Developing a reproducible workflow for batch geoprocessing social media in a cloud environment

RICARDO MORALES TROSINO
March 2019

SUPERVISORS:
Dr. F.O. Ostermann
Dr. O. Kounadi

Developing a reproducible workflow for batch geoprocessing social media in a cloud environment

RICARDO MORALES TROSINO
Enschede, The Netherlands, March, 2019

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.

Specialization: Geoinformatics

SUPERVISORS

Dr. F.O.Ostermann

Dr. O. Kounadi

THESIS ASSESSMENT BOARD:

Dr. MJ. Kraak (Chair)

Dr. E. Tjong Kim Sang (External Examiner, Netherlands eScience Center, Amsterdam)

DISCLAIMER

This document describes work undertaken as part of a programme of study at the Faculty of Geo-Information Science and Earth Observation of the University of Twente. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the Faculty.

ABSTRACT

The main objective of this research is to deliver workflow scenarios that can process and geoprocess social media with batch data. The research focused on defining useful tasks and sub-tasks to explore and analyze batch social media and to deliver a prototype able to reproduce the workflow. Two architectural scenarios were identified. One scenario designed for newcomers in a local machine and another for more advanced users in a cloud environment. A local machine scenario developed to explore a stored data set with a sample of the data set, and a more complex scenario to explore the complete data set in the cloud and with a big data framework such as Spark. A prototype was designed to test the workflow and to achieve reproducibility. To test the prototype, a data set was provided with the intention to search for tick bites events in the Netherlands. The results showed that, following the workflow, the example data set contains some noisy words and the processing in the cloud environment was relatively cheap and efficient.

ACKNOWLEDGEMENTS

I want to thank ITC teachers and staff, to share all their knowledge with me. To my fellow students who worked with me side by side.

My supervisors Frank and Rania, without your guidance this work would not be possible, thank you for your advices, but mostly to share with me your time and knowledge. To Luis Calixto and Raùl Zurita that shared with me some of their experience with Hadoop. Rosa and Mowa that help me with to improve my proposal.

To Massyel to be with me in my brightest days and darkest knights, thank you gatita. To my mom, sister, and father that gave me remote spiritual support.

I want to acknowledge my sponsor, CONACyT and Alianza FIIDEM, for giving me the opportunity to join their international scholarship program, and for providing all the necessary resources; to get to the Netherlands and the chance to grow professionally.

And to my friends of GFM, sorry guys to interrupt every class with my silly questions and comments. I just wanted to break my own bubble to understand and learn more about our World.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

API	Application Programming Interface
AWS	Amazon Web Services
CLI.	AWS Command Line
DAG	Directed Acyclic Graph
DBMS	Database Management System
DNS	Domain Name System
EC2	Amazon Elastic Compute Cloud
EMR	Elastic Map Reduce
GPU	Graphing Processing Unit
HDD	Hard Drive Disk
IaaS	Infrastructure as a Service
IAM.	Identity Access Manager
NLP	Natural Language Processing
PaaS	Platform as a Service
RAM	Random Access Memory
RDD	Resilient Distributed data set
S3	Amazon Simple Storage Service
SaaS	Software as a Service
SSH	Secure Shell
VM	Virtual Machine
VPN	Virtual Private Network

1. INTRODUCTION

1.1. Motivation and Problem Statement

Social media (SM) has become a channel to share information between users, in recent years experiencing exponential growth. Researchers have been interested in the study of this type of information due to its complexity, immediate generation of knowledge, its spatial and time characteristics, and the volume that social media can generate. For a spatial scientist, it is relevant to study how to extract information from this type of source, what type of spatial procedures can be applied and in which platforms and what spatial valuable information can be extracted from this type of data source. It is important to generate usable workflows to make SM information reproducible and more accessible to people with minimum experience in spatial analysis or computer science analysis. Over the last decade, the information generated by social media has evolved; with changes triggered by several factors like mobile phones accessibility, global positioning services or WIFI/4g connectivity. Nowadays, users share immediate data which may be personal, informative or even a geolocated event. Some of these can provide valuable information or services that can be used as a product. Social Networks like Foursquare, Flickr or Twitter can be used as a powerful tools to analyze trends of behavior, disasters, events, health outbreaks with a geographical location context (Mazhar Rathore, Ahmad, Paul, Hong, & Seo, 2017).

Social media and cloud computing conceptually seem to be related, but they are quite distinct. The cloud is described as a model that has several computing resources like storage, applications, networks or services. By definition the cloud has five essential characteristics, three service models and four deployment models (Mell & Grance, 2011). The five characteristics of the cloud are On-demand self-service, Broad Network Access, Resource Pooling, Rapid elasticity, Measured Service; four deployment models: private cloud, community cloud, public cloud, hybrid cloud; and three service models Cloud Software as a Service (SaaS), Cloud Platform as a Service (PaaS), Cloud Infrastructure as a Service (IaaS) (Rountree & Castrillo, 2014). Social media is defined by the several platforms that exchange information, the platforms have several channels of interaction like blogging, networking or multimedia content, they share the same goal, to provide services to exchange information between users. Emerging computing paradigms like cloud computing show high penetration rates in social and geosocial developments, due to the structure, costs, and flexibility (Ramos, Mary, Kery, Rosenthal, & Dey, 2017). Big Data as a Service (BDaaS) is a term conceived by cloud providers to give access to common big data-related services such as processing or analytics. This type of services are based on the user necessitates such as infrastructure, platform or software, but with the difference that the framework, tools, and environment are designed to process massive amounts of information (Neves, Schmerl, Camara, & Bernardino, 2016).

Organizations and researchers are looking for new models that unify real time and easily access information, providers like Amazon, Azure or Google can offer tools to implement different types of cloud models, depending on the necessities of the users, these can become elastic and scalable with promises of increasing productivity and reduce costs. The high amounts of information that the cloud can storage and process are one of the most significant benefits of the cloud (Chris Holmes, 2018)but the cloud also have some drawbacks just as the security, format inflexibility or the unexpected downtime that the cloud might have (Larkin Andrew, 2018).

In recent years the words Volume, Variety, Velocity, Veracity, and Value are described as the five V's, these words are associated with the concept of Big Data, which is related to the generation, processing, and storage of vast amounts of information (Neves et al., 2016). At some point the generated data can surpass the capacities of a local machine or even a computing cluster, social media streams usually create this amount of data. This enormous amount of information has popped up some research questions in recent years like, How to process, store and extract valuable information from it, how to make it more efficient, how to make this information more accessible. The social media data produced by the mixed social networks has been termed Social Media Big Data (Stieglitz, Mirbabaie, Ross, & Neuberger, 2018).

Social media data can be georeferenced by the service provider or even by the user, this leads to a new generation of data sets with spatial information generated by users or providers, this is called geosocial media. When the information contains a spatial reference, a spatial researcher can analyze this type of information with a geoprocessing procedure, which can be defined as a framework of tasks and tools that process and automate information from a Geographical Information System (ESRI, n.d.). Previous studies utilized the geoprocessing tasks in their research (Ostermann, García-chapeton, Kraak, & Zurita-milla, 2018) & (Yue, Zhang, Zhang, Zhai, & Jiang, 2015) they suggest an advisable geospatial analysis tasks for points, some of the tasks are: pattern analysis and clustering, methods of spatial association, pattern recognition or classification just to mention a few, this information may be a start point to explore the geoprocessing possibilities.

Geosocial media streams are relatively new in studies of big data on the cloud, and combined with geoprocessing have open unexplored paradigms of research. Due to their characteristics, these type of data have sparked some research questions like what are the possible tools to process social media in spatial context or how to make this information more accessible and reproducible to a wider audience of researchers. (Ostermann & Granell, 2017) reported the replicability and reproducibility in 58 papers through 2008-2009 in which he mention that 58 percent of the papers declare a null or limited reproducibility and replicability. The challenge then consist in define the appropriate tasks and procedures to ensure replicability.

This analysis arises some questions like, which are the top infrastructures to start processing social media data, what are the methods for geoprocessing available on the cloud, how to process large-scale data sets of stored data, how to make this research usable for other researchers. To answer these questions, it will be

necessary to study cloud environments, workflow designing, big data spatial analysis, and analyze the data set on a local machine and in the cloud. Therefore, research of a workflow design and how this could be implemented on different scenarios will allow starting a scouting of important or relevant data that in further research stages can provide automatized platforms already proved that can facilitate the information analysis.

1.2. Research Identification

The main objective of this research is to analyze and design workflow scenarios with capabilities of processing and geoprocessing geosocial media stored data. This research identifies the lack of information about technicalities and limitations when geoprocessing social media is about, what are the options to process this type of information, and how to geoprocess that information on a local machine or in the cloud. In computing analysis, a workflow is used as a way to interpret, communicate and inform a path to achieve a computational analysis (Hettne et al., 2012). A reproducible workflow could provide a trustworthy diagram to explore geosocial stored streams, this workflow will include tasks like, data storing, data management and data analysis, with techniques of filtering, cleaning or tagging. The importance of developing a workflow is to ensure their usability by other users or researchers. This is an opportunity to explore the possibilities of storage and processing into the cloud and also design and implement a workflow with geoprocessing capabilities.

It is important to explore and define the databases and tools to use in each scenario, one of the challenges is to define the parameters of the scenarios, the differences in the filtering and clustering, and the storage of the data set. It is important to make a review of the available information. In literature there exist some researches that explored topics like to model geosocial media with big data, or to identify risks based on crowd sourcing information, or extracting valuable information from social media (Mazhar Rathore, Ahmad, Paul, Hong, & Seo, 2017b; F. O. Ostermann et al., 2015; Sun & Rampalli, 2014), just a few papers incorporate the geoprocessing into their research and they have their own challenges such as insufficient geotagged information to accomplish a scientific analysis (Yue et al., 2015).

