index : 4, mot : "sale", nature : "ADJ",
        index_adv : [3],
        pol: { pos: 0.02, neutre: 0.02, neg: 0.96 }}]
```

Propagation de la polarisation

Ordre de propagation

- $ADV \rightarrow ADJ, ADV \rightarrow VERB$ (Moyenne)
- $ADJ \rightarrow NOM, VERB \rightarrow NOM$ (Moyenne ou subtitution)

Exemple

```
[   {index : 0, mot : "la", nature : "DET",
        pol: { neg: 0, pos: 0, neutre: 0 }},
    {index : 1, mot : "chambre", nature : "NOUN",
        index_verbe : [2], index_adj : [4],
        pol: { pos: 0.01, neutre: 0.01, neg: 0.98 }},
    {index : 2, mot : "était", nature : "VERB",
        pol: { pos: 0.9, neutre: 0.06, neg: 0.04 }},
    {index : 3, mot : "très", nature : "ADV"},
    {index : 4, mot : "sale", nature : "ADJ",
        index_adv : [3],
        pol: { pos: 0.01, neutre: 0.01, neg: 0.98 }}]
```

Étude quantitative

Protocole

- 14 avis
- 82 termes de notre ontologie

Résultats

- Accuracy = 51,12
- Précision_{pos} = 0.67
- Précision_{neg} = 1
- Rappel_{pos} = 0.69
- Rappel_{neg} = 0.22

		Références		
		1	0	-1
Déduit	1	34	0	17
	0	15	1	8
	-1	0	0	7

Table 2 – Matrice de confusion

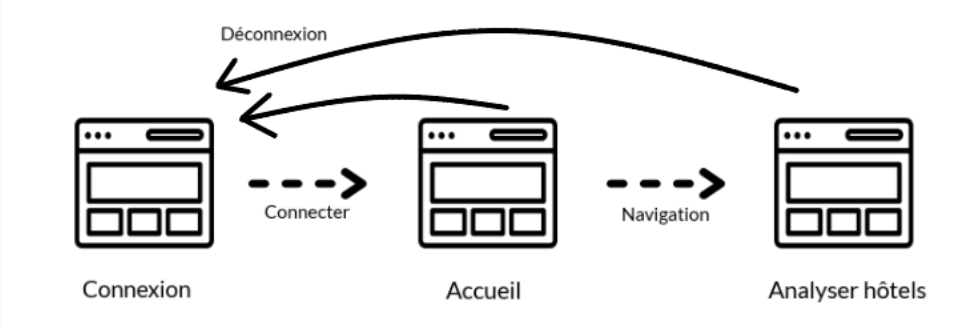
Exemple avis

