



*Rapport du projet EISD*

# **Infochien : un chatbot sur les chiens**

**Enseignants :** Sophie ROSSET, Thomas LAVERGNE

**Membres du groupe :** Morgan FEURTE  
Jeffrey GONCALVES  
Adrien LAVILLONNIERE  
Corentin MANSCOUR  
Hien Minh NGUYEN

**Rédigé le :** 01/03/2019

**Nombre de pages :** 15

# Introduction

Saviez-vous que le berger allemand pouvait être jusqu'à six fois plus lourd que son homologue, le berger australien ? Eh bien maintenant, oui ! Et ce, grâce à notre chatbot : **Infochien**.

Dans le cadre de ce projet, nous avons conçu un chatbot permettant d'en apprendre plus sur les chiens. Il s'appuie sur une base de donnée qui compile plusieurs types d'informations sur les chiens : leur taille et leur poids moyen, leur origine et leur utilisation. Cette base de donnée est extraite automatiquement de Wikipédia à partir d'un ensemble de patterns.

Le type de système qui nous intéresse est un chatbot mixte, sachant restituer les connaissances de la base de donnée et comparer certains éléments. Il pose également des questions dans les cas ambigus. Il peut dire bonjour et agrémente la conversation de phrases pré-écrites.

Ce rapport présentera tout d'abord le corpus utilisé, l'extraction et l'analyse de texte. Il parlera ensuite du chatbot et des différents moyens utilisés pour générer une réponse à la question de l'utilisateur. Nous avons également placé en annexe une conversation complète avec le chatbot, ainsi que le déroulement de la démonstration faite lors de la présentation finale.

# Table des matières

<b>Introduction</b>	<b>2</b>
<b>L'analyse des textes</b>	<b>4</b>
Remarques sur le corpus utilisé	4
Extraire des information	6
Statistiques de l'extraction	7
Utilité du chien	7
Taille	7
Poids	7
Origine	7
<b>Le système de dialogue</b>	<b>8</b>
Compréhension	8
Gestion du contexte	8
Distance de Levenshtein	8
Génération du dialogue	9
<b>Conclusion</b>	<b>11</b>
Conclusion et perspectives	11
Suites du projet	11
<b>Annexes</b>	<b>12</b>
Statistiques calculé dans le code	12
Déroulé de la démonstration finale	12
Un dialogue type du chatbot	13

# L'analyse des textes


## Remarques sur le corpus utilisé

Comme dit précédemment, les informations sur lesquelles se reposent notre ChatBot proviennent de diverses pages Wikipedia. Il y a au total 18 pages analysées, une par race de chien. Suite à la présentation final, ce nombre a été porté à 27, avec l'ajout des chiens possédant des données incomplètes.

Ces articles wikipedia sont construit comme suit : d'un côté il y a un texte principal, divisé en plusieurs paragraphes, et de l'autre, il y a un petit tableau regroupant des informations structurées sur le chien en question.

Chaque paragraphe est précédé d'un titre qui indique très clairement le type d'information qu'on va pouvoir y trouver ("Histoire", "Santé", "Caractéristiques physiques", etc). Le premier paragraphe de chaque article est systématiquement un paragraphe d'introduction ayant pour titre le nom de la race. Cependant les points communs (au sens strict) entre chaque article s'arrêtent là. En effet, même si les mêmes types de paragraphes ont tendance à être souvent présents, chaque article étant écrit indépendamment des autres, et chaque race recevant plus ou moins d'attention que les autres, les articles ne sont pas tous aussi fournis, et ne sont pas tous structurés de la même façon. On peut prendre comme exemple les deux articles suivants : le premier très fournis (figure 1), et le deuxième à peine rempli (figure 2).