Technical challenges like the usage of non relational database, the creation of an efficient batch processing, spatial indexing with big data, generating overviews with sampling reduction strategies, and the flexibility of computing storage resources on the cloud would be encountered during this research. Some of this challenges are described by Yang (2016), that reported some geospatial studies that developed some significant geospatial challenges related to the Big Data V's and a cloud computing processing or storage. This study has the potential to contribute and provide a beginners guide to explore unexplored data sets in different scenarios such as the cloud environment.

1.3. Research Objectives and Questions

On this section, the objectives and research questions were established. Briefly, the first objective was related to the workflow, the second associated with the infrastructure, a third objective linked with a prototype system and the last one connected with the reproducibility of the research. Therefore some questions were established, in total there are nine questions, and each question try to answer specific challenges of the study.

- To specify the principal tasks and techniques to transform regular social media into geosocial information and incorporate them in a workflow proposal.
- To analyze the relationship between stored geosocial media and infrastructure characteristics to define architecture scenarios.
- To implement a prototype system that analyzes the study case information.
- To evaluate the reproducibility of the workflow and performance of the prototype.

Research Questions

1.
    1) Which tasks and techniques are necessary to incorporate geosocial media, geoprocessing and a cloud environment in the same workflow?
    2) How to operationalize the workflow integrating the required tasks and techniques?
2.
    1) Which scenarios can be defined based on the stored data, geoprocessing tasks, and system infrastructure?
    2) Which and why different type of technologies are required for each scenario?
    3) What are the advantages and disadvantages between the selected scenarios?
3.
    1) Which type of limitations from the study case data set and the proposed scenarios will affect the prototype
4.
    1) How do the characteristics of input data and scenarios affect the reproducibility of the workflow and the re-usability of the prototype?
    2) Which techniques or benchmarks are the most feasible to evaluate the performance of the prototype?
    3) How the results from the social media spatial analysis can be used for further research?

1.4. Thesis Outline

The project is divided into three main stages, the first one describes the design of the workflow based on the required tasks and techniques, the second phase will be to create the prototype and analyze the available data set, and the last step will be dedicated to evaluate the prototype and the workflow.

The document is divided into seven chapters. The first chapter describes the motivation of the research combined with a brief description of the research and the research questions and objectives. The second chapter depicts the basic concepts and research associated with social media, the cloud environment, and big data. Chapter three report the design of the workflow with detailed information on the suggested tasks and sub-tasks. Chapter four reports the implementation of the tasks and sub-tasks in the prototype. In chapter five are described the results of the prototype appliance with a study case data set. Chapter six discuss the outcomes of the research critically. Finally, chapter sever provides a brief conclusion and answer each research question.

Figure 1: Brief description of chapter contents

2. RELATED WORK

Social media has become an interesting topic for social science researchers who are seeking to find valuable information provided by the users. To comprehend the importance of this research, there exist some studies where social media data was used and helped to understand stages of an outbreak: in Nigeria 2014, social media information reflected an Ebola Outbreak before the official announcement (Edd & Rn, 2015). In 2016 a Zika outbreak in Latin America was tacked with Google searches and Twitter information. As it can be appreciated this approach may be useful to track a virus and forecast new spreading areas from social media information and could be incorporated to enhance the usual epidemiological attention of an outbreak (McGough, Brownstein, Hawkins, & Santillana, 2017).These type of researches have a direct impact on communities, but to achieve this, it is necessary to study the characteristics of the data produced by social media. It is essential to explore how to transform the data efficiently, and in which cases is essential to analyze, processing and storage data when big data and social media is in the picture.

Batrinca & Treleaven, 2014 published a review for social science researchers interested in the necessary techniques, platforms, and tools to analyze social media. In the report they describe the initial procedure to explore social media (SM), such as, the file formats expected like HTML or CSV, the main social media providers divided in free and commercial sources,tools for processing and analyze text from a language perspective or even storing data in a file or in a Database Management System (DBMS), just to mention a few examples. In 2013 Croitoru developed a system prototype to explore information from geosocial media in which he reports a conceptual model based on two systems, the first one to ingest feeds into a system and a second multi-step approach to analyze the social media information. For example, an automatic event detection using big data and a machine learning approach developed by Suma (2018) in which concludes that there are some improvements in the management and processing of data, but there are still challenges in the event detection. Both projects are useful, and their methodologies could be applied in this research as an example of a prototype with social media characteristics for processing massive amounts of data.

Twitter is one of the most used social media microblogs sources to obtain information provided by the users. Some of the analyses incorporate trending words to predict specific behavior of financial markets, or traffic congestion in conglomerated areas or even analyze data to detect events without any prior information (Lansley & Longley, 2016). Some of these studies are based on tasks to process text, to manage and process text generated from the users may be a risky task due to the number of users and the complexity of each text. Luckily some research has been done on this field. The Association and Journals such as the American Association for Artificial Intelligence, Stanford Natural Language Processing Group, Machine Learning or Journal of Artificial Intelligence Research have been working on speech and

language processing for decades in computing science; this is called Natural Language Processing (NLP) (Jurafsky & Martin, 2007). Some of the techniques and tasks used to process text from social media are related to clean text for a more specific analysis, the tools and platforms vary depending on the purpose of the research or project. Lately, some studies focus around the generation of massive amounts of text from Twitter, studying issues such as data management and query frameworks, or challenges like the necessity of an integrated solution that combine an analysis on Twitter with a big data management perspective (Goonetilleke, Sellis, Zhang, & Sathe, 2014).

Nowadays, there exist several tasks in NLP such as tokenization, language detection or stemming to analyze social media microblogs like Twitter, these powerful tools allow to monitor user activities that have been impossible to observe until now, some of these tools are described by Preotiuc-Pietro (2012) which provide an open source framework to process text, describing some tasks implemented on Twitter. Other researches have focused on tracking fake accounts (bot) on Twitter by trying to clean the information and locate patterns on the fake accounts (Wetstone, Edu, & Nayyar, 2017). These solutions are not bullet proof, but until now some of them has proved to have good performance analyzing text from the web, research continues by addressing the challenges generated in this field(Sun, Luo, & Chen, 2017).

A percent of social media records are geolocated, when a record includes this type of information from the source is called geotagged data, but when this spatial information needs to be translated from a the text source to coordinates is called geocoding. Some studies has used geolocation and geocoding to study mobility and dynamics between border countries (Blanford, Huang, Savelyev, & MacEachren, 2015) use geolocated tweets as proxy of global to determine human mobility (Hawelka et al., 2014), check land uses in cities by their tweet activity (Frias-Martinez, Soto, Hohwald, & Frias-Martinez, 2012).

2.1. Cloud Environment

In computing, it is vital to comprehend some concepts such as node, core, processor, and cluster. A node refers to one individual machine in a collection of machines that are connected and form a cluster (Roca & Cited, 2001). Each node typically has one Central Processing Unit (CPU), which in turn has one or more cores. A core collects instructions to perform tasks. The performance of a computing cluster depends on its components, but the most efficient choice of number of nodes, number of cores per node, and size of node memory is not always straightforward. A common, practical solution to select the components in a cluster is to monitor the time spent to achieve one task with a representative sample of the data and check the usage of nodes in the cluster. Another solution is to compare the size of the data set and calculate the capacity of the considered nodes and then compare the values (Amazon, 2018). The cloud has proved to provide the tools to process massive amounts of information, using the cloud services available it is possible to manage, access, process and analyses social (C. Yang, Yu, Hu, Jiang, & Li, 2017).

In cloud computing there exist four types of environments public, private, hybrid and community. The public cloud environment is defined as a variety of services offered by organizations which specialize in providing infrastructure and technologies for different purposes, this type of environment is focused on the external consumers (INAP, 2016). The private environment is often deployed within companies, similar to an intranet this type of environment is usually used internally to manage internal affairs, this can be classified as the most secure form of cloud computing. The hybrid approach is a mix between private, public or even a community environment, where each member remained as a unique entity but bounded with different standardized technologies. Jimenez (2018) combined the services of public cloud such as IaaS and SaaS with a secure private network environment to provide multimedia services for mobiles. The community environment is a combination between private and public with the aspiration of combine resources to provide grid computing and sustainability green computing.

Public cloud service providers offer several products and services such as databases, gaming, media service, analytics or computing. A possible appliance is the virtualization of data centers using the infrastructure as a service (Moreno-Vozmediano, Montero, & Llorente, 2012). A clear example of SaaS is pictured in some learning platforms that offer an interface for teaching and learning (Gurunath & Kumar, 2015). .Garg (2013) developed a index to evaluate different services that are available in public clouds, his evaluation are based on the following attributes: Accountability, agility, cost, performance, assurance, security and privacy, usability; these attributes are may be considered to select a service cloud provider.

One of the services that the cloud provide is related with infrastructure as a service, which include the provision of architecture for processing, the selection of architectural resources is based on the type of necessity of each user. The architecture selection on the cloud is associated with the time of processing, the necessity of scalability, management of groups or profiles, and the availability of particular frameworks or services such as Spark (Kirschnick, Alcaraz Calero, Wilcock, & Edwards, 2010). In recent years Hadoop and Spark have been used to process and analyze massive amounts of data. Hadoop has been an option to process this type of information in the cloud. International Business Machines (IBM) define Apache Hadoop as an "Open source software framework that can be installed on a cluster of commodity machines so the machines can communicate and work together to store and process large amounts of data in a highly distributed manner." It is common to use Hadoop, and also an object file system for storage and scalability of the clusters or nodes on the cloud. Apache Hadoop relies on a Map/Reduce model that separate the tasks by mapping and reducing. Map/Reduce developed by Google in 2004 resulted a reliable model to reduce significantly the execution time in a cluster. A combination of a cloud services and models such as Map/Reduce can reduce utilization of computing resources utilization and handle workload spikes automatically by increasing the resources when it is required (Z. Li, Yang, Liu, Hu, & Jin, 2016).

