**Synthèse: étude exploratoire du fichier logs**

Le fichier de données brutes *logs.json* contient la totalité des textes reçus par le chatbot, provenant des utilisateurs de Flybot, ainsi que les entités et “intents” qui en étaient extraites. Au total, le fichier contient plus de 17M de lignes.

L’objectif de cette première étude est d’évaluer la qualité des données textuelles contenues dans le fichier (taille des messages reçus, précision/taux de confiance liés à chaque entité détectée, variance et distribution des messages reçu, etc). Cette étude permettra d’**effectuer des choix judicieux pour le prétraitement des données** et de **décider quant à l’utilité ou non d’augmenter le jeu de données d'entraînement**.

**1 - Informations générales**

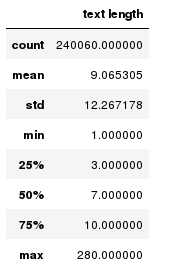
* Le fichier contient **17 million de lignes (et non 17M de textes de conversation)**
* Les 17M de lignes représentent les outputs (sous format JSON) du chatbot après traitement des textes/messages reçus. Ces messages sont au nombre de **240 060**. Ainsi, nous disposons d’un jeu de données de **240 060 exemplaires annotés** (entrée + sortie du chatbot). Est ce que ces annotations sont correctes? Est ce que la distribution des données est bonne pour atteindre les performances souhaitées?

* **Format output: exemple**

{'**at'**: '2018-07-26T10:17:36-07:00',  
 **'conclusion'**: {'\_text': 'casablanca',  
 'outcomes': [{'\_text': 'casablanca',  
 'confidence': None,  
 'entities': [{'body': 'casablanca',  
 'confidence': 0.93602,  
 'end': 10,  
 'name': 'location',  
 'start': 0,  
 'value': {'type': 'value', 'value': 'casablanca'},  
 'wisp': '53679cd0-314c-499a-915c-d0879ffc028d'}],  
 'intent': 'default\_intent',  
 'intent-id': 'default\_intent\_id'}]},  
 **'context'**: {'timezone': 'UTC'},  
***'diag': {'\_text': 'casablanca',  
 'outcomes': [{'\_text': 'casablanca',  
 'confidence': None,  
 'entities': {'location': [{'body': 'casablanca',  
 'confidence': 0.93602,  
 'end': 10,  
 'entity': 'location',  
 'start': 0,  
 'value': {'type': 'value', 'value': 'casablanca'}}]},  
 'intent': 'default\_intent',  
 'intent-id': 'default\_intent\_id'}]},*** **'from'**: 'msg',  
 **'instance-id'**: '5b5892d0-f12f-44e0-a9d3-efb5ece2f0d6',  
 **'instance-name'**: 'flybot-refactor',  
 **'uid'**: 'github:10999207',  
 **'user-instance'**: 'revfyawo\_ece2f0d6',  
 **'version'**: '20160516'}

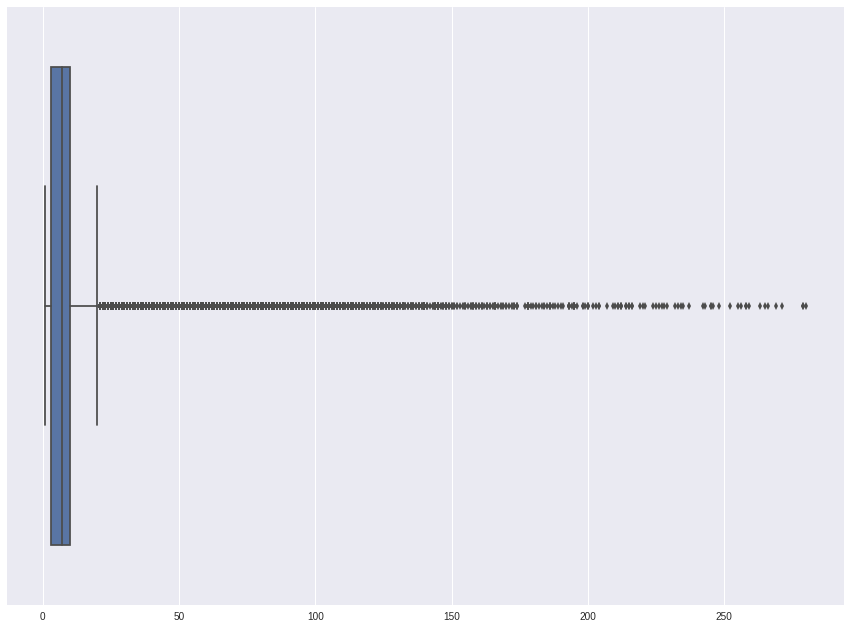
En vert italique la partie de la sortie qui nous intéresse et qui peut servir à la construction du fichier d’entrainement.