<p>9.2 Héros de fiction</p> <p>10 Voir aussi</p> <p>10.1 Références</p> <p>10.2 Articles connexes</p> <p>10.3 Liens externes</p>	<p><b>Histoire de la race</b> [modifier] [modifier le code]</p> <p>En 1878, les éleveurs allemands de chiens à aptitudes bergères réalisent une première tentative de regroupement dans un but d'amélioration de leurs chiens. Ceux-ci sont très variés d'un point de vue phénotypique, notamment selon les régions : par exemple, le type Wurtemberg porte les oreilles droites, alors que celui de Thuringe a les oreilles tombantes. Il existe également des chiens différents en Bavière ou dans la Hesse. Mais leur principal point commun est leur caractère qui fait d'eux de bons gardiens de troupeaux et de biens : intelligence, obéissance, vigilance, incorruptibilité. On retrouve également chez tous robustesse et rusticité puisque jusqu'alors la sélection empirique qui a été réalisée ne visait qu'à obtenir de bons livres généalogiques et tentent d'établir une société, le Phylax, mais le projet n'aboutit pas.</p> <p>C'est le capitaine de cavalerie Max Emil Frédéric von Stephanitz qui sera le véritable « père » de la race. Après avoir longtemps admiré les aptitudes des chiens de berger, il décide d'en acheter un le 3 avril 1899. Son nom d'origine est Hektor von Linksrhein qu'il rebaptisera ensuite Horand von Grafath. Celui-ci est gris et jaune, plutôt de type berger de Thuringe (mais à oreilles droites). Von Stephanitz et Arthur Meyer créent ensuite le club de race le 22 avril 1899 à Karlsruhe (« Verein für deutsche Schäferhunde » ou SV), le capitaine sera à sa tête pendant trente-cinq ans.</p> <p>Le 28 septembre 1899 est publié le premier standard de la race, et en 1900, Horand inaugure le livre des origines du SV (« Zuchtbuch »).</p> <p>Ce qui fera plus tard la force de la race, c'est la largesse de ses dirigeants dans la première définition du berger allemand : « tout chien de berger vivant en Allemagne qui, grâce à un exercice constant de ses qualités de chien de berger, atteint la perfection de son corps et de son psychisme dans le cadre de sa fonction utilitaire ». Les buts sont clairs, c'est donc par et pour le travail que la sélection du berger allemand commence.</p> <p>Le club présente rapidement une activité importante et organisée. Dès 1902 un journal est édité pour tous ses membres, et en 1903 un registre de sélection voit le jour avec la compilation des performances des reproducteurs. Les progrès seront rapides car bien dirigés à l'échelon national et bien suivis par les efforts des éleveurs (1 215 membres en 1906). On commence à rechercher des femelles du même type et à organiser des expositions pour uniformiser la race, et notamment une exposition nationale d'élevage qui permet à tous les éleveurs du pays de choisir des reproducteurs homogènes.</p> <p>Peu à peu on trouve au berger allemand d'autres utilités que la garde des troupeaux (ceux-ci ayant vu leur effectif diminuer). Ses qualités de robustesse, son flair hors pair et son obéissance à toute épreuve encourageront la police allemande à l'utiliser. En 1914, le SV et l'armée organisent une démonstration des possibilités du berger allemand en temps de guerre, à laquelle il paya ensuite un lourd tribut.</p> <p>En 1922, l'examen de Köring<sup>2</sup> est mis en place pour la sélection des reproducteurs ; ceux qui sont déclarés aptes à la reproduction sont inscrits dans le registre Körbuch. En 1926, le livre des origines compte déjà 346 000 chiens inscrits. Dans les années 1950, l'épreuve du coup de feu<sup>[réf. nécessaire]</sup> et le test de caractère font leur apparition. Le SV est renommé pour le drigisme qu'il impose à l'élevage : nombre de saillies limité, choix de l'étalement en accord avec le surveillant d'élevage, interdiction de faire saillir une femelle recommandée par un mâle non recommandé, etc. Toutes ces mesures visent à guider la sélection pour le mieux.</p> <p>Katzmaier et Funk succèdent à von Stephanitz, puis le D<sup>r</sup> Rummel en 1971. En 1974, est créée l'Union mondiale des Associations de Berger Allemand (WUSV), grâce à la volonté de regroupement du D<sup>r</sup> Rummel, qui encourageait « un dialogue fructueux, aussi bien pour l'élevage que pour l'utilisation ». En 1982, Hermann Martin (élevage von Amminius) devient président du SV. Le premier championnat de travail WUSV se déroule à Münster<sup>[L. a quel ?]