2.2. Big data Framework

Three characteristics define big data, the massive amount of information, unstructured data and the requirement to process this information in real time, one common generator of this type of information are social media users (Maynard, Bontcheva, & Rout, 2012). There exist different approaches to handle and examine a big data, some important aspects to consider are the data source, the data management, the data analysis tools available, type of analyses and very important the framework to process big data such as Hadoop or Spark (Rao, Mitra, Bhatt, & Goswami, 2018). The big data infrastructure is defined as the mechanisms to collect the information, the computer program and physical storage to collect it, the framework and environment that allows to process and canalize the information, and finally the and the foundation where the result will be backup and stored(Tozzi, 2018).

In 2010 Spark was developed by the AMPlab in the University of Berkeley, is a flexible in-memory framework that allows to process batch and real-time processing. This framework is compatible with MapReduce model and also can be seen as a complement of Hadoop framework. The Spark utilize a Master and Workers schema for the nodes of the clusters allowing parallel computing. The innovation of Spark is the inclusion of a distributed collection of objects in a set of computers named Resilient Distributed data sets (RDDs) and Directed Acyclic Graph (DAG). The main difference between Spark and Hadoop relies on the method of processing information. Spark process the information in-memory (RAM) with the use of RDD and Hadoop read and write the information in an HDD called Hadoop Distributed File System (HDFS). An RDD is a collection of read-only objects partitioned through different machines in a cluster array(Zaharia & Chowdhury, 2010). HDFS reorganize the files in small chunks and distribute the files in different nodes (Verma, Hussain, & Jain, 2016). Both frameworks Spark and Hadoop use as the resource management and job scheduling a technology called YARN, this technology distribute the work between the different cluster nodes.

When spatial data and big data is in the picture some recommendations are: to implement data reduction strategies or to provide, a computational method that minimize the computational necessities in the cluster (Armstrong, Wang, & Zhang, 2018). Some of the technologies available for huge spatial data sets are, Spatial Hadoop which enriches the Map/Reduce framework by adding two levels of spatial indexes and it contains spatial operations such as spatial join or kNN range queries (Eldawy & Mokbel, 2015). The other is GeoSpark which also provide Map/Reduce model and support for geometrical and spatial objects with data partitioning and indexing, also supports spatial querying . In practice some researchers have evaluated the performance between platforms like GeoSpark or Spatial Hadoop testing the scalability and performance (Yu, Jinxuan, & Mohamed, 2015), some interesting findings are related with the incorporation of spatial indexes instead of traditional indexing (Eldawy & Mokbel, 2015). Some of them explore the index efficiency for spatiotemporal frameworks dividing the input file into nodes, mimicking a spatiotemporal index (Alarabi, Mokbel, & Musleh, 2018). Some of these studies revealed that

Spark has better performance, scalability, and stability compared with Hadoop (Reynold, 2014). But this type of processing and analysis is still in a development stage, some challenges remains in, 1) the spatial indexing of and models to process real-time information, 2) quick assessments to calculate the propagation of errors and, 3) study of efficient methods to visualize big data sets into an understandable and communicable displays (S. Li et al., 2015).

3. CONCEPTUAL DESIGN

The workflow design is described in two steps, a general and the implemented workflow. The first workflow in this document describes the social media processing and analysis in a conceptual manner, embracing general concepts that apply to a broad social data sets and user affairs which is called “General Worklfow”. The second step focuses on to go look, and report the importance of techniques and technologies and how to implement a workflow to the current study case, called “Implemented Workflow”. The first workflow tries to provide simplicity, predictability and reproducibility in the design, providing suggestions of implementation in the cloud environment depending on the user necessities, in comparison to the second workflow which the structure is more rigid and more specific in the sub-tasks and technologies suggestions.

A workflow give infrastructure to initialize a process in an ordered way, with the promise of increase the productivity, reduce errors, diagnose and cut down errors of the process among other benefits (Integrify, 2017). The workflow design should focus on achieve clarity, simplicity, recordability, reportability and reusability, with this characteristics a problem can be analyzed in a systematic manner and also integrate intensive computing analyses (McPhillips, Bowers, Zinn, & Ludäscher, 2009). The usage of workflows is a common practice to pictorially express an abstraction for an automated process in sequential order. The generation of a workflow provides a tool for further analysis and with a good design opens the possibility to reproduce the work for new research. In social media, cloud computing and big data some researchers have developed workflows that express some of the problems shown above. Some of them developed workflows to understand some geosocial challenges, like the developed by (Zhang, Bu, & Yue, 2017) in which he developed a workflow and a tool to provide an open environment for geoprocessing raster and vectors in a friendly context. The WIFIRE tool which implements a workflow model to analyze fires based on data-driven modeling and geospatial information Altintas (2015) and Wachowicz (2016) who develop a workflow with geotagged tweets focused on the data ingestion, data management and, data querying. (Suma et al., 2018) research focused on the event detection in cities using Twitter for spatiotemporal events, in his study he developed a workflow for this type of event detection.

To develop a workflow is necessary to incorporate and define tasks and sub-tasks, in this case storage, process and analyze the geosocial media data will be carried out. As part of the workflow is essential to evaluate some characteristics of storage and processing environments, in general, some features like the supported services of the cloud, networking capabilities and performance of workflow solution needed to be reviewed and evaluated in the context of geosocial media. One crucial reflection based on the work of Hettne (2012), is based on the idea to avoid workflow decay, described as the missing factors to make a workflow executed or reproducible, one recommendation to avoid the decay is to compare functional

workflows and reuse them, extracting the main idea and restate it in your current work. The main tasks proposed on the general and implemented workflows are based on the research of Wachowicz (2016), Suma (2018) & Mazhar (2017), in which the principal tasks are, collection or data pool, preprocessing, processing or data management, data querying or data analysis. Previous work visualizes four different main tasks, for this research the main tasks will be data management, pre-processing, processing, and analysis.

The reason of designing two workflows is to propose separate perspectives to describe the geosocial media geoprocessing in a cloud environment. A general and more broad applicable approach was developed in the general workflow, which is a simpler, flexible and more robust proposal. The implemented workflow describes in a detailed scheme the technical and technological tools to fulfill the analysis.

The design of both workflows consist in two scenarios with a sample and complete data set and several branches on a local machine or the cloud environment. One scenario on both workflows is based on a sampled data set and an optional local machine architecture; the second is designed to implement an analysis with the complete data set and an optional cloud environment. The branches are designed for specific users, a novice user which objective is to explore a data set with limited sources and knowledge, and another scenario for a user who has more computational resources and is familiar with cloud computing, big data frameworks and social media conceptual analysis. The workflow branches have different degrees of difficulty when implemented, the degrees are related to the architecture selected, the data set characteristics and the research question.

3.1. General Workflow

This section describes a general overview of the workflow, in a conceptual form this workflow is defined in two ways, one perspective thinking in the availability of resources such as architectural resources or skills, and the second one with the type of data evaluated in the volume, variety, and velocity. This project considers two types of workflows, a general workflow which is designed for a general case with a broad spectrum of appliances, and a second appliance workflow where the workflow turns into a more exhaustive and detailed flow; the workflow appliance tasks are described in the following sections.

The workflow (Figure 2) contemplates a stored social media data set; the main process looks up to answer a research question within the data set. Briefly, the questions are designed to split the workflow into three sections, one section for a local machine processing and analysis, a second section that contemplates the use of the cloud, and a third one that uses the cloud but also integrates a big data environment. The first question is related to the current resources of each user, a second question focuses on the skills and the financial means, at this point the workflow is divided in two branches, a cloud branching which seeks for specific characteristics of the data set for example, volume, variety and velocity necessities and a sampled branch which looks for answering the research question with a portion of the data set; the third question

in both scenarios seem to find an answer for the research question. The research question can be defined as a question that needs to be answered concerning the data set, some questions may have a specific topic, or some questions might be more general. Also, the question should be related to the fields or characteristics of the data set.

First Question: The first question provides a branching in which the user has the possibility of process and analyze all the information in a local machine, and continue the analysis in his own machine, this scenario is designed for relatively small data set.

Second Question: Depending on the characteristics of each data set the second question provides two branches one with a sampled data set with limited resources, and another with the complete data set with extra resources, depending on budget and experience. The sampled branch uses the same methodology as the local machine branch.

In the case of selecting a cloud environment, the branching asks three new questions related with some characteristics of the data set, since each data set and user has very specific necessities the workflow only suggest some scenarios on the cloud, based on these scenarios the next branching will be applied.

In the cloud division, three factors affect all the questions, time, resources and user experience, depending on each data set and user necessities, these factors influence the decision of branching, the following questions in the workflow are also affected by these three factors. They are related with the characteristics as the volume (a) of the data set which in this case the contemplated volumes are gigabytes, terabytes, and petabytes, following by the structure of the data(b), and finally a time-related question (c). The resulted answers characteristics split the workflow between a cloud processing and analysis or the big data environment. It is important to mention that the characteristics are only suggestions due to the specificity of each user data set, these suggested characteristics may change. The workflow does not contemplate a semi-structured data set in the suggested characteristics, due to the complexity of combinations that the addition of this option represent, but it is a depicted in the question and suggested options.

Figure 2 shows the complete workflow; the third question split the image in two possible results, the upper part designed for a local machine analyses, and the lower part of the model designed for a cloud approach. The cloud approach branching was designed to give a solution to the processing and analysis with additional computational resources and a second approach in the case that data is massive and unstructured.

