Bangladesh Army International University of Science and Technology (BAIUST)

Department of Computer Science and Engineering

Lab report

	Course Title: Artificial Intelligence Sessional ***Course code: CSE- 404*** ***Name: Ahamed Jamil Bhuiyan*** ***ID: 1107026*** ***Level-4, term-1***	

0bjectives: Finding innovative ideas where Natural Language Processing can be implemented

Introduction

Natural language processing is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. Natural language processing is among the hottest topic in the field of data science. Companies are putting tons of money into research in this field. Everyone is trying to understand Natural Language Processing and its applications to make a career around it. Every business out there wants to integrate it into their business somehow. Because just in a few years' time span, natural language processing has evolved into something so powerful and impactful, which no one could have imagined. To understand the power of natural language processing and its impact on our lives. There are many application of NLP. This lab report I will be represent my idea of email classification.

Email classification

An email classification system is a monitoring system from within an email inbox, designed to monitor the real-time flow of email into an organization. By extracting the email sender and content, and learning user behavior, the system sorts emails into the preferred folder

Methodology: NLP is a subfield of artificial intelligence that allows machines to manipulate human natural languages. NLP has been exploited in many fields. In this report, I will be going through the steps of classifying an email weather it is a spam or not using NLP techniques

Problem statement-

Most of us should be familiar with spam emails. Cisco defines it as unwanted junk email sent out in bulk to an indiscriminate recipient list. Typically, spam is sent for commercial purposes. It can be sent in massive volume by botnets, networks of infected computers. Therefore, spam email filtering is an essential feature for email services such as Outlook and Gmail. Services providers are extensively using Machine learning techniques to filter and classify them successfully.

The dataset are using is a public text data, it contains two columns including text (email) and spam (label). The spam column has two values (0 and 1), the text is labeled 1 if it is a spam or 0 otherwise.

In computing, a pipeline, also known as a data pipeline, is a set of data processing elements connected in series, where the output of one element is the input of the next one. The elements of a pipeline are often executed in parallel or in time-sliced fashion

The dataset are using is a public text data, it contains two columns including text (email) and spam (label). The spam column has two values (0 and 1), the text is labeled 1 if it is a spam or 0 otherwise. ETL (Extract, Transform and Load) Pipeline refers to a set of steps that allow us prepare our data for the machine learning model. Transforming the data allows to clean it

Machine Learning Pipeline

In this step, I am going to build a NLP pipeline, this will include:

1. Text Processing

Tokenize→Clean→ Normalize

Tokenize

Given a plain text, we first normalize it and convert it to lowercase and remove punctuation and finally split it up into words, these words are called tokenizers.

Clean

Remove stop words to reduce the vocabulary

Normalize

In order to further simplify our text data, we can lemmatize or stem in this step. Lemmatization and Stemming usually refer to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma

2.Evaluation

In order to evaluate the performance of our machine learning model, I am using the following metrics:

Accuracy: the fraction of predictions the model classified right

Confusion Matrix: a summary table to evaluate the accuracy of a classification, it breaks down the number of correct and incorrect predictions by each class.

The purpose of the pipeline is to assemble several steps that can be cross-validated together while setting different parameters.

3.SKLearn Pipeline

The components of sklearn pipeline are the following:

The CountVectorizer () Where we pass in our tokenize function to provide vocabulary and encode new documents using that vocabulary

The TfidfTransformer() Transforms a count matrix to a normalized tf or tf-idf representation. Tf means term-frequency while tf-idf means term-frequency times inverse document-frequency. This is a common term weighting scheme in information retrieval, that has also found good use in document classification.

Discussion

In this report I built an ETL pipeline that helps to prepare the data in order to use it for building a machine learning model that classifies emails.

I implement a basic NLP techniques and machine learning and found. The results are not bad as the accuracy of this model without hyper-parameter tuning nor adding another as we have a very small dataset and using the basic techniques of NLP and ML.

Reference

[1] ETL Pipeline

[2] Udacity Data Scientist Program

[3] sklearn.pipeline