</sup> du 16 au 18 septembre 1988. En 1994, Peter Messier prend la tête du SV jusqu'en décembre 2002 ; lui succède alors Wolfgang Henke. Les membres du club font preuve de beaucoup de motivation et de dynamisme pour voir évoluer leur race. Ils se retrouvent très fréquemment lors de réunions pour parler de leurs chiens.</p> <p><b>En France</b> [modifier] [modifier le code]</p> <p>Dès 1910, l'importation de bergers allemands commence en France et augmente d'année en année : 4 132 chiens arrivent en France au cours du premier semestre 1912. C'est Georges Barais (élevage de Beauchamps) qui tiendra une place capitale pour le berger allemand en France. Il crée en 1913 le club du berger d'Alsace puis, en 1920, la Société du Chien de Berger d'Alsace (SCBA) qui structure réellement l'élevage. Dès sa constitution, celle-ci édite un bulletin mensuel malgré les faibles moyens de l'époque. Le 7 mars 1920 une première exposition de berger d'Alsace a lieu à Bordeaux, jugée par Georges Barais, et le compte-rendu de la journée est déjà disponible dans le bulletin du 1<sup>er</sup> avril.</p> <p>Ce n'est que le 6 octobre 1922 que les Français reconnaissent officiellement l'origine allemande de leur chien favori (ils considéraient auparavant que cette race française avait été volée par les Allemands en 1870) et la SCBA devient la Société du Chien de Berger Allemand. Elle tient son livre d'élevage et publie des pedigrees jusqu'en 1958 (création de la SCC). Lorsque Georges Barais disparaît en 1955, c'est Marcel Olive (élevage de Fort-Réal) qui lui succède.</p> <p>La SCBA a été et reste le premier club de race français. Elle aligne sa politique d'élevage sur le pays d'origine pour obtenir les meilleurs résultats possibles.</p>	<p><b>Caractéristiques</b></p> <table> <tr> <td><b>Taille</b></td> <td>60 à 65 cm (mâle) 55 à 60 cm (femelle)</td> </tr> <tr> <td><b>Poids</b></td> <td>30 à 40 kg (mâle) 22 à 32 kg (femelle)</td> </tr> <tr> <td><b>Poil</b></td> <td>Court ou long, dense et souple</td> </tr> <tr> <td><b>Robe</b></td> <td>Noire marquée de taches brun-rouge, brunes ou jaunes jusqu'à gris clair. Noir et gris unicolore (charbonné) ou noire</td> </tr> <tr> <td><b>Tête</b></td> <td>Cunéiforme, proportionnée, stop peu prononcé, truffe noire</td> </tr> <tr> <td><b>Yeux</b></td> <td>En amandes, obliques, le plus sombre possible</td> </tr> <tr> <td><b>Oreilles</b></td> <td>Droites et asymétriques, de taille moyenne, pavillon vers l'avant</td> </tr> <tr> <td><b>Queue</b></td> <td>Portée tombante déclinant une légère courbe</td> </tr> <tr> <td><b>Caractère</b></td> <td>Équilibré, courageux, sûr de lui, protecteur et joueur</td> </tr> </table> <p><b>Nomenclature FCI</b></p> <ul style="list-style-type: none"> <li>groupe 1       <ul style="list-style-type: none"> <li>section 1           <ul style="list-style-type: none"> <li>n° 166</li> </ul> </li> </ul> </li> </ul> <p><small>modifier</small></p>	<b>Taille</b>	60 à 65 cm (mâle) 55 à 60 cm (femelle)	<b>Poids</b>	30 à 40 kg (mâle) 22 à 32 kg (femelle)	<b>Poil</b>	Court ou long, dense et souple	<b>Robe</b>	Noire marquée de taches brun-rouge, brunes ou jaunes jusqu'à gris clair. Noir et gris unicolore (charbonné) ou noire	<b>Tête</b>	Cunéiforme, proportionnée, stop peu prononcé, truffe noire	<b>Yeux</b>	En amandes, obliques, le plus sombre possible	<b>Oreilles</b>	Droites et asymétriques, de taille moyenne, pavillon vers l'avant	<b>Queue</b>	Portée tombante déclinant une légère courbe	<b>Caractère</b>	Équilibré, courageux, sûr de lui, protecteur et joueur
<b>Taille</b>	60 à 65 cm (mâle) 55 à 60 cm (femelle)																			
<b>Poids</b>	30 à 40 kg (mâle) 22 à 32 kg (femelle)																			
<b>Poil</b>	Court ou long, dense et souple																			
<b>Robe</b>	Noire marquée de taches brun-rouge, brunes ou jaunes jusqu'à gris clair. Noir et gris unicolore (charbonné) ou noire																			
<b>Tête</b>	Cunéiforme, proportionnée, stop peu prononcé, truffe noire																			
<b>Yeux</b>	En amandes, obliques, le plus sombre possible																			
<b>Oreilles</b>	Droites et asymétriques, de taille moyenne, pavillon vers l'avant																			
<b>Queue</b>	Portée tombante déclinant une légère courbe																			
<b>Caractère</b>	Équilibré, courageux, sûr de lui, protecteur et joueur																			