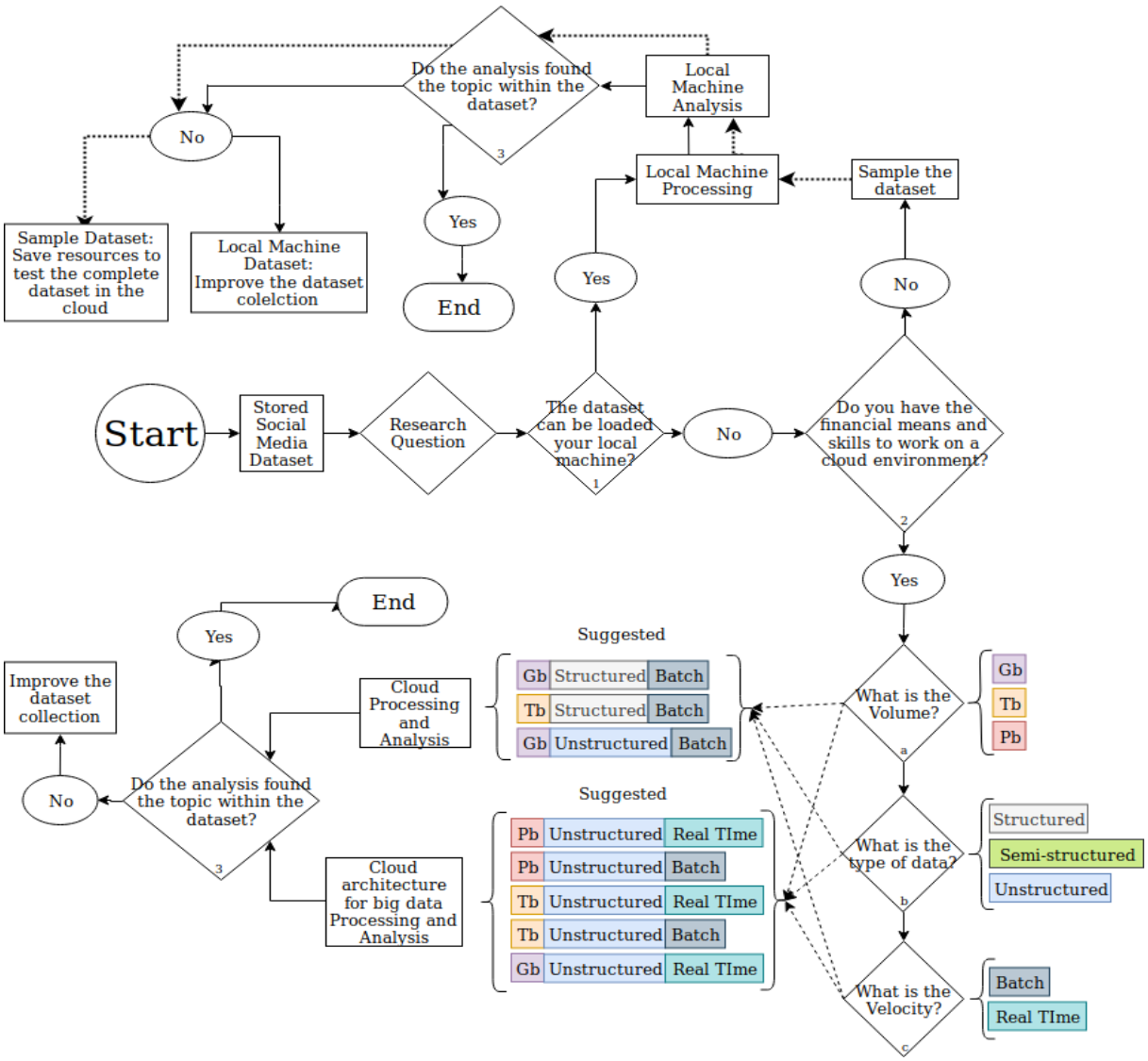

Figure 2: General Workflow

The general questions are numbered from 1 to 3 and the cloud questions have letters in the bottom, from a to c.

The Third question: this question provides an optional resolution related with the data set topic, in the case of a negative answer the suggestion is to start to collect data with the new insights from the previous analysis. In the case that the analysis is not successful in the sample branch, then the recommendation is to save resources to test the complete data set in the cloud.

3.2. Implemented Workflow

The main tasks of the implemented workflow are displayed in Figure 3 which are data management, pre-processing, processing and analysis, each task contemplate one or several sub-tasks. The tasks and sub-tasks are different from the general to the implemented workflows. Both workflows have the same objective but different purposes, the general workflow purpose is to provide a route to pick the appropriate architecture for each user, while the implemented workflow purpose is to provide details for

processing and analysis of the data. The architectural limitations and necessities established the idea of two scenarios, one scenario uses a sample, only the geotagged tweets on a local machine, and the other scenario includes the complete data set within a cloud environment.

The collected data set contains tweets with selected keywords that may represent a topic or an event, one of the objectives of this work is to verify if these words are related to the initial research question. The data set proposed for this workflow is classified as a Retrospective Event Detection (RED) because it was stored and then analyzed in a batch processing mode. A common approach to explore a data set topic is by using a frequency of words analysis, or adopting an unsupervised methodology by exploiting NLP methodologies to analyze topics or applying a supervised method to classify topics. On this workflow, the frequency of words and an unsupervised methodology are suggestions to be incorporated into the workflow workload.

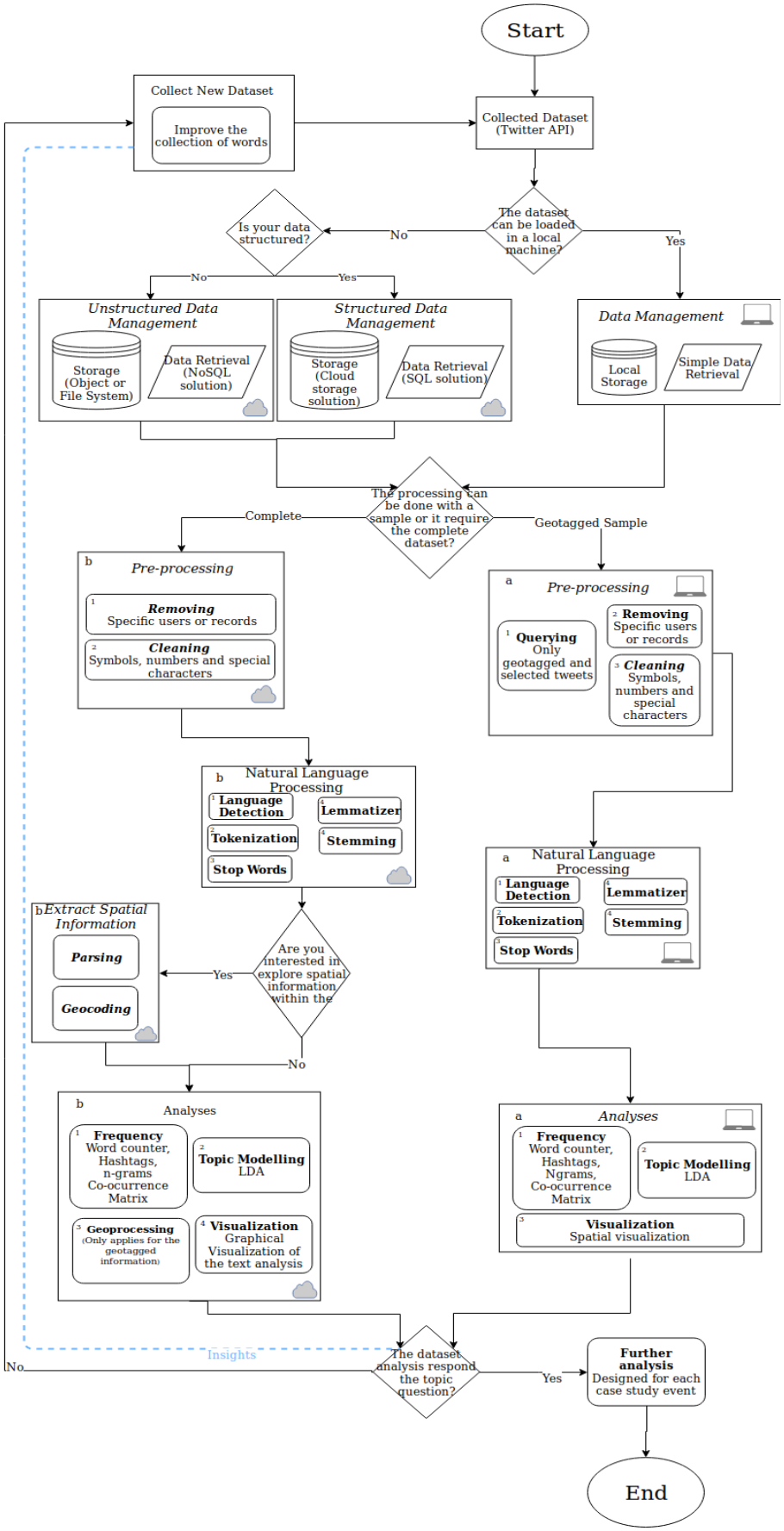

Figure 3: Implemented Workflow

The following workflow sections describe the scenarios, geotagged sample and complete data set, with a local and cloud environment characteristics. Each part describe the scenarios in detail, adding sub-tasks and technologies suggestions to the workflow. First with a introduction to the scenarios, followed by a data management description, and finally the detailed description of each scenario.

3.3. Scenarios Introduction

Two scenarios are described in the following paragraphs, one scenario in a local machine, with a sample of the data set designed to explore the data set, and a second scenario in the cloud with the complete data set and designed to process massive amount of social media stored information.

The workflow represented in Figure 3 is divided in two branches. The first branch of Figure 3 assumes a local machine scenario (depicted with a laptop logo in each workflow form), within a sampled geotagged data set (depicted with letter “a” in each workflow form), while the second branch represents the cloud environment (depicted with a cloud logo in each workflow form), within the complete data set (depicted with letter “b” in each workflow form). Some sub-task are dependent on the result of the previous task while others could be reordered or even skipped. But, on the workflow the sub-tasks have been numerated as a suggestion to follow.