Salle de bain sale, baignoire bouchée et poils dans la grille... Cuvette des WC jaunit par l'urine et les années... Planché bancale et bruyant... Aucune intimité cloison non isolées et un jour de 10 cm entre le sol et la porte d entree... Sans parlé de la prestation restaurant : accueil et service très moyen, plats sans goûts et pourtant la maison vante les produits du terroir... (indigestion garantie...) et pour terminer un chien couché à l'entree qui empeste l accueil et nous empeche le passage au comptoir.

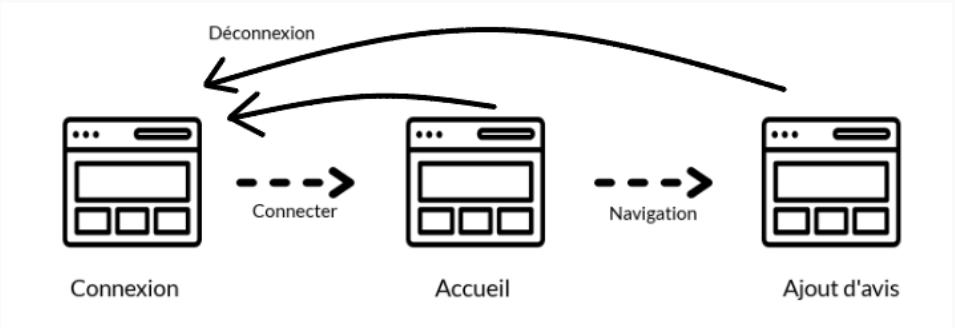
Application Angular

Workflow

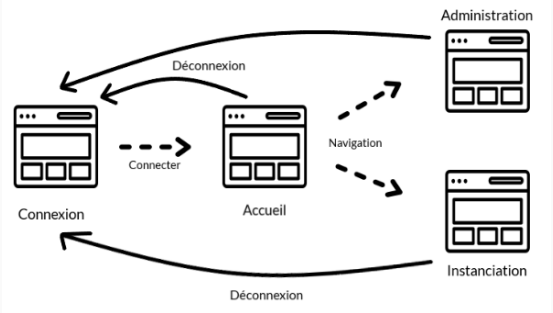
Gérant



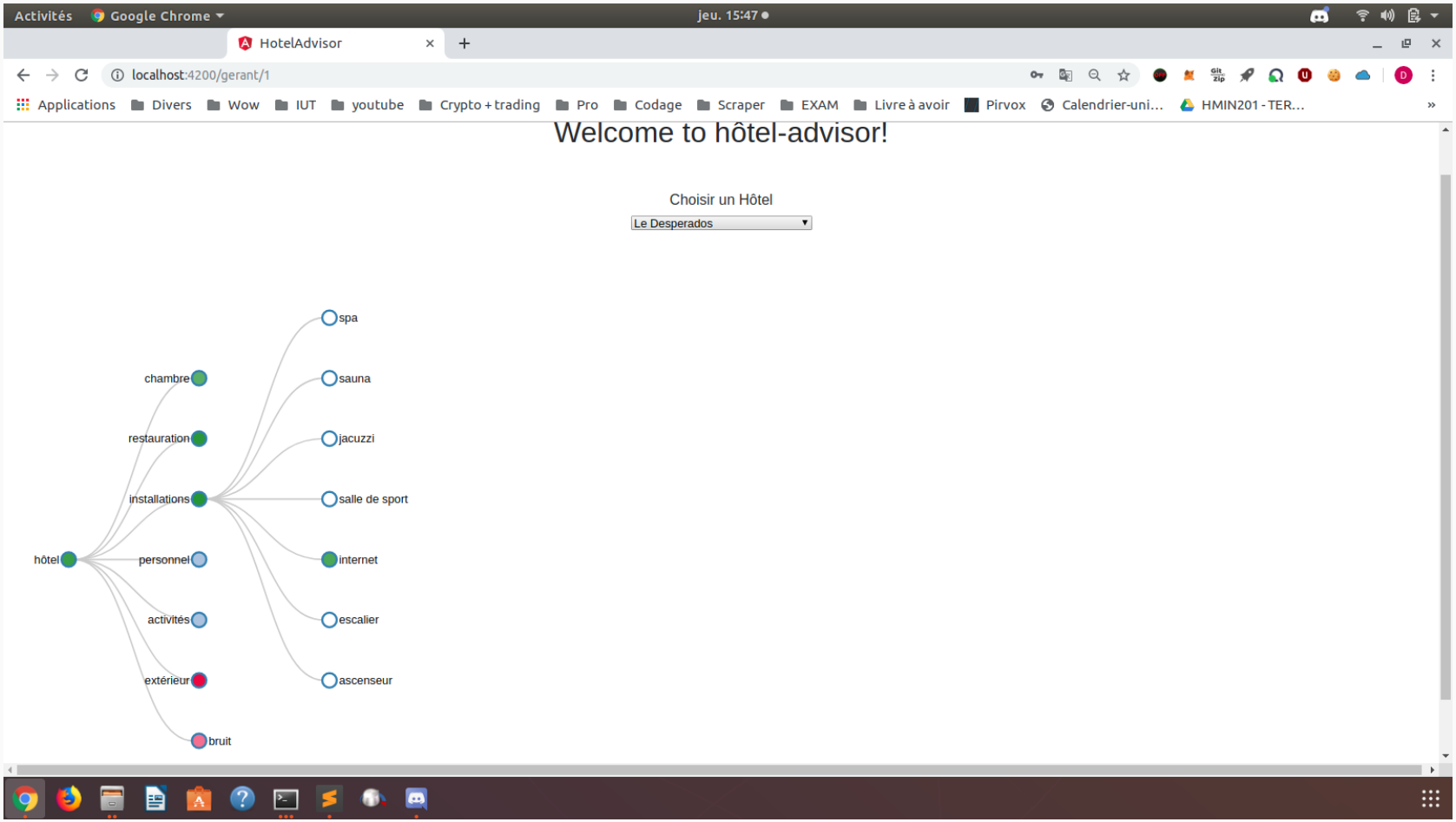
Émetteur



Admin



Ontologie


Activités
Google Chrome
jeu. 15:47
HotelAdvisor
localhost:4200/gerant/1
Applications
Divers
Wow
IUT
youtube
Crypto + trading
Pro
Codage
Scraper
EXAM
Livre à avoir
Pirvox
Calendrier-uni...
HMIN201 - TER...
Welcome to hôtel-advisor!
Choisir un Hôtel
Le Desperados
hôtel
chambre
restauration
installations
personnel
activités
extérieur
bruit
spa
sauna
jacuzzi
salle de sport
internet
escalier
ascenseur

Limites et évolutions

Limites et évolutions

Limites

- Fautes de syntaxe très difficilement réparables
- Fautes d'orthographes
- Étiquetage grammaticale imparfait

Évolutions

- Amélioration de l'interface
- Amélioration des pré-traitements
- Utilisation d'une autre propagation de polarité en utilisant un arbre syntaxique
- Gestion temporelle de l'ontologie