**Figure 1** : capture d'écran de la page Wikipédia du berger allemand<sup>1</sup>.

<sup>1</sup> [https://fr.wikipedia.org/wiki/Berger\\_allemand](https://fr.wikipedia.org/wiki/Berger_allemand)

## Pumi (chien)

[Pour l'article homonyme, voir Pumi.](#)

**Cet article est une ébauche concernant les canidés.**  
Vous pouvez partager vos connaissances en l’améliorant ([comment](#) ?) selon les recommandations des projets correspondants.



Le **pumi** est une race de chiens d'origine hongroise. En 2011, sous le gouvernement de Viktor Orbán a été adoptée en Hongrie une *taxe* sur les chiens dont le pumi est exempté au motif d'être « de race hongroise », au sens de la Grande Hongrie<sup>1,2,3</sup>.

**Description** [[modifier](#) | [modifier le code](#)]


Chien de recherche et secours anciennement ressemblant à un terrier. On l'utilise aussi pour conduire le gros bétail. Son flair est bien développé. Il a fait ses preuves dans le combat contre les cambriagers et les rongeurs. C'est un excellent chien de famille qui ne peut vivre dans un appartement. Il a besoin de beaucoup de mouvement ; c'est un chien de sport parfait.

**Liens externes** [[modifier](#) | [modifier le code](#)]

- (PDP) Le standard de la race sur le site de la SCC? [[archive](#)]
- How to pronounce Hungarian dog names? [[archive](#)]
- le Pumi sur Commons

**Notes et références** [[modifier](#) | [modifier le code](#)]

- ↑ (hu) Magyar faják után nem szednek ebadókat [[archive](#)] de la chaîne privée TV2
- ↑ Une taxe qui s du chien? [[archive](#)], Vincent Baumgartner, 7 décembre 2011 Hu-Ita
- ↑ Et la Hongrie inventa la préférence nationale canine? [[archive](#)], 23 décembre 2011 Le Monde



**Pumi**

**Caractéristiques**

**Silhouette** 10-15 kg (M)idéal 12-13kg ; 8-13 kg (F)idéal 10-11kg.

**Taille** 41-47cm (M)idéal 43-45cm ; 38-44 cm (F)idéal 40-42cm

**Poil** ondulé et bouclé

**Robe** gris, noir, fauve (fakó), blanc

**Tête** longue et étroite

**Yeux** moyennement séparés en position légèrement oblique, expression vive et intelligente

**Oreilles** attachées haut et dressées, tier supérieur de l'oreille est plié vers l'avant, grandeur moyenne formant un V

**Queue** forme un large arc de cercle au-dessus de la croupe

**Caractère** ce chien de berger est assez vif, d'un tempérament infatigable, extrêmement audacieux mais méfiant avec les inconnus. Chien assez bruyant.

Nomenclature FCI

Figure 2 : capture d'écran de la page Wikipédia du Pumi<sup>2</sup>.

Ces disparités d'un article à l'autre nous ont poussés à faire des choix vis-à-vis de quels articles nous allions utiliser, et de quel type d'information nous allions extraire. Nous avons tout d'abord décidé d'exclure toute race ayant une page "vide", tel l'exemple 2 ci-dessus. Ensuite nous avons établi une liste des informations qui nous semblait apparaître le plus fréquemment :

"Historique, Description, Caractère, Santé, Utilité"

Enfin, nous avons parcouru une bonne centaine d'articles, afin de récupérer ceux qui contenaient la majorité des informations décrites par la liste ci-dessus.

Pour conclure, mêmes si ces étapes nous ont permis "d'uniformiser" un peu les différents textes qui composent notre corpus, certaines disparités subsistent encore (informations manquantes, vocabulaire différent d'un article à l'autre...).

<sup>2</sup> [https://fr.wikipedia.org/wiki/Pumi\\_\(chien\)](https://fr.wikipedia.org/wiki/Pumi_(chien))

## Extraire des information

Pour plus de simplicité dans la recherche d'attributs, nous avons fait le choix de prendre un corpus en français. Nous avons également décidé de travailler entièrement avec des lettres minuscules, la casse n'étant pas importante pour notre analyse. La dernière modification du corpus faite pour l'extraction des informations a été de rajouter un espace avant et après chaque ponctuation, pour pouvoir profiter du pattern basic de dark sur nos pipelines.

Pour extraire nos attributs dans le texte non structuré, nous avons tout d'abord décidé de nous aider des titres présents dans les différents textes de notre corpus. Par exemple, nous avons utilisé uniquement la pipeline avec le pattern servant à extraire l'origine d'un chien dans le paragraphe Histoire d'un texte.