The local machine scenario design contemplates querying and filtering data, the use of NLP techniques, and appliance of basic analysis techniques. The data set was sampled by their spatial characteristics or the geotagged records, typically only 1 – 3% of tweets have been geotagged, this is the information used in the sample scenario. The complete scenario has a similar design but with some differences in the tasks and sub-tasks. Initially, the purpose of the complete scenario is to extract spatial information embedded in the text to increase the number of geotagged tweets. Finally, use an improved geotagged data set to look up for the research question with new spatial information that may provide extra spatial details to answer the research question.

The implemented workflow consists in a few questions, the main purpose of the questions are to separate the scenarios and provide options to each user; the first question focused on the management of the data, the second question focused on the data set size, the third question (only in the complete data set branch) refers to a quest of spatial information within the text, and the final question looks to answer the research question. At the end of the last question there is an option to provide insights, this is a loop question, in case of a negative answer this question was designed to improve the data set by providing insight from the first loop analysis, this is represented with a dotted line in Figure 3. Since is relatively easier to apply the scenario with only geotagged tweets, running this scenario might provide valuable information without the necessity of apply the complete scenario.

First Question: The first question splits the management task. The question tries to find if the current system can load the information collected from the social media source. In both branches the data management is composed of storage and retrieval, both sub-tasks are different due to the necessities of each user, the current system and the size of the data set. In the workflow, the data management is divided between a local machine and the cloud management which is divided in structured data management and unstructured data management. A database management system is usually used to retrieve information; this can be installed in a local machine, cluster or the cloud, depending on the resources and the volume of the data set.

The characteristics or each system define the necessity to the usage of a DBMS and the technology associated. In some cases, the use of a DBMS is not required, and the information can be processed from a unique file, however, in the workflow the first question open the possibility to load the data set in a local machine or a cloud environment, depending on the characteristics of the data set and the requirement of the user. Most of the cloud providers have a DBMS integrated a services, structured and unstructured, also the files can be stored and retrieved from the cloud without a DBMS.

A DBMS organize the data set on logical structures defined by a model, this allows to query a database and filter information from the source, which are two sub-tasks of the workflow. Without a database, query and filter task become more complex and time consuming. The disadvantages of a DBMS are the complexity, size and performance, some DBMS require additional resources and even an administrator, designer and developer, this type of databases are expensive and complex.

Since social media is associated with the generation of massive amounts of information a DBMS would be a good option to store and retrieve information. There exist some considerations to select a DBMS such as the structure of the data set. There exist three types of data sets, structured, semi-structure and unstructured. A structured data set is typically associated with a SQL relational structure, a unstructured data set is associated with a non relational NoSQL structure, the semi-structured data set can be associated with both. A NoSQL is classified in types of databases that follows a data model such as column, document, key-value or graph. A SQL approach is associated with a traditional rows and columns schema. In the case of a structured data set some recognized technologies are Oracle SQL or PostgreSQL, and in the case of a non-structured data set MongoDB or Cassandra are one of the most popular.

The cloud provide some services that make it relatively easy to set up, maintain and administrates a DBMS. The advantages of a DBMS on the cloud are the scalability and performance. A DBMS in the cloud is dynamic and allowing simple or complex data sets, the resources on the cloud are easily scalable, unlike a physical server or cluster which a regular database upgrade and administration may have expensive costs. Some of the disadvantages of the cloud is that the administrator loose the complete

control of the servers, the data set is totally dependent on the provider, to transfer information require a good internet connection, and to switch between providers have several issues.

The workflow contemplate the cloud storage and retrieval as a service, the complete data set scenario have two options a structured and unstructured data management, the selection depends on the characteristics of the data set such as type of structure, volume and velocity of retrieval.

3.3.1. Sampled Geotagged Scenario

Second Question: The second question of the implemented workflow split the scenario in the sampled and complete scenarios. This scenario contains three sub-tasks; each sub-tasks is focused on process and provide inputs for the following tasks; the tasks in here are pre-processing, NLP and Analysis. The geotagged information represent a fraction of the complete data set, this scenario contemplates the information collected from the streaming API, depending on the characteristics of the search and keywords the information can be distributed in different parts of the globe. The main goal of this scenario is to seek for an answer to the research question or provide insights for further analysis.

Pre-processing Task

The pre-processing task focus on querying, filtering, removing, and cleaning. In the workflow, these tasks are numbered in the superior left corner, this numbering is only a recommendation and may be subject to changes depending on each data set characteristics.

The first sub-task consist in querying the database or file to get only the records that have valid coordinates of latitude and longitude. The advantage of querying is to reduce the amount of records by soliciting only the information that is required; this saves time in the following procedures. A query may be simple or complex, depending on the user necessities, in some cases an efficient query may help to solve the research question almost in the first stages of the workflow. It is important to mention that in the situation where there are no geotagged tweets, it will be required to collect more information.

The sub-task named removing, focus on locating some specific patterns from some social media users. Some users are excluded from the data set analysis. Some accounts may be fake or may be skewing the sample by adding the same tweet several times with the same sentence and words, because of this reason it is necessary to do locate this type of users or records, the analysis may focus on the users and their behavior.

The last suggested sub-task on the workflow is called cleaning, which look up for special characters, symbols, numbers and URL's that are not essential for the NLP analysis and remove them from the text. It is important to remember that the data set comes from social media, this type of information has several sources, and for this reason tends to be noisy. Tweets may have several types of words, symbols, URL's, emoticons or numbers that do not represent any contextual message on the tweet, removing and cleaning this type of data might be challenging. Fortunately people in several institutions have been working on the comprehension of human language by computers, this work includes tasks to analyze text, and remove the additional unnecessary information from it.

Natural Language Processing Task

The processing task will use NLP techniques such as tokenization, stemming, stop words and lemmatizer, the following are a brief explanation of each method and their use on this workflow. It is essential to clean and filter the processed words before this task is applied.

The language detection sub-task classify the information depending on the the language of each sentences or document. The language detection is applied prior the stop word and tokenization, the reason is that the stop word sub-task look for specific words in a language list, prior apply a stop word list, the language should be detected. There exist some words that do not contribute with any information to the analysis, these words are removed from each line of text in the tweets.

The tokenization sub-task is applied to separate every word in each tweet; this is a necessary step to simplify the text and analyze the message that each tweet contains. The tokenization purpose is to separate each word into different terms; this technique can separate specific symbols, on microblogs such as twitter. There exist several pre-trained tools that are capable of fulfilling the tokenization of a sentence. Typically after the tokenization and language detection, the stop words or stop-list technique is applied, the removed words are widespread and do not contribute to the semantic analysis. On the workflow, it is suggested to use it after the tokenization.

The stemming and lemmatizer (Figure 3) techniques have the same order numbering on the top left corner; this denotes that the order is not significant when applied this procedure in the workflow. Both techniques are used to normalize the text in each tweet, in the workflow they are suggested after the stop word technique. The stemming procedure shortens or reduces the verbs into their morphological root while the lemmatizer analyses specific dictionaries to return the lemma of each word. The Table 1 display examples of the suggested sub-tasks for the workflow.

Sub-task	Example
	@miley This tweet is an examples of Natural Processing Language 12345, Http://www.google.com.mx
Cleaning	This tweet is an examples of Natural Processing Language
Language Detection	english
Tokenization	['this', 'tweet', 'is', 'an', 'examples', 'of', 'natural', 'processing', 'language']
Stop List	['tweet', 'examples', 'natural', 'processing', 'language']
Lemmantizer	['tweet', 'example', 'natural', 'processing', 'language']
Stemming	['tweet', 'exampl', 'natur', 'process', 'languag']

Table 1: Tweet processing example of NLP sub-tasks

Analysis Task

The task contains three sub-tasks, frequencies, top modeling, and visualization, numbered and applied in the same order. One of the sub-tasks in the workflow is called frequencies, the goal to analyze the frequencies is to understand the distribution of the words collected examples of these are the most mentioned words, the most common hashtags, words combinations with n-grams, and by sentence with a co-occurrence matrix. The expected result is to visualize and comprehend the weight and behavior of each queried word in the data set.

The Frequency provides a first insight of the data set, it is a powerful tool to look up for information related to the research question, in this case, the suggested analyses with the most frequented mention words are: hashtags, N-grams and a specific combination of words, this analyses will help to understand what are the most frequent terms in the data set but also their relationship with the initially queried keywords. The co-occurrence matrix is part of the frequency analysis, this type of analysis look up to find the most common terms with a specific word.

The workflow incorporates a machine learning unsupervised clustering as a suggestion, to analyze the data set text and test if the data set contains valuable information for the case study. Based on the experience of Yang (2016), this method proved to be adequate to cluster all the information and find a topic giving the distribution of the words. Due to the lack of information about the data set topic, the workflow sub-task topic modeling will categorize all the terms in different topics. The LDA model is a suggestion which typically is used in NLP for unsupervised topic modeling. It uses documents as a collection of words which categorize in semantic groups, due to their distribution of each semantic group a word is classified into a set of words that contains a clustered topic; the purpose of usage is to find the diversity of topics that the data contains. Nevertheless, this technique require some understanding of probability distributions such as Poisson and defining the input parameters of the model is not an easy task (Blei, Jordan, & Ng, 2003).

It is essential to visualize the information generated by the frequency analyses and also the spatial information, typically social media data is skewed by cities where the population density is high. The main purpose of the visualization sub-task is to represent the information from the social media with a space and time context. Nevertheless visualize this type of information may reflect a different behavior in

space and time, this graphical visualization is necessary to see different patterns , this sub-task expect to display some extra information on the research question.

3.3.2. Complete Data set Scenario

This scenario is divided by the second workflow question, which denotes the user preference of a data set and type of environment. This scenario contemplates the complete data set within a cloud architecture due to a necessity to explore the data with additional computational resources. This branch focus on the same research questions that the previous scenario but including other frameworks and techniques integrated into new sub-tasks. Some of the differences with the sample scenario are the inclusion of sub-tasks, the order of appliance of some sub-tasks, and the inclusion of the "Extract of Spatial Information" as a main task. This addition is focused on spatial users who look to go further in search of spatial information within the micro-blog text. The main tasks of the complete data set scenario are pre-processing, NLP, Extract of Spatial Information and Analyses (Figure 3). It is essential to mention that the workflow contemplates insights from the geotagged scenario that can be interpreted as valuable information to initialize the complete scenario. In the case that both scenarios fail to answer the research question, the suggested option in the workflow is to improve the collection of words for further research.

