Question extraction and intent classification using COVID-19 Twitter data

Ivaylo Krumov
BSP 2 - BiCS Semester 2 (2021-2022)
ivaylo.krumov.001@student.uni.lu

Project objectives

Scientific objectives

The scientific part of this project aims to answer the question “Can a program process human-written questions to identify their intent?”. The scientific deliverable will complement the technical one by acting as its theoretical counterpart and each of them will provide an answer to the scientific question in one way or another. Alongside this answer, the scientific deliverable is expected to inform about the various ways questions in languages can be structured, as well as give insight into the NLP domain, particularly into the topic of intent classification.

Technical objectives

From a technical point of view, the purpose of this project is to create a program that is able to process a large set of unsorted data and sort it by classifying its contents in terms of shared similarities. In particular, the program is expected to sort out a set of unordered questions by autonomously recognizing their intent and map each question to an appropriate intent group. The technical deliverable will also complement the scientific one by acting as its practical counterpart and will be used to help answer the scientific question. The final product is expected to fulfill the functionality of a tool that a user is able to utilize for the purpose of extracting and grouping different questions with similar intents.

Method

Pre-requisites

The complete realization of the project, both the scientific and technical deliverables, will require the possession of several skills and competencies that need to be acquired before beginning work on the project. The following is a list that briefly describes them all (in no particular order):

- Experience in programming using the Python programming language;
- Ability to use several Python packages (libraries) for the development of a product, specifically pandas, spacy, tweepy, among others.
- Basic knowledge about the NLP field of study and its subdomains;
- Theoretical knowledge about the definition of a question and the forms it can take;
- Fundamental theoretical and technical knowledge about intent classification.

These prerequisites, in combination with the additional knowledge obtained during the work on the project, will be used to fully develop both parts of the project.

Method description

To answer the scientific question “Can a program process human-written questions to identify their intent?”, a thorough research about the domains it concerns will be done. A high level view of the field of natural language processing (NLP) will be provided, while a more detailed exploration of text and intent classification will be given thereafter. In this scientific deliverable numerous smaller questions like “What is an intent?”, “How does natural language understanding (NLU) relate to intent classification?” and “What is required to create an intent classification model?” will be answered. Each answer will contribute to the formulation of the definitive answer to the scientific question.

The technical deliverable will be a Python program that will be trained to classify questions based on the similarity between their intents. The program will use a large set of IDs of Tweets related to the COVID-19 pandemic as a test sample. The respective Tweets will go through pre-processing, where only the English-written Tweets will be ultimately selected as the

test sample set to be used. This will be done with the goal of providing a lighter but also more focused and concise work on the already complex enough program. Afterwards, each Tweet will be analyzed and any question it may contain under some shape or form will be extracted and put up for the intent classification process.The program will use a predefined data sample, trained for classification using a predetermined set of intents, as a reference to assign the proper intent to each extracted question. Questions with similar or identical intents will be grouped up together under the label of the respective intent they have in common. The outcome of the classification process is expected to consist of several intent labels, each housing a certain number of questions corresponding to that intent. The produced result will be visually displayed in the form of a table containing all the relevant data, which will also be written to an external file for documentation purposes.

The technical work will rely on several technical dependencies. As previously mentioned, the program will be written in Python, which allows for the usage of many versatile tools and libraries. One such tool is spaCy package, which will be used for the Tweet selection and question extraction processes and will also serve as the back-end of the program's intent classification process. The classification itself will be developed with the help of the Rasa NLU package. Tweepy is another vital library, which will be required to be able to work with Tweet IDs in the first place. Google Colab will be the coding environment of choice due to its easy accessibility and the utility that Colab notebooks offer for machine learning projects in particular.

Evaluation

After a proper implementation of the project's described research and development method, it is expected that all of its objectives will be fully achieved. The quality of the project's final results will be evaluated based on several criteria, e.g. the completeness of the answer to the scientific question, the correctness of the program's intent classification results, etc.