**2 - Analyse de la taille des messages**



La taille des messages reçus par le chatbot varie entre **1 caractère** (un chiffre, 1,3, par exemple) et **280 caractères** (une phrase/long texte). Cependant, la distribution des messages sur cette fenêtre (1 - 280) est très dissymétrique (cf. tableau ci-contre). **Pour les trois quarts des textes reçus par le chatbot, la taille des messages ne dépasse pas 10 caractères (un mot?)**.

Le boxplot ci-dessous reflète cette dissymétrie dans la distribution des tailles des messages reçus:

**Figure 1:** boxplot taille des messages reçus par Flybot. La majorité des messages ne dépassent pas 30 caractères. 

Cette faible présence de messages longs pose un premier problème pour l'entraînement du chatbot avec ce jeu de données.

Ce phénomène peut être justifié par le fait que le chatbot propose aujourd’hui 2 façons pour communiquer avec les utilisateurs: par messages directs des utilisateurs ou en proposant des choix aux utilisateurs qui sont amenés à cliquer sur une réponse (parmi une liste de choix), leurs réponses (ou leurs messages) sont donc des mots.

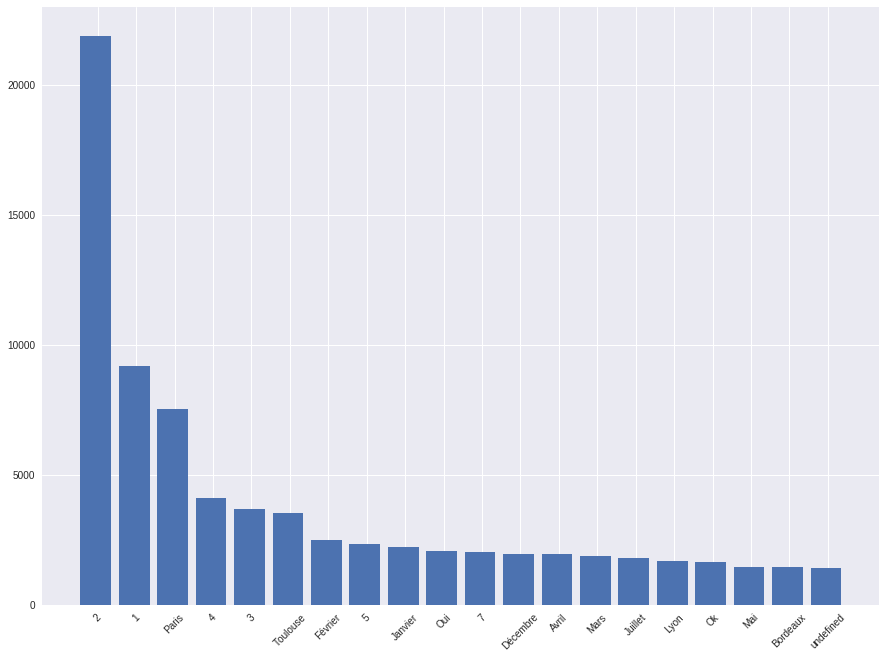
**3 - Analyse des redondances dans les messages**



Le jeu de données contient également beaucoup de redondances.

Le tableau ci-contre illustre la redondance de certains messages qui, de plus, sont des mots simples (une chaîne de caractère ne dépassant pas 10 caractères). Par exemple, le message le plus présent dans le dataset est **“2”** (le chiffre), qui représente 10% des 240k messages.

La figure ci-dessous est un histogramme des chaînes de caractères les plus fréquemment retrouvées dans le jeu de données.