The cloud environment can provide access to more resources, in this case, the necessity of increase the processing and storage capacity is implied in the second question. The scenario was designed for the users with means and knowledge that desire to explore a stored data set in the cloud. The cloud offers several services such as financial, processing and, storage. The workflow focuses on two components on the cloud, the storage, and the processing component, both have specific characteristics and are necessary to review and define specific configurations to provide the amount of processing power required to complete some of the tasks and sub-tasks of the workflow.

All tasks and sub-task are numbered in order of appliance, depending on the case some of them may not apply, and they can be skipped. The workflow design contemplates information from micro-blogging social network, other types of social media such as those that contain videos or images as their main objects of analysis may not apply for this workflow.

Pre-processing Task

Two tasks were defined on this part of the workflow, removing and cleaning, both tasks are designed to remove records and characters in order to make the processing more efficient and also to have confident inputs for the following models. In the cloud and with the complete data set this type of sub-tasks may

take some time without the right implementation. The removing sub-task consist in identify specific users or records and remove them, and the cleaning part keep some characters out of the analyses. The main difference with the local machine scenario is that the data set size on this scenario may be enormous.

The difference between the pre-processing part of the workflow and the previous in the geotagged section is that with the complete data set it is suggested to implement a big data processing approach such as Map/reduce or in-memory processing. In the case of an unstructured and massive data set, one possibility is to utilize Map/Reduce. The map phase is responsible for extracting the message field from each tweet and further processing it with the purpose of obtaining a "bag of words" representation of the tweet. In the bag of words model, a text (such as a sentence or a document) is represented as a multi-set of its words, disregarding grammar and even word order but keeping multiplicity (Moise, 2016). Each map task reads its assigned input block and processes each line of the block, corresponding to a single tweet. The in-memory approach is different but not exclusive from the map/reduce approach; this model uses the RAM to process the information avoiding the disk access, allowing to increase the speed of the processing.

This type of techniques combined with some regular expressions may find and remove special characters in the data set. A regular expression can be defined as a pattern defined by a sequence of characters, this type of patterns are commonly used in programming as tools. A regular expression can locate and remove patterns such as a) filter out all non-latin characters, b) remove numeric characters c) special characters or user-defined characters.

Natural Language Processing Task

The Natural Language Processing task in the complete scenario will focus on the same tasks that the previous scenario, but targeting to implement the task ans sub-tasks in a cloud environment with a regular framework or with a big data model framework. The task is split in 5 sub-tasks language detection, tokenizaition, stop words, lemmatizer and stemming. However the sub-tasks are very similar to the previous scenario, the main difference are in the the language detection that may require a different approach to define the language per record or the lemmatizer that in the big data framework require to be trained. One option is just to implement NLP on a cloud environment, with regular programming schema or to apply NLP for big data which require specific methods of programming and order, there are some libraries that are designed for this type of framework.

To apply this type of tasks within a cloud service have different levels of difficulty, depending on the architecture selected but mostly on the framework. It is important to distinguish that the framework add technical complexity to the sub-tasks, an example is the language detection which needs to be trained for before employ it or stemming which need specific designed libraries to be applied in a Spark framework.

Extract Spatial Information Task (Optional)

In order to collect spatial information from a text message it is important to identify the different entities to extract place names making use of geonames gazetteer, on the workflow this process may be simple or complex depending on the requirement of each case study. Parsing is defined as the process of analyzing text sentences composed by tokens, to determine the grammatical structure in agreement with formal grammar rules. The parsing sub-task is suggested to identify the type of form from a sequence of characters, some approaches or algorithm are more practical and other more powerful, this tool is usually performed previous to the geocoding in order to select only the words classified as nouns.

In the workflow a sub-task suggestion is geocoding, which can be defined as the procedure of transform a country, township, place or addresses into geographic coordinates (Google, 2019). The objective of this sub-task is to add spatial information to records that do not have it, this extra information is extracted from the context of each text message.

Nowadays there exist some tools that automatically apply both tools called geoparser. The geoparser has two main components, the geotagger, which is responsible for place name recognition, and the georesolver, which is responsible for georeferencing (Grover et al., 2010). This type of analysis it is not an easy task and requires a connection with a gazetteer with geographical names and the identification of the geographical entities in each Tweet.

AnalysisTask

This part of the analysis tries to answer the research question, the following sub-tasks are implemented, frequency, topic modeling, geoprocessing and visualization. There exist technical differences between the two scenarios, due to the complete scenario present a technical challenge since use a variety of frameworks or technologies. The frequency tasks suggest the word count, bigram analysis, hashtag counting and co-occurrence matrix, all this analyses are the first frontier to display the content of the information.

On the workflow the topic modeling utilize LDA as a suggestion to find the different topics in the data set, however implement this type of analysis on a big data framework may be challenging. The sub-task of geoprocessing on the workflow purpose is to process the geocoded information and provide inputs for the visualization sub-task. The visualization sub-task purpose is to provide to the user valuable and understandable information of the data set.

4. PROTOTYPE DESCRIPTION AND CASE STUDY CHARACTERISTICS

A prototype is defined as a preliminary interactive model based on an idea, the objective of a prototype is to explore and express an idea with an ordered structure (Houde & Hill, 1997). This research is considered a prototype because is expressing the idea of exploring social media represented with a sequence of stages or tasks, to explore the possibilities of adding a geoprocess in a structured sequence.

One objective of this research is to implement a prototype able to analyze the case study data set. To accomplish this objective, a workflow was designed with the purpose of checking the stored data set relation with a research question. Some investigations have generated prototypes associated with geoinformation. Some of these prototypes create information from social media feeds by the integrating, harvesting, processing and modeling geosocial information in a prototype (Croitoru et al., 2013). Another example related with geosocial media was developed by Chang who developed a real-time geocollaboration prototype in a geospatial environment (Chang & Li, 2013). The mentioned examples are presented as evidence that a social media idea or concept can be implemented in a prototype.

The prototype consists in two stages, one stage that analyze the collected information with a sample of the data in a local machine, and another stage to make a similar approach but in the cloud, with the complete data set, and with a big data infrastructure. The following sections describe the characteristics of the case study data set and the prototypes in a local machine and the cloud.

4.1. Case Study Description and Data set Characteristics

The case study focuses on answering whether a collected twitter data set contains relevant information on actual tick bites and tick bites risk. These tweets were collected using a query with Dutch keywords that may be related with a tick bites or outdoor activities that increase tick bite risk. The research question of the study case is “The data set contain information related with tick bite events?”, the prototype may answer the question or provide insights related with the data set.

The data set is composed by a set of Tweets that have been collected since 2015, the initial aim of the collection is to track tick bite events in the Netherlands. The data set used for the case study comes from the Twitter Streaming and Search Application Programming Interface (API), this standard API allows to search and collect tweets queried by location or specific keywords. The API collects information as Tweet Objects which are composed by more than fifty attributes such as tweet id, user id, text, date of creation,

coordinates, etc., the most significant attributes for this study are related with the spatial information and the text generated by the user, Table 2 describe the most meaningful attributes for this project.

Attribute	Type	Description
created_at	String	UTC time of Tweet creation.
id	Int64	Unique identifier for this Tweet as an integer number.
id_str	String	The string representation of the unique identifier for this Tweet.
text	String	Status update from the user in UTF-8.
user	User object	The user who posted this Tweet.
coordinates	Coordinates	Nullable. Represents the geographic location of this Tweet as reported by the user or client application. The inner coordinates array is formatted as geoJSON.

Table 2: Twitter Object Attributes

**Information extracted from Twitter developers*

This data set was collected not by using location, but by using the following searched keywords, from now on these are referred in the text as initial searched words:

Searched Terms	English Translation	Language
Fietsen	Cycling	Dutch
Kamperen	Camping	Dutch
Lopen	Walking	Dutch
Spelen	Play	Dutch
Teek	Tick	Dutch
Teken	Symptom	Dutch
Tekenbeet	Tick bite	Dutch
Tekenbeten	Tick bites	Dutch
Wandelen	Hiking	Dutch
Wandeling	Hiking	Dutch
Camping	Camping	English
Lyme	Lyme	English

Table 3: Set of words extracted from Twitter (Initial searched words)

The words were chosen based on typical outdoor activities related with increased tick bite risk. Although we can expect (actually aim for) Tweets from the Netherlands, several factors may lead to Tweets from

outside the Netherlands: the internationally used terms “camping” and “Lyme”, Tweets sent by Dutch tourists, and the closely related languages of Vlaams (spoken in neighboring Belgium) and Afrikaans (spoken in South Africa). Hence the majority of the records on the complete data set do not have any spatial reference, less than 1% is have spatial reference from Twitter (geotagged). For the remaining 99% a geocoding technique is applied to search for any spatial information within the context of the text.

The Geoinformation Processing department of the University of Twente has been running a query with the mentioned initial keywords since April 2015. The collected data stored in the original JSON format as text files, and selected fields were processed and stored in a Postgresql table. This table was used and queried only for the sample scenario in a local Machine. For the complete scenario in the cloud the raw data has been used. The data set have 31,944 text files (one per hour) have a combined size of about 175.1 Gb. The total amount of records with valid information is 34.5 millions, duplicate Tweets have been removed, resulting in 22.3 million records.

