

Voici votre titre

Les auteurs - c'est vous

La date

1 Ceci est une section

Vous pouvez créer des sections, des sous-sections, des sous-sous-sections et des paragraphes. Après la compilation, tout sera numéroté comme il faut ! Vous pouvez faire du **texte en gras** ou du *texte en italiques* ou en **typewriter**.

On peut aussi changer la taille du texte. Les tailles prédéfinies sont : `tiny`, `scriptsize`, `footnotesize`, `small`, `normalsize`, `large` et `Large`.

On peut changer la couleur du texte **comme ça**.

1.1 Ceci est une sous-section

Voici pouvez faire des listes avec des puces :

- Liste
- Non
- Numérotée

Ou des listes numérotées :

1. Liste
2. Numérotée

Et même des listes imbriquées :

1. Chien
 - Caniche
 - ◇ Fluffy
 - Candyfloss
 - Yorkshire
2. Chat
3. Oiseau
 - Moineau
 - Aigle

1.1.1 Ceci est une sous-sous-section

Voici un tableau :

Left	Centre	Right
Voici du texte	Beaucoup de texte	Et Encore du texte
a	b	c

Le même tableau mais centré !

Left	Centre	Right
Voici du texte	Beaucoup de texte	Et Encore du texte
a	b	c

Ceci est un paragraphe Pour faire des guillemets, regardez bien les caractères écrits dans la source L^AT_EX :

- « Des guillemets à la française ».
- “Des guillemets doubles à l’anglaise”
- ‘Des guillemets simples’

Pour afficher les caractères spéciaux :

- accolades } et {
- signe de pourcentage %
- anti-slash \
- tilde : ~ or ~

Insérez une image comme suit :

Tout petit et centré →

2 Maths

Pour faire une équation simple, utilisez l’environnement maths. Vous pouvez faire des maths inline comme suit : $a = \sum_{i=0}^N x_i + c$ ou comme suit :

$$a = \sum_{i=0}^N x_i + c + \frac{1}{2} + \text{some more stuff} \dots$$

Pour faire des équations alignées :

3 Et plein d’autres choses !

Quelques templates : <https://www.latextemplates.com>

Documentation : <https://www.latex-project.org>

Introduction

Le sujet qui a été choisi est celui du chatbot (ou assistant personnel). Explications sur le chat bot

L’ensemble des membres du groupe était relativement motivé par ce sujet, du chatbot. Le thème restait toutefois à choisir : pour ne pas avoir un simple robot générique

Difficultés

le corpus : trouvé un corpus qui permette de mieux répondre

4 Stratégie

La stratégie envisagée est assez simple. Elle consiste à récupérer de très nombreux corpus de textes, et d'identifier dans ces écrits les couples de questions-réponses. Ensuite, une question est attendue de la part de l'utilisateur. Cette question est analysée, puis à partir d'un dictionnaire, la réponse la plus adaptée est déterminée et affichée. Pour assurer la qualité de ce chatbot, plusieurs conditions sont à assurer :

- la capacité du programme à correctement comprendre la question de l'utilisateur. Dans ce but, les phrases entrées sont divisées en ensembles linguistiques, eux-mêmes étiquetés puis pondérés. Cette pondération (cf. figure 1) est particulièrement importante, puisque c'est elle qui permet de sélectionner la réponse la plus appropriée parmi celles possibles ;
- un nombre conséquent de couples de questions-réponses doit être disponible. En effet, pour répondre selon le token indiquant la question, il faut grouper l'intégralité des couples de question-réponse selon le ce token. En pratique, pour une question précise, cela divise la taille du corpus de texte utilisé par un facteur six (voire plus) ;
- d'autres critères ? J'en avais en tête, mais je les ai oubliés.
- Idem.

Type de token	Poids	Exemples
Questions	2	"When", "What", "Who", "Where", "Why", "How"
Verbes	1.2	"Borrow", "Expand"
Noms	0.7	"John", "Doe"
Pronoms	0.7	"You", "Themselves"
Adjectifs	0.25	"Awesome", "Magnificent"
Adpositions	0.25	"From", "With"
Adverbes	0.25	"Certainly", "Even"
Autres	0	"?", "mistake", "666", ...

FIGURE 1 – Tableau de pondération des éléments syntaxiques

4.1 Les différentes pondérations

Il paraissait évident d'attribuer différent poids aux différents, mais pourquoi de tels choix de pondération ? Voici une explication selon les types de tokens identifiés par *NLTK* :

Questions : Détermine la nature de la réponse et sa tournure principale. C'est pourquoi la pondération de ce type d'éléments a été majorée ;

Verbes : Fixe le type d'action et permet généralement d'identifier l'action ou l'évènement concerné(e) par la question. Il permet de réduire correctement les réponses possibles : c'est pourquoi il s'agit de la seconde plus grande pondération ;

Noms : Un nom fait souvent référence à un contexte précis, ou tout simplement à un interlocuteur, parfois direct. C'est pourquoi il est important. Toutefois, si un nom n'est pas trouvé dans la base de données, il devient immédiatement sans importance. Cela devrait être fréquent, puisqu'il existe une infinité de noms propres. De plus, certains noms sont beaucoup plus fréquents que d'autres. Si une importance trop grande était accordée aux noms, alors une réponse en rapport avec ce nom, mais totalement hors-sujet pourrait être fournie ;

Pronoms : Les pronoms sont bien plus permissifs que les noms en termes de réponses. C'est pourquoi leur poids n'est pas non plus extrêmement important. De plus, ils sont utilisés dans beaucoup de contextes, sans forcément avoir une importance majeure. Par exemple, lors d'une tournure rhétorique anglaise "..., don't you ?", l'information principale de la phrase est contenue dans la partie initiale de la phrase. Bien

qu'une réponse courte du type "I do" soit valide, d'un point de vue conversationnel, il est plus intéressant de s'intéresser aux autres éléments de la phrase ;

Adjectifs : La majorité des adjectifs peuvent être remplacés par d'autres adjectifs sans forcément changer le sens de la phrase. C'est pourquoi ils ont un poids relativement faible. Néanmoins, la présence d'un champ lexical pour connaître les adjectifs permettrait d'élargir le champ de recherche des réponses valides en évitant en même temps les adjectifs qui changent effectivement le sens de la phrase ;

Adverbes : Les adverbes correspondent à un cas relativement similaire à celui des adjectifs, expliquant leur poids faible ;

Adpositions : Les adpositions correspondent généralement à un complément d'information à l'action. Action qui reste l'un des éléments principaux de la phrase. Les adpositions ont donc une pondération faible ;

Autres : Les autres types de tokens comprennent les conjonctions, les déterminants, les nombres cardinaux, les particules, les mots issus d'une autre langue et enfin les fautes de frappe sont actuellement très peu importants à la réponse. En guise d'illustration, changer un nombre de valeur n'affecte que très peu le sens de la phrase (bien que les conséquences puissent être totalement différentes selon le contexte).

Les signes de ponctuation sont volontairement omis (pondération nulle) puisque l'entrée doit être similaire à une question pour que le programme trouve une réponse adaptée.

4.2 Une capture d'écran puis suite ?

Thème

Explication, on veut faire un chatbot avec pour thème D&D, pour cela on change le vocabulaire. Ce qu'on aurait pu faire et comment

si vous avez des idées de quoi et comment le faire : garder un lien entre les questions d'une conversation.

Champ lexical pour adjectifs et noms. Distance de Levenshtein pour les fautes de frappe, associées à un dictionnaire.