Cette méthode s'est vite retrouvé trop complexe pour le résultat obtenu. En effet, après avoir parcouru tous nos textes, notre base de données n'était que partiellement rempli. En effet, certaines informations dans nos fichiers ne sont pas stockées dans les catégories les plus intuitives. Nous avons alors décidé d'utiliser qu'une unique pipeline, comportant tous nos patterns, sur l'intégralité des textes. Cette nouvelle façon de faire nous a permis de récupérer des informations présentes dans d'autres catégories sur wikipédia, pour une perte quasi nulle de performance.

Une fois tous nos attributs complétés dans la base de données, on compare les informations avec les données structurées. Comme les données structurées ne sont pas standardisées entre chaque race de chien, nous avons pris le parti de réutiliser les patterns des données non structurées pour les extraire. Si une information est manquante ou si elle n'est pas conforme, alors on prends l'information structurée dans la base.

Une des particularité que nous a "offert" notre corpus, est que certaines informations peuvent être contenues sous des formes différentes d'un article à l'autre. Notamment l'information Taille, qui peut dans certain cas être une mesure ("de tant à tant centimètres"), et dans d'autres cas un qualificatif ("de petite taille" ou "grand"). Pour que notre chatbot puisse répondre à des comparaisons entre divers chiens, il a fallu standardiser cette donnée. Nous avons donc créé deux pattern, un pour extraire la mesure, et un autre pour extraire le qualificatif. Et lorsque la mesure n'est pas présente nous transformons le qualificatif en une mesure correspondante (par exemple "petite taille" devient 20 cm).

Enfin, nous avons fait le choix pour ce projet d'abandonner tout ce qui avait attiré à la thématique "Santé" présente dans pas mal d'article. En effet, cette catégorie est assez vague, et nécessiterait probablement d'être découpée en plus petite catégories pour avoir du sens. De plus, il s'agit sur wikipedia d'une catégorie un peu fourre-tout : difficile d'en extraire des données correctement rangées dans une table. Une piste possible aurait été d'établir un petit lexique de problèmes de santé courant chez les chiens, et d'enregistrer ces mots dans la base de données si ceux-ci était apparu dans l'article de telle ou telle race de chien.

Pour le stockage, nous avons attribué à chaque chien une valeur pour chacune des catégories présentes, ainsi qu'une valeur "par défaut" pour les catégories non présentes ou non détectées dans le texte. On vérifiera donc dans le chatbot la présence ou non d'une information afin d'afficher quelque chose de correct à l'utilisateur.

## Statistiques de l'extraction

Nous avons relativement peu de fichiers à cause du fait que la plupart des articles de chiens de Wikipédia ne sont pas assez fournis pour extraire des informations de manière consistante. Même au sein de notre corpus de 18 chiens, certains n'ont pas de poids ou de taille (même dans les informations structurées). Cependant, pour les textes que nous avons parsés, nous avons une très grande exactitude de nos informations.

### Utilité du chien

On trouve en moyenne 2 utilités par chien dont 8 chiens qui ont pour seule utilité "chien de compagnie" (de manière explicite ou implicite : pas de catégorie utilité). Après vérification humaine, toutes les utilités extraites du corpus correspondent aux utilités réellement présentes dans le corpus.

### Taille

La taille extraite est divisée en deux métriques, une textuelle (exemple : de grande taille) et une chiffrée (X cm). On retrouve une taille chiffrée dans les données non structurées de 61% de nos fichiers. Pour les autres fichiers, on extrait la donnée des informations structurées. Lorsque la donnée existe en non structurée, on la compare avec ce que l'on trouve en structuré (un intervalle); dans 81% des cas, la donnée trouvée en structurée se trouve dans cet intervalle.

### Poids

On ne trouve le poids que pour 50% de nos fichiers dans les données non structurées, à quoi on peut ajouter 16% de poids trouvés dans les données structurées. Lorsque les deux données sont présentes, on effectue la même vérification que pour la taille, cette fois-ci 100% des résultats sont confirmés par les données structurées.

### Origine

On trouve une origine pour 88% de nos fichiers. Après vérification humaine, lorsque l'on ne trouve pas d'origine c'est qu'il n'y a pas de phrase indiquant clairement l'origine dans le fichier. Cependant, dans des cas comme celui du Bichon Maltais, il serait possible d'implémenter un lexique d'adjectif d'appartenance à un pays (exemple : australien, allemand, maltais etc.) permettant ainsi de déterminer l'origine du chien à partir de son nom, s'il possède un tel adjectif.

Cependant, pour le Bichon Maltais, il semblerait que ce soit trompeur puisque son origine semble être la Sicile.

