**Synthèse: métriques d’évaluation d’un chatbot**

Ce document contient un bilan du travail de recherche sur les métriques utilisées dans l’industrie pour l’évaluation des performances des robots conversationnels, réalisé dans le cadre de la phase 5 du projet.

A l’issue de cette étude, une liste des métriques les plus significatives sera rédigée. Parmi ces métriques, certaines seront retenues pour l’évaluation du chatbot développé. Le choix des métriques retenues repose sur les critères suivants:

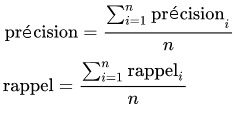
* **Pertinence de la métrique** dans le cadre de l’application développée et du service final (chatbot tourisme, service proposant des itinéraires/vols aux utilisateurs à partir de conversation avec ceux-ci, service se positionnant comme un apporteur d’affaires pour les sites de réservation)
* **Accessibilité de la métrique**: c’est-à-dire, est ce qu’il est facile de mesurer cette quantité? Certaines métriques, bien que pertinentes, nécessitent par exemple l’intégration du chatbot développé dans le service actuel pour être mesurées (durée moyenne d’une conversation, durée moyenne avant proposition du chatbot, etc), et sont, de ce fait, pas accessibles aux membres de l’équipe

**================================================================**

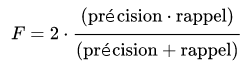
**Technical Metrics**

* **Précision:** La précision pour une classe **i** (une entité, un intent) est le **nombre de classifications correctement attribuées à la classe i** rapporté au **nombre total de classifications attribuées à la classe i**
* **Recall: (Rappel)** Le rappel pour une classe **i** est le **nombre de classifications correctement attribuées à la classe i** rapporté au **nombre total d’éléments appartenant à la classe i proposés au chatbot**

Dans le cadre multiclasse (plusieurs entités par exemple), ces deux valeurs sont calculées en moyennant sur les valeurs correspondants à chaque classe:



* **F-Score:** Mesure populaire qui **combine la précision et le rappel**. C’est leur moyenne harmonique.



L’utilisation de cette dernière mesure (F-score) est plus judicieuse et reflète mieux les performances d’un chatbot (ou un programme de classification). Il est possible d’avoir un programme ayant une précision presque parfaite, on croirait ainsi que le programme est performant. Cependant, ce même programme peut avoir un rappel très faible, et est de ce fait loin d’être parfait. La F-mesure est plus significative, puisqu’elle prend en compte les deux métriques précédentes. **Il est ainsi intéressant d’étudier la F-mesure pour connaître la performance globale du chatbot, mais aussi d’étudier les deux autres métriques pour mieux comprendre son comportement et cibler les éléments à améliorer.**

Ces métriques sont calculées à partir d’un jeu de données. Le jeu de données est généralement séparé en jeu d'entraînement, et jeu de test, ce dernier contenant des exemples que le chatbot n’a jamais vu. La robustesse du chatbot est mesurée en évaluant la performance de celui-ci sur le jeu de test. Cela permet d’évaluer à quel point le chatbot réussi à capturer les patterns qui nous intéressent, de manière générale, et vérifier l’absence de surapprentissage. **Ces valeurs sont mesurées sur un dataset donné. Elles dépendent donc du choix et de la construction du dataset.**

Les métriques suivantes reflètent l’engagement des utilisateurs avec le chatbot, et peuvent donc servir de proxy pour évaluer la qualité et la performance du chatbot dans le cas d’usage pour lequel il a été pensé.

**User Metrics**

* **Total users:** Nombre de personnes qui utilisent le chatbot. Mesure intéressante à connaître, mais pas très utile à elle seule (sans d’autres user metrics)
* **Active users:** Utilisateurs qui lisent un message du chatbot sur une durée donnée (de manière régulière)
* **Engaged users:** utilisateurs qui communiquent avec le chatbot régulièrement
* **New users:** Cette mesure permet d’évaluer une campagne promotionnelle de la startup mais aussi d’analyser les tendances de croissance de la startup

Les métriques suivantes reflètent l’efficacité et la performance du chatbot dans le contexte des objectifs de la startup: proposer des itinéraires et des vols adaptés aux utilisateurs afin que ceux-ci cliquent dessus.

**Bot Metrics**

* **Retention rate:** le pourcentage des utilisateurs qui ont recours plusieurs fois au chatbot, après une première utilisation
* **Goal Completion rate:** Le pourcentage des utilisateurs ayant eu une conversation avec le chatbot et qui ont atteint leurs objectifs. **Typiquement, pour Flybot, cela se traduit par le pourcentage des utilisateurs qui cliquent sur un lien proposé par le chatbot à l’issue d’une conversation.**
* **Goal Completion Time:** Pour évaluer la qualité et la compétitivité du service, il faut s’intéresser au **GCR** (métrique précédente) mais aussi à la durée moyenne d’une conversation réussie (click-through à la fin).

**================================================================**

Pour évaluer notre travail durant ce projet, nous allons retenir quelques métriques parmi celles listées ci-dessus. Toutes ces métriques sont pertinentes et contiennent des informations utiles. Cependant, certaines ne sont pas accessibles ou sont difficilement mesurables par notre équipe (bot metrics and user metrics), et il n’est pas pertinent de s’y intéresser à ce stade du développement.

Ainsi, nous nous focaliserons sur les trois premières métriques: **précision, rappel et F-score** pour l’évaluation du chatbot développé. Ces métriques seront mesurées à partir d’un jeu de données contenant principalement des phrases liés au domaine du tourisme, et en particulier à la recherche de vol.

Ci-dessous le détail des méthodologies d’évaluation du chatbot développé.

**================================================================**

**Méthodologies d’évaluation et de validation**

* **Construction d’un dataset pertinent**

Notre équipe commencera par la mise en place d’un jeu de données pertinent pour l’application du client, c’est-à-dire contenant des exemplaires liés au domaine du tourisme et plus spécifiquement à la réservation de vols, et de bonne qualité, c’est-à-dire avec une bonne distribution des exemplaires choisis. Pour cela, nous utiliserons deux approches: le prétraitement du fichier logs.json fourni par le client, la génération combinatoire de phrases par algorithme.

Le dataset contiendra des exemplaires et les annotations **ground truth.**

* **80% / 20% split**

Le dataset sera séparé en une partie pour l’entraînement, et une partie pour le test. L’évaluation sur le jeu de test permettra d’évaluer les performances mais aussi la robustesse du chatbot et de vérifier qu’il n’y a pas eu de surapprentissage.

* **K-fold cross validation**

La validation croisée K-fold consiste à séparer le dataset en K parties, utiliser une partie pour le test et les K-1 restants pour l’entraînement, et répéter ce processus K fois en choisissant à chaque fois une nouvelle partie parmi les K pour le test. Cette méthode permet d’obtenir des valeurs plus fiables sur les performances du chatbot.