2020/2021

A picture containing shape

Description automatically generated

Student

K.M. Mudith Chathuranga Silva (kogsi273)

Examiner

Prof. Marco Kuhlmann

**732A92 Text Mining Final Project**

**Chatbot for a service company**

Table of Contents

[Abstract 2](#_Toc58876831)

[1.Introduction 3](#_Toc58876832)

[1.1 Motivation 3](#_Toc58876833)

[1.2 Objective 3](#_Toc58876834)

# Abstract

Since the development of social media networks, more and more businesses have been using chat services to attract more customers by providing real-time information for their potential or existing customers. Medium or large scale businesses have a dedicated customer service person to cater to their online customers’ needs. But for small-scale businesses are unable to compete with their rival companies since they have a limited amount of employees and unable to pay for an extra person for online customer handling. And also there are some special service-related businesses in which the owner needs to directly interact with customers. So they can’t use an employee to handle their customer because of the business model or they can’t expose customer information with another party.

Over the last few years, we could see that Text Mining and Natural language processing has been spreading to varies kind of industries. So we could use the Text mining knowledge to address the above situations and create an automated chatbot. We could extract existing customer's text messages from a specific industry and create a simple chatbot that could communicate with customers and solve their problems in real-time.

# 1.Introduction

## Motivation

I have a friend in Sri-Lanka who is running a small-scale saloon. And also he has unique customers such as Actresses, Businessman’s … etc. Most of his customers make/cancel and remove appointments through WhatsApp and Facebook Messenger (Private profile and Business profile). By the way, he won’t use the phone when he is providing a service to a client. Therefore sometimes he can’t reach the phone to reply to his exiting customer or a new customer who is asking general questions or even ask appointments over the chat. Some of the important clients' required a reply instantly and new customers are also demanding a reply from the saloon as soon as possible. Since I’m aware of this problem, I thought that Text Mining may help to resolve this problem somehow.

## Objective

It is always recommended to have a look at the dataset before implementing the solution to the problem. So sample chat records were requested for initial analysis. Then several question groups were identified. Automated replies could be generated from the saloon side by checking the client question type. But the tricky part is to extract customer information for the appointment placement. Then generate an appropriate text according to the customer's reply. The chatbot would be an ideal solution for this kind of scenario. A chatbot could be trained by using previous customer questions and provide a general reply. And also python libraries such as Spacy could be used to identify customer replies and store required data.

# 2.Theory

There are some frequently used terminologies in Text mining and it is required to get some idea about those terms and the theory behind them. Generally ‘Spacy’ and ‘sci-kit’ libraries has been using for the project and the below terms are corresponding with the library functionalities.

Tokenization

During processing, spaCy first tokenizes the text, i.e. segments it into words, punctuation and so on. This is done by applying rules specific to each language. For example, punctuation at the end of a sentence should be split off – whereas “U.K.” should remain one token. (Spacy - Tokenization, n.d.)

Stop Words

A “stop list” is a classic trick from the early days of information retrieval when search was largely about keyword presence and absence. It is still sometimes useful today to filter out common words from a bag-of-words model. To improve readability, STOP\_WORDS are separated by spaces and newlines, and added as a multiline string. (Spacy - Stop Words, n.d.)

Count Vectorizer

Number of words appear in the given document and represent the count in a vector form. This approach may be ignore rare words when creating the model. To overcome this problem inverse document frequency method introduced. (scikit - CountVectorizer, n.d.)

Inverse Document Frequency

Overall weighted document of a word. This helps to detect most frequent words and penalize them. TfidfVectorizer weights the number of word count in the document by calculating the frequency of the word. In scikit-learn, the tf–idf weight is computed as:-

* 𝑁 \_denotes the number of documents in the collection.

# Bibliography

*Spacy - Tokenization*. (n.d.). Retrieved from https://spacy.io/usage/linguistic-features

*Spacy - Stop Words*. (n.d.). Retrieved from https://spacy.io/usage/adding-languages#stop-words

*scikit - CountVectorizer*. (n.d.). Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.feature\_extraction.text.CountVectorizer.html