4.2. Sample Scenario in a Local Machine

The local machine prototype was developed mainly with python and SQL languages. The most relevant tools and technologies associated with the developing of the prototype are related with data structures and analysis (pandas library python), retrieve and filter information (SQL querying), NLP analysis and implementation (nltk topic modeling python library) and tools for visualization and mapping (matplotlib and folium). The prototype on this scenario retrieves a plots with the language records founded, the most mentioned words, and display the most mentioned hashtags in Dutch and English, also display the co-occurrence of each initial searched term.

The raw format from Twitter is in multiple JSON objects in one document. The prototype extract the information from a server with PostgreSQL interface, this technology were used to store and retrieve the information on this scenario. The display of the database is in rows and columns, splitting each record in rows and attributes in columns. In raw format each geotagged Tweet comes with a field named coordinates which contains the latitude and longitude, in the database they are stored as rows and columns, the sample scenario focus on this records.

The information extracted from the DBMS are text, user id, latitude, and longitude. In a first approach, the database was queried only with valid geotagged information. The query only applied to the rows of latitude and longitude. In order to extract only the records within the Netherlands country, a bounding box coordinates can be ingested in the query; another option is to apply a spatial query within a boundary of a geometry. Both approaches have advantages and disadvantages, the bounding box approach only needs two coordinates to filter the data but is not completely accurate, and some records outside the country boundary remain in the filtered data set. The geometry boundary filter all the records outside the

boundary of an area, polygon or country. One disadvantage is that the data set and the boundary must have a column with their geometries in the DBMS and PostGIS must be installed.

The following step is to locate specific users that may bias or influence the analysis. Users who tweeted more than 3000 times in 4 years and the users who repeated the same tweet three or more times were removed for further analysis. These analysis is based on the work of Lansley (2016) who remove users from the sample by the tweet frequency of more than 2.16 tweets per hour, and repeated messages with more than three times. The prototype locates and count all the tweets per user and add the user's id into a list; the same process is applied with the repeated tweets. Eventually, the tagged users are excluded from the analysis.

The next step is to clean specific characters and symbols that are not useful for the analyses but to keep those who express a particular sentiment. The program removes specific characters and symbols from the Tweets text, this step was made with a regular expression which remove all characters except numbers, letters, spaces and hashtags, these last were not removed to track the most used hashtags per word.

The next step is to determine the language of each tweet. Language classifiers are pretty accurate when the input do not have serious misspelled words and there is only one language on the phase or sentences. Even though most of the searched words in Twitter were in dutch, in microblogs and especially in Twitter there exist a mix of languages, Table 4 showed two examples classified with different tools, the purpose of these table is to exemplify the variety of languages that each Tweet can contain, and the difficult of classify some records:

Example	Classifiers		
	langdetect	**langid**	**polyglot**
We think so too! Great for hiking, backpacking, camping, etc..	English	English	English
Tighadouini: 'Spelen in Primera División droom die uitkomt	Spanish	Afrikaans	Dutch

Table 4: Example of Language Classification python tools

The Tweets were classified differently by three different language classifiers, because of these, the tools were tested with a sample. Three python tools were tested to detect the language of each tweet (langdetect[1], langid[2] and polyglot[3]), with a sample of the 300 tweets the three tools were tested and ranked, with an accuracy of 90% of the sample polyglot proved to be the more accurate classifying sentences and the langid was more accurate when classifying individual words. The classifier selected for this part of the prototype was polyglot, this is a pre-trained language identification tool, which used word embedding to classify languages. This method uses the words as vectors of a distributional semantic model, and words with similar vectors are classified in the same language or vector, this is the principle of word embedding (Al-rfou & Perozzi, 2013). The prototype counts all the records that were classified in English, Dutch and the records classified with another language, and report it in a chart.

1 https://pypi.org/project/langdetect/
2 https://github.com/saffsd/langid.py
3 https://polyglot.readthedocs.io/en/latest/Installation.html

On the local machine prototype each record is classified by the language and then applied the NLP sub-tasks. Once the language of a tweet is defined, then the task of tokenization starts. The tokenization splits the sentence into individual elements or unigrams, in the prototype an array was created with all the tokens per record. There are two different stop word lists one list for English and one for Dutch. If any unigram matched any stop word, it is removed. The stop words list is predefined per language After removal of stop words. Then the text in Dutch was stemmed using a snowball stemmer[4] and for the English records a porter[5] stemming tool was used. Both stemmers belongs to the Natural Language Toolkit (nltk[6]) library, an example of this process is depicted in Table 1. The lemmatizer was not included in the prototype due to difficulties of installation of the lemmatize tool frog[7], this seems to be the only tool available to lemmatize in Dutch.

The program analysis focus on the frequency of words, n-grams and co-occurrence matrix. These analyses are applied to check which are the most common words and to provide first insights of the data set. The first analysis is to count the most frequent terms in the sample, following by the contiguous sequence of words in a bigram analysis, and a general a co-occurrence matrix analysis for the initial queried words.

Finally the prototype use a interactive tool for visualization, this tool produce a HTML file that show the daily tweets in a heatmap (Figure 4). The tool used is called folium[8] which is a python adaptation of Leaflet, that is defined as an interactive platform for JavaScript mapping.

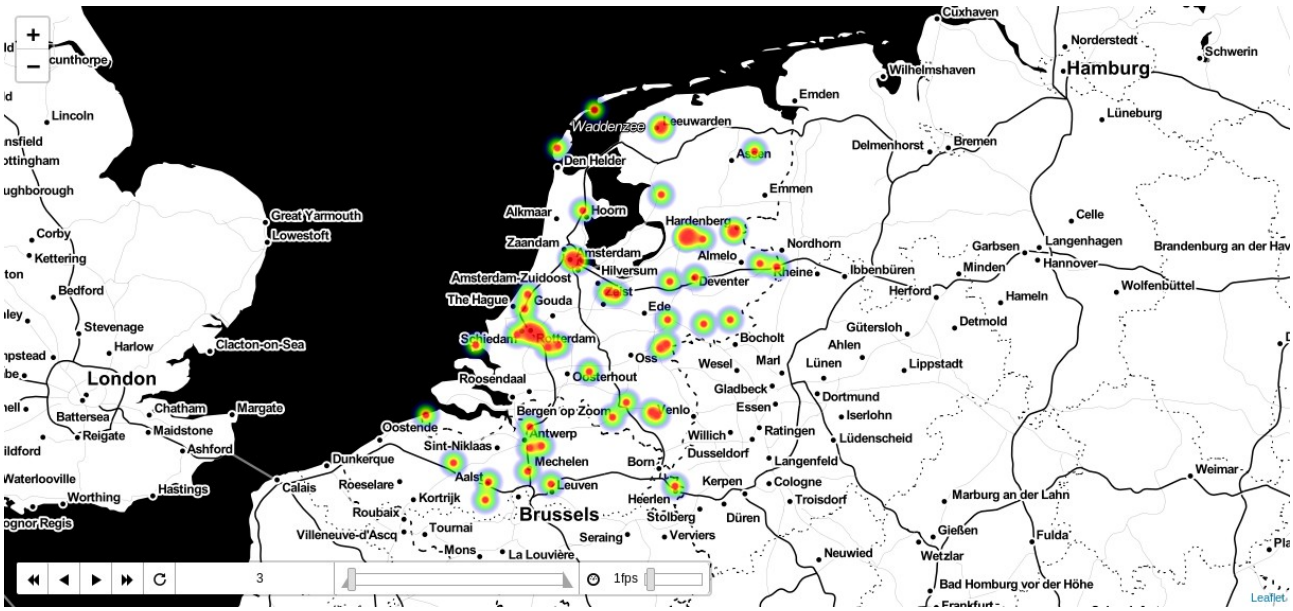

Figure 4: Example of the interactive visualization created by the prototype

4 https://snowballstem.org/
5 http://snowball.tartarus.org/algorithms/porter/stemmer.html
6 https://www.nltk.org/
7 https://languagemachines.github.io/frog/
8 https://github.com/python-visualization/folium

The LDA analysis was not implemented on this prototype due the time limit of this research, but with some modifications the data set might be usable for an LDA analysis, some of the most popular tools are gensim[9] or scikitlearn[10].

4.3. Complete Data set Scenario in a Cloud Environment

The prototype in the complete data set scenario used the cloud service EMR from Amazon and big data framework, due to the possibility of process massive amounts of data. The complete raw data set has 172 Gb of Twitter records, this may not be considered big data, but social media easily can generate more records, thinking on this possibility the prototype was designed on this type of framework. Amazon provides many cloud services, some of them specialized in storage, processing and processing big data, the prototype uses these three types of services on the cloud.

Briefly, the prototype makes use of Spark Python API (Pyspark), within Pyspark the following tasks were performed, removal of records, cleaning symbols and special characters, Tokenization, Stop Words, Stemming and a geocoding approach based on matching data set unigrams and a filtered gazetteer. The architecture is based on a cloud cluster which performs parallel computing on a Spark framework. The following sections describe in detail each one of the services used and the prototype application.

4.3.1. AWS Introduction

For this prototype, Amazon Web Services (AWS) was selected as the cloud provider for logistical reasons (available funds). The described functionality is also available (under different trade names) from other cloud service providers. The project requires services for storage and computing facilities; the use AWS Elastic Compute Cloud (EC2) which is considered a Platform as a service provides a configurable compute service in the cloud, Simple Storage Service (S3) provide storage, and Elastic Map Reduce (EMR) which provide a per-configured virtual machines able to do parallel computing with a Spark framework. These three amazon services were implemented to store, process and analyze the complete scenario data set.

The EC2 service of Amazon is based on the virtualization concept, which allows emulating a computer system such as Windows Server, and a few Linux distributions. In Amazon, this type of virtual machine is called instances. Each instance has internal components such as cores, GPU, HDD, and RAM. Another feature used by the prototype was the S3, called bucket, which allows storing files as objects instead of a traditional file system, and share between instances within the cloud environment, this type of storage is defined as an object store which manages the data as objects. The EMR service is the pre-configured platform that provides access and resources for processing vast amounts of data. EMR uses EC2 instances

9 https://radimrehurek.com/gensim/
10 https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html

for processing and S3 for storage, besides depending on the requirements of each user, the number of EC2 instances varies. The pricing depends on the components of each instance and the region it is located, the range of prices in EMR varies from USD $0.026 - $0.270 per hour of usage. For a batch process is recommended a general architecture, focused on the CPU. To run a frameworks such as Spark the memory selected is critical and the storage.

