Chapter 2
Twitter Data Analysis Using Apache Streaming

Lavanya Sendhilvel
Vellore Institute of Technology, India

Kush Diwakar Desai
Vellore Institute of Technology, India

Simran Adake
Vellore Institute of Technology, India

Rachit Bisaria
Vellore Institute of Technology, India

Hemang Ghanshyambhai Vekariya
Vellore Institute of Technology, India

ABSTRACT

Real-time data from social network sites like Twitter or Facebook has been a popular source for analytics and researchers in the recent years due to various factors like large amount of data, structured-ness, and popularity. Analyzing data is a very common requirement today, but such requirements become difficult when there is a bulk of data which needs to processed and analyzed in real time. Analyzing large number of tweets from Twitter to get different patterns and extract useful information is a massive challenge. Apache Spark is a platform that can be used to handle big data efficiently, and it offers faster solutions compared to Hadoop. This chapter addresses the issue of real-time analyzing and filtering the tweets as per the user's requirements from among the millions of other streaming tweets and classifies them into various categories. It creates an interactive automatic system that splits data based on important keywords and displays a graphical representation of connected tweets using Apache Spark.

DOI: 10.4018/978-1-6684-5264-6.ch002

INTRODUCTION

Twitter is a popular social media site where people communicate abouts news, topics of interest, grievances using short messages commonly referred as tweets. Twitter users can express or share their opinions, information regarding events, products in anything in their tweets. Hashtag is the convention of prefixing a word in a tweet with the symbol '#' which indicates a keyword or topic of the tweet. It is used for categorization of tweets based on topics and helps in searching. Keeping up with users and their tweets, trending hashtags help us understand what is going around in the world and people's sentiment on it. Tweets often contains latest information, and it is frequently updated. Tweet analysis can reveal useful information which can create a practical and immediate application in the life of common man.

Due to the benefits of networking sites like twitter, users find it easy to share information or opinions regarding any event, products etc. instead of publishing them in print or online media which saves cost, time and efforts. This paper investigates the problem of real time analysis and filtering those specific tweets which a user wants without having any twitter account. Because social media material is unstructured in comparison to other sources, big data technology like Spark can manage the processing and analysis of unstructured data. The tweets will be streamed and processed in real time using Apache Streaming and TCP client socket programming. Aggregated tweets under categories such as sports, news, traffic jams, complaints etc are stored locally making it easy for users to keep a track of topic/s they are interested in.

The goal of this work is to make a Twitter Data Analysis programme available to the public as a service. We have utilised Apache Spark to use a developer API to extract live tweets from Twitter, classify the tweets, and show them on the user interface. IntelliJ, an integrated development environment has been used to run this programme. Two services have been included, one for classifying real-time tweets and the other for visualising the data from archived tweets.

LITERATURE SURVEY

Twitter trend analysis is done by 2 methods-first using normal execution environment in which latent dirich let allocation, cosine similarity, k-mean clustering and Jaccard similarity techniques and second using big data Apache spark tool implementation. ("Apache Spark", 2016). They both were compared and conclusion was made spark are better and faster than normal execution environment ("Apache Hadoop", n.d.). The twitter streamed data in Apache spark the data in clustered to achieve less computations time sparks works in 2 phase first by creating viable clustered utilization by using fuzzy c-mean clustering and it is further improved by adaptive particle swarm optimization