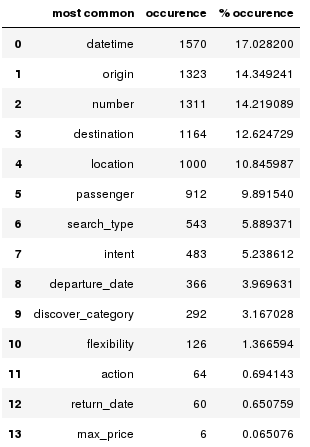
**Figure 2:** messages les plus fréquemment retrouvés dans le jeu de données initial (logs.json). La totalité de ces messages sont des mots.

**Bilan 1**

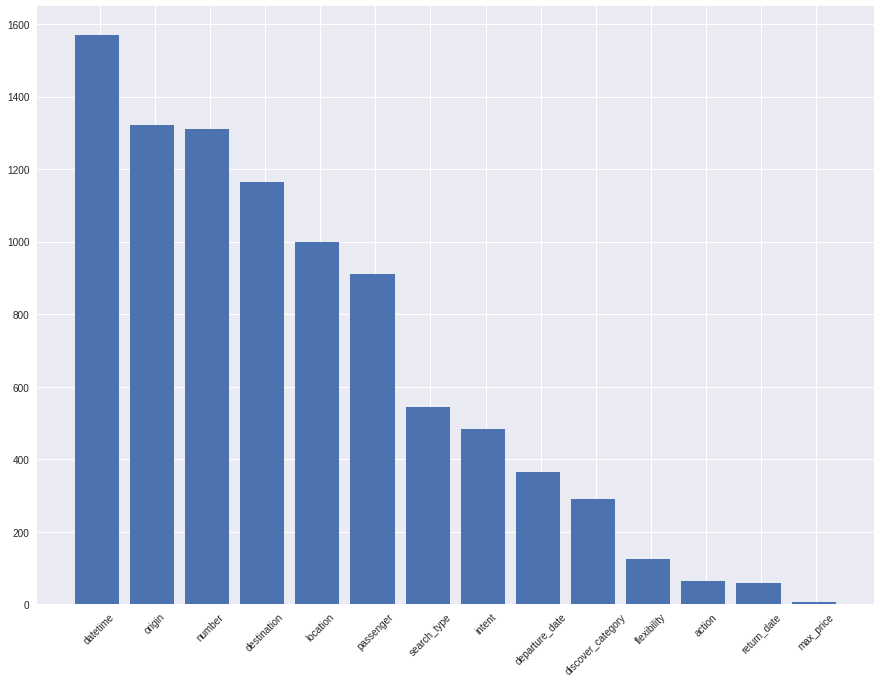
A partir de cette première étude, il est apparent que le fichier *logs.json* ne peut pas servir d’unique jeu de données d’entraînement pour le chatbot pour atteindre de bonnes performances. Les messages dont la taille dépasse 50-60 caractères (des phrases complètes et pas des mots ou des parties de conversation) seront analysés dans l’étape qui suit pour conclure quant à leur utilité pour l’entraînement du chatbot. Quelque soit le résultat de cette deuxième analyse, il est impératif de construire d’autres jeux de données pour entraîner le chatbot. Les pistes recommandées par notre équipe sont: **la génération d’un jeu de données à partir des entités spécifiées par le client et en utilisant un algorithme, l’utilisation de jeux de données open source.**

**4 - Analyse des phrases (plus de 50 caractères) reçues**

* Le nombre de messages dont la taille dépasse 50 caractères est **3771 messages.**
* Sur ces messages, **9220 entités** ont été détectées
* **Les entités les plus détectées/présentes dans ces messages:**

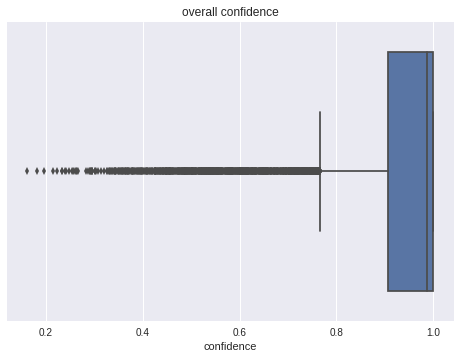


Les entités les plus détectées sont les dates, les destinations/origines et les nombres/passagers. On s’intéressera par la suite au taux de confiance liés à ces entités. On peut dès à présent mieux cadrer la phase de data augmentation: **se focaliser sur des entités peu détectées dans le jeu de données actuel**.