## Le système de dialogue

Outre le fait de pouvoir parser des articles Wikipédia, le programme est également capable de restituer les connaissances via sa fonctionnalité de chatbot. A la manière d'une conversation naturelle, il répond aux questions de l'utilisateur et peut lui en poser quand il rencontre un cas ambigu.

### Compréhension

Le chatbot reconnaît les questions grâce à un système de mots-clés. Ces mots-clés sont regroupés par thèmes : `helloLexicon` (salutations), `sizeLexicon` (taille d'un chien), `useLexicon` (utilité), `originLexicon` (origine), `weightLexicon` (poids), `compareLexicon` (question comparative), `listLexicon` (lister), `qualifTaille` (qualificateur de taille : grand, petit), `qualifPoids` (qualificateur de poids : gros, lourd..), `placeList` (liste des origines) et `raceList` (liste des races).

### Gestion du contexte

Le chatbot a besoin d'un contexte pour pouvoir répondre de manière naturelle aux questions de l'utilisateur.

Lorsque l'utilisateur emploie un mot des lexiques cités ci-dessus, une fonction active la variable de contexte de ce lexique. Ces variables sont stockées dans une table (`contextTable`). Les variables de contexte se réinitialisent au bout de 3 questions de l'utilisateur, pour éviter qu'un contexte déborde trop sur le reste du dialogue. Ce contexte est ensuite utilisé pour générer les questions.

### Distance de Levenshtein

L'un des challenges principaux de ce projet a été d'inclure une détection des fautes de frappe de l'utilisateur grâce à la distance de Levenshtein.

La fonction permettant de calculer distance de Levenshtein renvoie un nombre entier positif permettant de mesurer la différence entre deux chaînes de caractères. Elle représente le nombre de caractères à changer pour passer d'une chaîne à l'autre<sup>3</sup>.

---

<sup>3</sup> [https://fr.wikipedia.org/wiki/Distance\\_de\\_Levenshtein](https://fr.wikipedia.org/wiki/Distance_de_Levenshtein)



Cette fonction est facilement trouvable sur internet. Nous avons choisi l'implémentation de Yonaba<sup>4</sup> et récupéré le fichier `levenshtein.lua` directement dans notre projet.

Faute de temps, nous n'avons pas pu trouver d'autres moyens que de comparer chaque mot de la phrase de l'utilisateur avec tous les mots présents dans les lexiques du programme. Nous aurions pu établir une méthode statistique pour détecter quel mot correspond à un nom de chien dans une question. Le désavantage de la méthode actuelle est que si le nombre de mots dans le lexique augmente, la complexité en temps de calcul de notre fonction risque d'augmenter très rapidement.

La fonction `lev` de notre chatbot appelle la fonction `lev_iter` de ce fichier, qui renvoie la distance pour chaque paires de chaînes. Cette distance est ensuite comparée à coefficient. Si elle est inférieure à celui-ci, on remplace le mot de l'utilisateur par le mot du lexique correspondant. Sinon, on passe au mot suivant. Notons également que nous ne vérifions que les mots supérieurs à 3 lettres, pour éviter de comparer les déterminants ou mots de liaison.

Nos tests montrent qu'un coefficient fixé à 2 donne des résultats corrects (c'est-à-dire qu'il peut y avoir deux lettres de différence entre les deux chaînes). `Kabrador` est ainsi correctement transformé en `labrador` (1 lettre de différence), de même que `begke` → `beagle` (2 lettres de différence). Un coefficient supérieur provoque des effets étranges : par exemple, le mot `poids` est identifié comme faisant partie du lexique de l'utilité.

## Génération du dialogue

Le chatbot utilise les lexiques et le contexte pour générer une réponse à l'utilisateur. Des éléments de lexique, comme celui de la salutation et du listage des chiens connus, sont vérifiés même avant le traitement du contexte, car les réponses sont triviales et ne dépendent que peu de l'utilisateur.

Certains contextes sont vérifiés en priorité : par exemple, la comparaison est quasiment vérifiée en premier car elle détermine énormément l'issue de la conversation qui suit. La comparaison fait aussi intervenir de multiples questions à l'utilisateur.

Deux contextes peuvent être combinés : celui de la race et celui d'une information. Par exemple, si le contexte de la race et du poids sont présents (les deux sont à `true` dans la table), alors le chatbot génère la réponse correspondant au poids du chien.

Analysons quelques exemples de questions et de réponses :

---

<sup>4</sup>[https://github.com/kennyledet/Algorithm-Implementations/blob/master/Levenshtein\\_distance/Lua/Yonaba/levenshtein.lua](https://github.com/kennyledet/Algorithm-Implementations/blob/master/Levenshtein_distance/Lua/Yonaba/levenshtein.lua)

Utilisateur : Parlons du beagle.

