CETM 46: Data Science Product Development
Assignment 2
A Domain Specific Data Science Product Development Project: Product Prototype and Project Report

[Chatbots in the hotel industry]

Teacher: Ming Jiang
Student Registration ID: 209334704
Daniyal Rizwan Masood Butt
Word Count: 2043

Introduction

This report details design and development of a custom data science product for a data science start-up company. A chatbot for the hotel industry is developed that can conduct conversations with the user by providing answers to frequently asked questions. A chatbot is software that can conduct conversations with human users in natural language. Chatbots can answer customer's questions about services and products. They aim to replace human staff doing the same job. They cannot hold completely natural conversations and therefore in case advanced support is required the questions can be forwarded to a human agent. (X Luo et al. 2019)

The end product is provided in the form of html web page that allows users with little to no knowledge to access the chatbot within their web browser and simply type in their questions and get answers. The report discusses three sections of the development process i.e. product design, product development and project management. Product design mentions data sources, functional/non-functional requirements, use case and overall design of the product. Product development mentions the programming languages, libraries, software methodology and testing approach. Finally the time management strategy for the project is mentioned. The report is concluded with a brief discussion of further work & advanced techniques.

Product design

An initial design concept is important as it helps plan the product development strategy and the final product. (Elliot 2000) Therefore an initial design for the chatbot was created to facilitate the development process. The front-end view of the chatbot is a simple layout html and css page. The page also has links to an about and feedback section. The about section provides details about data source

and possible conversation topics while the feedback form can be used by the user to provide information on their experience. Once sent the feedback form data is emailed to the private email of the developer for follow up and a thank you message is produced for the user.

The chatbot created for this assignment is a task based chatbot which aims to solve conversational tasks with regard to customer queries within a hotel environment. These chatbots are also known as intent based chatbots as they work by identifying intents of the user from the user text and select a response based on that intent. In one instance the chatbot can chat with one single user and the quality of the responses is dependent on the available data. (Duvenhage 2020)

The data source for the chatbot is a json format file that contains different tags (topics e.g. greeting, policies, reservations, etc) and corresponding patterns, responses and contexts. These data points are populated with the help of frequently asked questions available at the Radisson hotel group website. (Radisson hotels) Within the context of a chatbot a tag refers to a certain topic while patterns are types of phrases that are expected to be entered by the user and responses are the possible replies of the chatbot. The dialogue data in the json file is static and doesn't contain all possible intentions of the user. This can lead to a failure to identify the correct intent of the user as the classifier is dependent on the provided data. Potential ways of improving availability of data would be to consider more FAQ datasets and take into account any feedback provided by the user. Another option could be to use existing free dialogue datasets however finding one specifically for the hotel guest customer service context would require further research work.

Requirements

The product allows the user to enter a phrase or question in the English language about services available at the hotel and receive a possibly helpful response. It provides a simple web based interface for the user to interact with the chatbot. A user can be defined as anyone who is interested in starting a conversation to gain information about the hotel. In addition to chatting the user also has the option to submit a feedback form with their contact details and a message to the developer about their experience. The form data is instantly sent to the developer's email for further follow-up. Since this product is based on static data and no user interaction is saved there is very little security risk.

Functional Requirements

- Chatting
    - Allow chat
    - Inform if answer not available
- Searching
    - Information about hotel policies
    - Information about reservations
    - Information about support and more
- Feedback
    - Allows user to submit feedback message

Non-functional requirements

- User interface
    - Easy to use
    - Accessible on all web browsers and different devices
- Scalability
    - Can be extended to further databases
    - Can be used by multiple users in the form of a web link

Use case model

There are primarily two agents in this product i.e. the user and the developer. The user is anyone who the user makes conversation with the chatbot to get information. While the developer is the one who can make changes to the chatbot with access to the code and the feedback email. Primarily the user can chat with the chatbot and provide feedback using the form. Developer receives the feedback data and can view, add, delete and update information including the data available for training the classifier.

Product Development

The chatbot is built in python using deep learning and the flask web application framework. The chatbot is trained on the .json data file containing intents, patterns and responses. (buffml) Machine learning is used to predict category of user's text and based on the classification a random response is selected from the available options. The chatbot's performance is dependent on the available training data and the choice of the training algorithm. Machine learning is used for several different applications e.g. speech recognition, virtual assistants, spam filters, recommendation engine, customer service, etc.

Usually there are high level and low level tools available for classification. High level tools like IBM Watson Assistant (IBM Cloud 2020) offers classifier configuration without knowledge about technologies and algorithms. Such a high level tool is usually built using low level tools such as a programming language library. Low level tools allow much more customization but also require deeper knowledge of machine learning. Examples of low level tools include programming languages e.g. Python, C/C++, Java, R, Javascript. (Vision mobile 2017)

The main reason for selecting python for this assignment is the availability of different libraries and the diversity of functionality. In addition to computation libraries like numpy and matplotlib, Python has several machine learning libraries like scikit-learn, tensorflow, keras, pandas, pytorch, etc. Python is the most popular language in the literature for NLP implementation and therefore is used for this assignment. (Raschka 2020)

The two main libraries used are NLTK and Tensorflow. Tensorflow is an open source machine learning platform. Tensorflow has extensive documentation for machine learning and is regularly updated with state of the art methods (tensorflow 2020). In simple terms the library uses tensors i.e. a generalization of vectors and matrices, as inputs and outputs for a neural network model where it is possible for the network to have multiple layers. Keras is a library built on top of tensorflow and is a popular choice for beginner chatbot implementation. It is used to create a model with multiple layers, compile, fit and save the model. (Keras 2020)