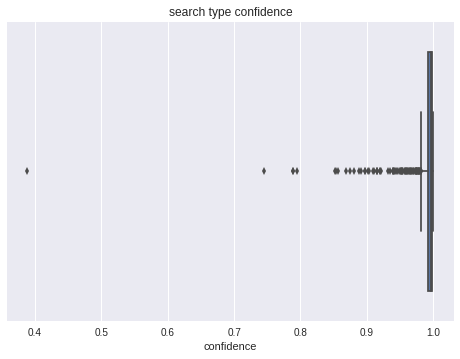


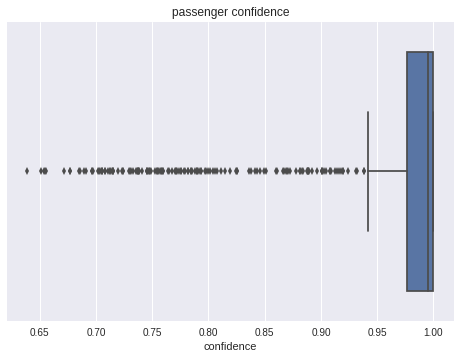
**Figure 3:** les entités les plus détectées dans les messages dont la taille dépasse 50 caractères.

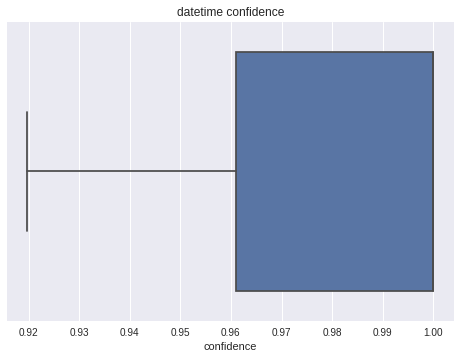
La variation du *confidence rate*, toute entité confondue:

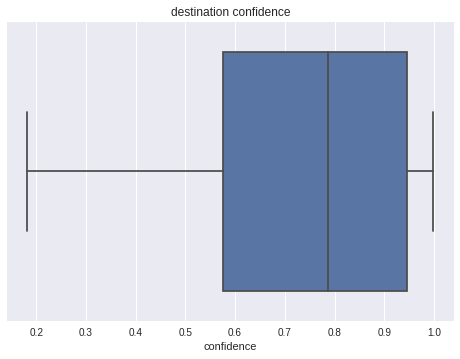


On s’intéresse ensuite au taux de confiance, en fonction de l’entité détectée:









Sur les entités **origin, destination**, le *confidence rate* varie significativement (de à 0.2 à 0.8).. Sur la totalité des autres entités, le taux de confiance est élevé (supérieur à 0.8).

Le fichier source contient des programmes pour analyser en détails ces quantités.

**Bilan 2**

Le fichier logs contient 3771 exemplaires qui peuvent être utilisés pour l’entraînement du nouveau chatbot. Cependant, le jeu de données d’entraînement ne peut pas être réduit à ces exemplaires. En effet, certains outputs sont erronés, les entités ne sont pas représentées de manière uniforme (1570 occurrence de *datetime*, 912 pour *passenger*, 292 pour *discover\_category*, etc.), et ce nombre d’exemplaires (3771) reste faible pour entraîner un chatbot et atteindre de bonnes performances.

Ainsi, les solutions retenues pour la suite du projet:

* **La génération d’un jeu de données à partir des entités spécifiées par le client et en utilisant un algorithme. Pour chaque message généré, l’output *ground truth* sera généré aussi. On obtient ainsi un jeu de données structuré**
* **Prétraiter et inclure les 3771 exemplaires dépassant 50 caractères dans le jeu de données généré**
* **L’utilisation de jeux de données open source**

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

Le notebook **explore\_logs.ipynb** contient les codes qui ont permis d’effectuer cette étude et de générer les différents graphes.