Le mot-clé beagle est reconnu : l'entrée du tableau des contextes pour la race est mise à True et la valeur de la race à "beagle". Le chatbot génère une réponse appropriée du type :

Infochien : Parlons du beagle.

Utilisateur : Quel est son poids ?

Le mot-clé poids est reconnu : l'entrée du tableau des contextes pour le poids est mise à True.

Le contexte de la race est gardé en mémoire pour encore deux tours. Le chatbot sait donc de quel chien on parle : du "beagle". Il peut donc générer une réponse correspondant à la combinaison de beagle et de poids :

Infochien : Le poids du beagle est 17kg.

Il en est de même pour les autres combinaisons de contextes. Le chatbot est capable de retenir jusqu'à deux types de chiens (stockés dans les variables globales previousRace et currentRace). Voici un exemple de dialogue impliquant deux races de chiens, stockées dans ces variables, ainsi qu'un contexte de comparaison :

Utilisateur : Je veux en savoir plus sur le beagle et le labrador retriever

Infochien : Parlons du beagle et du labrador retriever.

Infochien : Des questions sur les chiens ?

Utilisateur : quel est le plus grand ?

Infochien : Comparons le beagle et le labrador retriever.

Infochien : Le labrador retriever (de taille moyenne) est plus grand que le beagle (de taille petite à moyenne).

Le chatbot est de type mixte, c'est-à-dire qu'il peut également poser des questions. Par exemple, la question précédente peut être interprétée de manière différente lorsque le contexte des races est vide. Dans ce cas, le chatbot pose la question pour lever l'ambiguïté :

Utilisateur : Quel est le plus grand chien ?

Infochien : Voulez-vous comparer deux chiens (dites 'ouaf'), ou faire une comparaison absolue (dites 'ouaf ouaf') ?

Utilisateur : ouaf ouaf

Infochien : Le plus petit chien de la BD est le bichon maltais.

Infochien : Voulez-vous en savoir plus ?

Ici, le chatbot détecte le contexte de la comparaison grâce aux mots-clés le plus grand. Il n'a pas de contexte de race et demande donc si l'utilisateur veut comparer deux chiens, ou trouver le plus grand chien de la base de donnée (BD). Pour l'instant, c'est le seul cas de question qui a été implémenté, mais d'autres types d'ambiguïtés pourraient se résoudre avec ce processus.

## Conclusion

### Conclusion et perspectives

Ce projet nous a tout d'abord permis de nous familiariser avec le langage Lua, appris à la volée durant ces deux derniers mois. Similaire au Python, il propose une approche intéressante en ce qui concerne les tableaux.

Il nous a également été possible d'en apprendre plus sur l'extraction de données à partir de patterns et les difficultés qui accompagne cette méthode. Quant à la partie chatbot, nous avons déjà travaillé sur un chatbot en ET4 : nous avons donc pu consolider nos acquis et réfléchir entre autre sur la gestion des contextes.

### Suites du projet

A l'heure actuelle, notre méthode pour appliquer la distance de Levenshtein est obsolète : nous pouvons l'optimiser avec une méthode statistique moins gourmande en calculs. Nous avons également laissé de côté de nombreux attributs tels que la santé ou le caractère du chien. Afin d'améliorer le projet, il faudrait donc parser le reste des attributs.

Il serait également possible d'ajouter des options au chatbot : des questions du type "donne-moi les chiens avec tels attributs" ou "quels sont les chiens de plus de x kg ?" pourraient être intéressantes à poser pour l'utilisateur.

## Annexes

### Statistiques calculé dans le code

```
{
  use = {
    average = "2 uses per dog",
    ["pet_dogs"] = "8 dogs with only use : pet",
  },
  height = {
```

```
    successes = "81% accuracy (comparing non-structured heights with  
structured heights)",  
    structured = "found 7 new heights in structured data",  
    found = "61% on 18 dogs",  
  },  
  weight = {  
    successes = "100% accuracy (comparing non-structured weights with  
structured weights)",  
    structured = "found 3 new weights in structured data",  
    found = "50% on 18 dogs",  
  },  
  origin = {  
    found = "88% on 18 dogs"  
  }  
}
```