NLTK is the natural language toolkit library in python and for this project is used for tokenization and stemming of the text data. Tokenization iterates though the patterns and breaks the text into small parts. Stemming involves conversion of words to their stems (roots) followed by a pickle file to store the python objects. Stemming reduces vocabulary of the machine learning model and therefore reduces training time. The text is further converted into numerical values for processing with the neural network architecture i.e. this method is known as the bag of words where a bag represents each text pattern as a numerical array. After data preparation a neural network model using the keras library with 3 layers is implemented. The neural network takes a shuffled bag of words implementation and outputs intent for the input data. The model is trained over 200 iterations and shows a near perfect accuracy based on the available dataset. (Zhang et al 2010)

A function for getting the chatbot response based on the predicted intent is created. This function is used in the flask application to generate a response based on the user text. Three separate flask routes and html pages are created for the three web pages i.e. home, about and feedback. The home page uses the get method and the

chatbot response function to generate a response based on the user text. The about page is a simple text based page that provides details about what topics the chatbot is trained to answer. The feedback page uses both the get and post methods. The get method is used to return the html form template while the post method uses if else condition to first check whether all form fields are entered correctly. If validated it returns a success message and otherwise flashes an error message on the same page. (Flask 2021)

Finally some testing and evaluation was performed to check if this simple chatbot application is working as desired. A simple testing strategy to check if the chatbot behaves as expected was used e.g. when user types text an answer is generated, when user submits feedback form a thank you message is generated and the developer receives an email, when a user incorrectly submits feedback for an error message is generated. For evaluation purpose the type of questions were categorized into four categories i.e. relevant, irrelevant, no response and poor response questions. A relevant question is already available in the dataset, irrelevant is a question not related to hotel queries or not available, no response means the question is valid however no answer exists while a poor response means the algorithm fails to classify the correct response. User data collected for these categories over time and via feedback can be used to primarily improve the json dataset which includes intents and responses.

Software development methodology

The waterfall design and development methodology is used for this project. The model follows certain pre-defined steps during the development process. First step of the process is to collect the requirements. (Waterfall model)

This is followed by the design of the web interface and actual coding of the product.

The evaluation and testing phase is the final step if required.

Based on all these steps the product is accepted or rejected.

One can only move to the next step of the model if the previous step has been successfully completed.

Project management

Time management is critical when working on product development and there exist tools that help track project progress. For this project a gantt chart was used for tracking and recording time taken for different steps. Gantt charts can be useful

for planning, tracking and identifying tasks that can be improved within a project (Ong, Wang, Zinon 2016) Primarily the first month was spent reading introductory tutorials for beginners about python, flask, html and css. With very little experience in software engineering it was important to familiarize oneself with the different syntaxes of these low level tools and developing an understanding of how different languages work together for development of a data science product. First 2-3 weeks of the second month were used to implement and design the chatbot using online tutorials and documentation. While the last week was used for final dataset selection, testing, evaluation and addition of a feedback feature.

Conclusion

The chatbot created for this assignment is a simple task based chatbot that uses a deep learning model to predict intent based on input user text. The dataset used in json format is static i.e. it doesn't change based on new user interactions as no history is saved. Possible improvements include use of more data from multiple sources and use of algorithms that dynamically learn from each user interaction. This can allow models to improve their predictions and responses as more data is collected. Improvements to the design and layout of the webpages and embedding chatbot into an existing messaging app like whatsapp or slack can also be further improvements.

References

B. Duvenhage, "Natural language understanding for task oriented chatbots,"
Available: https://medium.com/feersumengine/natural-language-
understanding-for task-oriented-chatbots-6b494a16d7a7

Elliott, J. (2000). Design of a product-focused customer-oriented process.
Information and Software Technology, 42(14), pp.973-981.

https://www.buffml.com

https://flask.palletsprojects.com/en/2.0.x/quickstart/#a-minimal-application

https://www.radissonhotels.com/en-us/faq

IBM Cloud, "Watson Assistant," Available: https://www.ibm.com/cloud/watson-
assistant/

Keras, "Keras: The Python deep learning library," Available: https://keras.io/

Ong, H., Wang, C. and Zainon, N. (2016). Integrated Earned Value Gantt Chart
(EV-Gantt) Tool for Project Portfolio Planning and Monitoring
Optimization. Engineering Management Journal, 28(1), pp.39-53.

S. Raschka, J. Patterson, and C. Nolet, "Machine learning in Python: Main
developments and technology trends in data science, machine learning, and
artificial intelligence," Information, vol. 11, no. 4, Apr. 2020.

TensorFlow, "TensorFlow — An end-to-end open source platform for machine
learning," Available: https://www.tensorflow. org/

The waterfall model. Available at: http://www.selectbs.com/analysis-and-design/what-is-thewaterfall-model

VisionMobile, "Developer economics: State of the developer nation Q1 2017," 2017. Available: http://s3- eu-west-1.amazonaws.com/de-blog/uploads/2017/03/VisionMobile-DeveloperEconomics-State-%E2%80%93-of-the-Developer-Nation-Q1-2017.pdf

X. Luo, S. Tong, Z. Fang, and Z. Qu, "Machines versus humans: The impact of AI chatbot disclosure on customer purchases," Marketing Science, Jun. 2019.

Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: A statistical framework," International Journal of Machine Learning and Cybernetics, vol. 1, pp. 43–52, 12 2010