Apr 27, 2023

Plagiarism Scan Report

7% Plagiarized

93% Unique

Characters:6570	Words:982
Sentences:44	Speak Time: 8 Min

Excluded URL	None

Content Checked for Plagiarism

Question Answering Model based on Topic Identification Hari G,Kanumuri Sri Charan, Gurukiran Hanamantappa Veerapur, Kingshuk Karmakar Dr.Mamatha.H.R Abstract—In today's world, we are faced with an overwhelming amount of information. With the vast amount of data available online, it can be challenging for users to find accurate and relevant answers to their queries. This project aims to address this modern problem by developing a question-answering model based on topic classification. The model will use data from Stack Overflow, a popular online platform for developers to share their knowledge and expertise. The data will be scraped from the website and classified into different topics using known web scraping techniques. By leveraging the vast knowledge base available on Stack Overflow, this model will provide users with accurate and relevant answers to their queries, helping them navigate the overwhelming amount of information available online. Introduction In the modern world of rapid development of deep learning technology, we can see a significant rise in research and development of question answering systems mainly for the purpose of human assistance. The presence of softwares such as chatbots could be a well known example of this current situation. Quite a few of these services come along with the concept of topic classification. The reason this plays an important role is because it helps the system understand the context of the user's query and can then provide more accurate and relevant answers. Deep Learning is a well established branch of Machine Learning which acts as a core technology in this field due to its ability to learn from data.In the context of QA models, deep learning methods can be used to improve the accuracy of answers by better understanding the user's query and the context in which it is asked. Overall, the use of deep learning methods in QA models and topic identification can help improve the accuracy and relevance of answers by leveraging advanced machine learning techniques to better understand and process user queries. Related work There have been several related works in the field of question answering (QA) models and topic classification. One such work is the development of MultiModalQA (MMQA), a challenging question answering dataset that requires joint reasoning over text, tables, and images1. This dataset was created using a new framework for generating complex multi-modal questions at scale, harvesting tables from Wikipedia, and attaching images and text paragraphs using entities that appear in each table. Another related work is the evaluation of classification models for question topic categorization using a large real-world Community Question Answering (CQA) dataset from Yahoo! Answers. This study evaluated the performance of different classification methods on question topic classification as well as short texts, and found that certain aspects were important for question topic classification in terms of both effectiveness and efficiency. There are also several products and libraries available that can be used for question answering (QA) models and topic classification. One such product is BERT(Bidirectional Encoder Representations from Transformers), a pre-trained machine learning model developed by Google that can be used for a variety of NLP tasks, including QA and text classification. Another product is GPT(Generative Pre-trained Transformer), which is a language based prediction model that can be used for various NLP tasks, including text classification, sentiment analysis, text generation and question answering. Research statement Our research aims to improve the accuracy and efficiency of a question- answering model for technical and developmental domains, using StackOverflow as a source of training data. We explore various transformation based models and architectures such as - BERT(Bidirectional Encoder Representations from Transformers), T5(Text- to- Text Transfer Transformer) and DistilBERT to build a model that can accurately answer a wide range of technical questions. We also tried fine tuning data to reduce the amount of data required to train said model. The first model tested was trained on a combined number of 300 question-answer pairs. It had a higher accuracy than normal BERT model on the same dataset. Our model is based on DistilBERT which is faster and more accurate than the BERT model. Our results demonstrate the potential of leveraging user-driven knowledge sources such as StackOverflow for building effective question-answering models for technical domains. System Overview Coming to the architecture of the system, we consist of 5 main components : Datasets : Collected from the Stack Overflow website through web scraping, creating two datasets (labeled questions and question-answer pair) Input : The user would supply their question to the model Topic identification : This layer will identify the label of the inputted query Question Answering model : This section will generate the required answer based on the given input. 1)Dataset Formation The dataset for training this question-answering bot using StackOverflow data was done by getting the questions for certain tags using the stackexchange API. The answers were extracted using web scraping tools to extract the text and metadata for a large number of questions from StackOverflow. Once the data has been extracted, it is cleaned and preprocessed to remove HTML tags and tokenizing questions and answers. The resulting dataset is a pair of questions and answers in a JSON format making it easier for a BERT based model to learn. This dataset is then used to train the transformer based model for question-answering. The learning method is a combination of supervised and unsupervised techniques, these methods are used to optimize the model. Overall the dataset formation and cleaning is a complex and iterative process that involves multiple layers of preprocessing and cleaning. For more information on how to clean similar datasets, check out the github repo for this project -

https://github.com/smarton2k2/NLP-Project (private at the moment) 2)Topic Identification The purpose of this study is to implement topic identification in natural language processing using BERT-Uncased, a state-of-the-art language representation model. Training was implemented on a dataset with questions and their respective labels (NLP,API,Blender,Flutter,Unity), hence the abstraction of knowledge was implemented through the use of supervised learning.

Sources

2% Plagiarized

by A Talmor · 2021 · Cited by 59 — In this paper, we present MultiModalQA(MMQA): a challenging question answering dataset that requires joint reasoning over text, tables and ...

https://arxiv.org/abs/2104.06039

2% Plagiarized

by A Talmor · Cited by 58 — We create MMQA using a new framework for generating complex multi-modal questions at scale, harvesting tables from Wikipedia, and attaching images and text ...

https://openreview.net/forum?id=ee6W5UgQLa

2% Plagiarized

May 13, 2019 — We study the problem of question topic classification using a very large real-world Community Question Answering (CQA) dataset from Yahoo!

https://www.researchgate.net/publication/260282048_An_evaluation_of_classification_models_for_question_topic_categorization

Home Blog Testimonials About Us Privacy Policy