### Déroulé de la démonstration finale

#### **1. Est-ce que tu peux me lister tous les chiens dont tu peux me parler?**

\*liste de tous les chiens\*

#### **2.Quel est le chien le plus grand**

\*le lévrier afghan\*

#### **3.Chien le plus léger ?**

\*le bichon maltais\*

#### **4.Je voudrais en savoir plus sur le beagle**

\*parlons du beagle\*

#### **5.Quelle est sa taille?**

\*sa taille\*

#### **6.Entre ce chien et l'akita, quel est le plus grand?**

\*akita est plus grand que beagle\*

#### **7.L'utilité du labrador retriever ?**

\*liste d'utilités\*

#### **8.Combien pèse ce chien ?**

\*inconnu\*

#### **9.Et le lévrier afghan?**

\*son poids \*

## 10.D'où vient-il ?

\*Afghanistan\*

## 11.Quels sont tous les chiens qui viennent d'Allemagne?

\*liste des chiens venant d'Allemagne\*

### Un dialogue type du chatbot

Le dialogue suivant provient directement du chatbot. Les lignes commençant par Infochien sont générées par le chatbot.

Infochien : Bonjour je suis un chienbot ! Je peux parler de beaucoup de chiens.

Bonjour ! Peux-tu me donner une liste de tous les chiens que tu connais ?

Infochien : Bonjour !

Infochien : Voici la liste de tous les chiens de la base de donnée :

Le dalmatien, le akita, le golden retriever, le bouledogue, le epagneul breton, le chien de berger islandais, le samoyède, le lévrier afghan, le chow-chow, le bichon maltais, le terre-neuve, le lapphund suédois, le beagle, le spitz allemand, le labrador retriever, le berger allemand, le terrier écossais, le dogue de bordeaux sont dans la base de donnée.

Infochien : Des questions ?

quelle est la taille du chien de berger islandais ?

Infochien : Le chien de berger islandais est un chien de 46 cm.

Infochien : Ouaf.

et celle du beagle stp ?

Infochien : Le beagle est un chien de taille petite à moyenne (41 cm).

Infochien : Je sais tout sur les chiens !

Et celle du labrador retriever ?

Infochien : Le labrador retriever est un chien de taille moyenne (57 cm).

Infochien : Quelque chose d'autre ?

Et le poids du samoyède ?

Infochien : Désolé, mais je n'ai pas cette information.

Infochien : Des questions ?

et sa taille ?

Infochien : Le samoyède est un chien de taille moyenne (57 cm).

Infochien : Voulez-vous en savoir plus ?

et son origine ?

Infochien : L'origine du samoyède est : Russie européenne.

Infochien : Des questions sur les chiens ?

Quel est le plus grand chien entre le beagle et le samoyède ?

Infochien : Comparons le beagle et le samoyède.

Infochien : Le samoyède (de taille moyenne) est plus grand que le beagle (de taille petite à moyenne).

Infochien : Je sais tout sur les chiens !

Quel est le plus grand chien ?

Infochien : Voulez-vous comparer deux chiens (dites 'ouaf'), ou faire une comparaison absolue (dites 'ouaf ouaf') ?

ouaf ouaf

Infochien : Le chien le plus grand de la BD est le lévrier afghan.

Infochien : Voulez-vous en savoir plus ?

et le plus gros ?

Infochien : Voulez-vous comparer deux chiens (dites 'ouaf'), ou faire une comparaison absolue (dites 'ouaf ouaf') ?

ouaf ouaf

Infochien : Le chien le plus gros de la BD est le dogue de bordeaux.

Infochien : Quoi d'autre ?

Quel est le plus gros, entre le dogue de bordeaux et le dalmatien ?

Infochien : Comparons le dogue de bordeaux et le dalmatien.

Infochien : Le dogue de bordeaux (80 kg) est plus lourd que le dalmatien (32 kg).

Infochien : Je sais tout sur les chiens !

quel est le plus petit des chiens ?

Infochien : Voulez-vous comparer deux chiens (dites 'ouaf'), ou faire une comparaison absolue (dites 'ouaf ouaf') ?

ouaf ouaf

Infochien : Le plus petit chien de la BD est le bichon maltais.

Infochien : Ouaf.

d'où vient le labrador retriever ?

Infochien : L'origine du labrador retriever est : Royaume-uni.

Infochien : Voulez-vous en savoir plus ?

et le golden retriever ?

Infochien : L'origine du golden retriever est : Britannique.

Infochien : Quoi d'autre ?

Quels sont les chiens provenant d'Allemagne ?

Infochien : Les berger allemand, les spitz allemand viennent d'Allemagne.

Infochien : Des questions sur les chiens ?

et du Japon ?

Infochien : Les akita viennent de Japon.

Infochien : Voulez-vous des informations sur les chiens ?

quitter

Infochien : Au revoir ! Ouaf !