ML2019-Assignment 3

The Answer Of The Question One

Student Name: Youwen Zhang

Student Number: 12769687

Problem Description

With the proliferation of online services and mobile technologies, the world has entered the era of multimedia big data. A lot of research work has been done in the multimedia field, which analyzes different aspects of big data, such as capturing, storing, indexing, mining and retrieving multimedia big data. However, few studies provide a complete survey of the entire framework for multimedia big data analytics, including the management and analysis of large amounts of data, current challenges and opportunities, and promising research directions. To achieve this goal, we provide a comprehensive overview of the latest research in the field of multimedia big data analytics. We aim to bridge the gap between multimedia challenges and big data solutions under the current big data framework, discuss their applications in multimedia analytics, existing strengths and limitations of existing methods, and the potential future directions of multimedia big data analytics.

In the past few years, multimedia data has been rapidly and widely used, and the ease of use and usability of images, audio, video, and text, as well as multimedia resources, have had a major impact on the data revolution of multimedia management systems. Currently, multimedia sharing sites such as Yahoo Flickr (Flickr.Com 2016), iCloud (iCloud.Com 2016) and YouTube (YouTube.Com 2016), social networks such as Facebook (Facebook.Com 2016), Instagram (Instagram.Com 2016), Twitter (Twitter.Com 2016), etc., these multimedia giants are considered to have unique and valuable resource data. For example, to date, Instagram users have uploaded more than 20 billion photos, YouTube users have uploaded more than 100 hours of video per minute per day, and there are 255 million active videos. Internet traffic shared via multimedia reached 6,130 Petabytes per month in 2016. It is predicted that the digital data rate will exceed 40ZB by 2020, which means that everyone in the world will produce nearly 5,200 gigabytes of data.

Knowledge point

We will use the data on Twitter to combine Spark MLlib to realize the sentimental analysis of the two presidents of the United States and see how netizens in different parts of the United States think about them. Here we will use Twitter data, Spark MLlib sentiment analysis, Python map visualization tool Basemap.

Experiment process

First, we need to get the twitter stream data, then perform sentiment analysis on the tweet data, and finally visualize the analysis results.

Method-Machine learning

Nowadays, natural language processing (NLP), machine learning (ML) and other fields can be described as hot anomalies. With the explosive development of social networks, many researchers have naturally become *interested in text analysis on social media. And when it comes to social media, so. Does using some machine learning algorithms to analyze Twitter will produce some interesting results? With this in mind, this exercise started.*

The first step in the project is of course to download enough Trump tweet data. Twitter provides an official API for users to interact with the client in Python language; researchers can download twitter-related data directly by providing their own key information. For details, see the related documentation [1]. However, the official API has a big limitation, that is, when downloading, it can only trace back 3,200 tweets before the same user, and then it will not be given. . . This is a bit disappointing, saying that good big data is gone. Since the use of the official API is not very convenient, I actually call the higher-level tweepy[2], and then collect the relevant data of the @readDonaldTrump account, including the published tweet text, release time, and release language. , likes, number of forwarding, publishing platform, etc.

After getting the data, we can do some basic analysis, such as building a simple bag of words (BOW) model. Here we only count the original (non-forwarding) tweets of normal type tokens, and filter out stop words (stopwords, such as to, for, a, etc. without actual semantics), simple numbers and partial (meaningless) place names Wait. A total of 57,589 words, 31,503 non-stop words, and 5,206 unique words were obtained from the statistics. The most frequently used words are great (516 times), thank (397 times), hillary (362 times), clinton (261 times), people (241 times), trump (209 times), america (198 times), etc. , very consistent with our consistent impression of trump language. Finally, the words in the word bag are drawn into a word cloud as a visual display. Then we can use machine learning to introduce KNN, SVM and other classical algorithms to analyze data features.

Method-Business analysis

In addition to machine learning, we can also use the method of business analysis. Of course, this requires us to be quite familiar with Twitter's data business. Sometimes, directly analyzing business data may be more accurate than using machine learning. For example, a tweet of praises and comments will play an important role in whether or not this tweet can be forwarded.

Ethical and social consequences

Whether using machine learning or using data services, The premise is that we have a lot of data, because only the data is getting bigger and bigger, our accuracy will be higher and higher. Then when the amount of data becomes very large, there will be a problem, that is, it will involve the user's personal privacy issues. This is also an inevitable problem in the field of data science today. Therefore, more and more ethical problems arise in today's society. This is also a problem that we urgently need to solve.