A Secure Shell (SSH) is a protocol to administrate one or several computers in an insecure network remotely. Typically, a key pair is used to login into an EC2 instance with an SSH protocol. This protocol is commonly used to provide access for users, and also for file transfers or automated processes. The previously described method allows to connect an instance to a local machine, but the account administrator can generate user or roles to grant permits to a variety of services that the provider provide. In general, the Amazon root administrator carries out the security through the Identity Access Management (IAM) platform which is used to define users, groups, policies, and roles within an account. Typically the administrator of the account set a variety of functions that each user have, also policies related to the web restrictions of services, instances, and connectivity.

To manage Amazon services from a local machine, there is a tool called AWS Command Line (CLI) which provide commands to admin most of the Amazon services. The tool needs to be installed and configured, with IAM credentials the user can send commands from a local machine to services such as EMR or S3. This tool allows to create batch processing jobs in a cluster and also to set up and configures one or several instances in the cloud environment.

4.3.2. Prototype Application

This section is divided into two. The first section describes the configuration and implementation of the EMR cluster and the second section describes the sub-tasks developed in Pyspark. On this section two files are required, the complete data set files, and a filtered gazetteer containing only unigrams of the required area.

The complete data set scenario utilizes EMR service to achieve the processing and analysis of the complete data set in raw format. The following description applies for this prototype, but the arguments should be modified depending on the characteristics of the data set and the user requirements.

The programming was developed with Pyspark and AWS CLI interface. To create a cluster within Spark on EMR is necessary to know the type and number of instances required previously. Also the number of nodes, executors, memory per node and cores per node, this can be defined in a JSON Spark configuration file. This type of information is essential to tune Spark and YARN efficiently because every node requires a certain amount of cores and memory to be able to process the information efficiently in

the cluster. As an example, the configuration of a cluster with 10 instances with 16 nodes and 64 Gb of RAM, requires 7 executors per node, with 2 cores per executor and 8 Gb per executor (Shipman, 2016). The Spark configuration in the prototype was established with CLI and a JSON file looking to optimize the resources on each node. Once the cluster is created an id is displayed by the system, this id will be used to send jobs to the cluster.

Once a cluster is running, the next step is to send a job with a CLI command. The EMR CLI interface has several options to initiate a job in an EMR cluster, for this prototype the specific instructions were the following, 1) to start in a cluster mode and provide the cluster id, 2) the manager of the clusters is YARN, 3) the number of executors, cores, and memory used by each node, 4) the location of logs, Pyspark file script, input files, and output files in S3, 5) the action taken in case of failure, and finally 6) The extra packages installed in the cluster (databricksCSV[11])

The EMR cluster configuration was completed with CLI interpreter, to create a cluster from the command line is necessary to install and configure CLI. To complete the configuration is necessary to add information such as the EMR version, debugging options, log files, configuration files and number and type of instances (detailed instructions[12]). The EMR version or label utilized in the cluster was 5.20 which contains a Spark version 2.4.0 and Hadoop 2.8.5. The prototype allows debugging which allows collecting log files in an S3 bucket; this is a useful tool to track possible errors in the application or the framework. In here the type and number of instances are declared, depending on the type and number of instances the process ingested in EMR will be completed.

For this prototype, the files were stored in JSON objects within S3 service. There are 31,944 files which can be loaded and distributed into the cluster with some specific commands. Sparkcontext allows to connect and distribute to the cluster and SQLcontext to load the JSON files into a SQL data frame.

The prototype follows the workflow tasks and sub-tasks with some exceptions. The first sub-task on the workflow is to remove specific users or records. On the prototype, this sub-task remove the duplicate records of the complete data set, also remove the corrupted tweets. This sub-task was implemented dropping the exactly same text records, if a record contains the same message twice then one of the two records was removed from the data sets. The cleaning sub-task remove all special characters except the alphanumeric characters and the number (#) symbol. Then the sub-task removed the URL's addresses, retweets and user mentions with a regular expression (Table 5).

Original	'Good advice! RT @TheNextWeb: Lets go camping tomorrow http://t.co/lbwej0pxOd cc: @garybernhardt #lopen'
Output	Good advice What I would do differently if I was learning to code today #lopen

Table 5: Cleaning sub-task Tweet example

11 https://github.com/databricks/spark-csv
12 https://github.com/mrfiesta/ProcessAnalyze_Tweets_Local_and_EMRSpark

The following sub-tasks applied in the prototype were related to language processing. For this processing, a specialized library in machine learning was required (pyspark.ml[13]) the tasks performed in this section were Tokenization, Stop Words[14], these tasks were applied in a pipeline. Then the stemmer is applied. Once the pipeline and stemmer are finished, the resulted tokens arrays were split by elements and counted, but these elements kept track of the original text as is showed in the Table 6. Due to some difficulties with the installation and configuration on EMR, the language detector and the lemmatizer was not applied on the prototype.

Original Text	Elements
This is a Tweet example	This
This is a Tweet example	Tweet
This is a Tweet example	Example

Table 6: Tweet example with elements split but with the track of the original Tweet

The geocoding task used a unigram database matched with a unigram gazetteer data set. A unigram is defined as an n-gram of one element. The gazetteer was filtered by country, and the administrative division reaching the 4th order. This filter resulted in 430 township and provinces. The data set was loaded as a dataframe since the original file was in a CSV format, it was necessary to load the file as a dataframe. The package 'com.databricks:spark-csv_2.10:1.2.0' Allows to load CSV files as frameworks, this package was installed and used. Finally, the individual elements were joined with the unigram gazetter, adding coordinates for the matching records. The following example shows one Tweet that contains a location within the context of the text.

Pop Up Tv De Week van de Teek 13 tot en met 19 april schiedam
(Pop Up Tv The Week of the Teek 13 to 19 April schiedam)

The text of the previous Tweet were tokenized and split into separated elements. The unigram gazetteer contains places with latitude and longitude (Table 7), the selected places gazetteer only contains unigrams. Bigrams or trigrams were discarded from the database. Once loaded the gazetteer may contain duplicated records of places with the same name but different locations, it is recommended to manually check these records on the gazetteer and select the most appropriate for the analyses. However, the prototype is programmed to drop every duplicate record on the gazetteer.

13 https://spark.apache.org/docs/latest/ml-guide.html
14 https://spark.apache.org/docs/latest/ml-features.html

Place	Latitude	Longitude
borne	52.3112	6.74404
baarn	52.20602	5.27144
rotterdam	51.88246	4.28784
drenthe	52.83333	6.58333
leiden	52.15274	4.4836
helmond	51.47968	5.65559
assen	52.99635	6.55255
tilburg	51.57787	5.06555
haaksbergen	52.15514	6.75404
tilburg	51.57787	5.06555
schiedam	51.9265	4.38675

Table 7: Example of a unigram gazetteer data

Tokens	Join
Pop	No match
Up	No match
Tv	No match
De	No match
Week	No match
van	No match
Teek	No match
tot	No match
met	No match
april	No match
schiedam	Match

Table 8: Example of tokens with matched elements

Once every element of the tweet were tokenized then the element is compared with the elements of the unigram gazetteer. The elements that matched were joined with the initial text (Table 8), and the spatial information was appended.

Finally, the prototype contains code lines dedicated to query (via SQL[15]) the initial words from the data set. These queries have the objective to count the occurrence of each initial queried words, with this information the co-occurrence matrix was created manually. The Table 9 is the result of the initial words counter and this is the input of the co-occurrence matrix. The first record 'tekenbeet' is the queried word, and 'lyme' is the most common word in combination with this term with 904 occurrences.

wordscounted	**tekenbeet**	**count**	**4389**
wordscounted	lyme	count	904
wordscounted	teek	count	533
wordscounted	teken	count	471
wordscounted	ziekte	count	347
wordscounted	app	count	292
wordscounted	rivm	count	259
wordscounted	via	count	251
wordscounted	mensen	count	240
wordscounted	ziek	count	234

Table 9: Example of initial words counter

The prototype still missed to apply the Hashtags and N-grams analyses as they are suggested in the workflow. The visualization of the results from the geocoded tweets was made manually with Qgis software.

15 https://spark.apache.org/docs/1.6.1/sql-programming-guide.html

5. IMPLEMENTATION RESULTS

The following results represent the sequential instrumentation of the implemented workflow in a local machine followed by the results of the execution on the cloud. For this first section the geotagged information was presented, and in the second section, the results of the complete data set are displayed. The implemented workflow was tested on a local machine and in the cloud with a big data framework service such as EMR. The research question look for spatial information related with a tick bite or tick risk event, all within the context of each Tweet.

5.1. Local Machine Scenario with Sample Data set

Following the workflow the initial tasks were the retrieval of the data from the database, this was accomplished by querying the database. Once queried, the sampled data set spread all over the world, in specific parts of the world such as north and south America, Europe and in specific parts of southeast Asia (Figure 5). The sample data set distribution is related with the combination of words in English and Dutch. The geotagged records collected are 275,697, compared with the total amount of records, the geotagged sample only represent 0.7 % of the complete data set. The Figure 5 represents the extension of all the geotagged records, some of the records are displayed in remote places, some of them are located in islands such as Hawaii or Isla de Pascua, and even some records can be delivered from boats. The records outside the Netherlands are not useful to answer the research question, and they were removed for further analysis.

The selection of keywords plays an important role in the spatial location of the information, the data set filtered with a valid latitude and longitude coordinates returned records extended around the world (Figure 5), due to the usage of common English words.

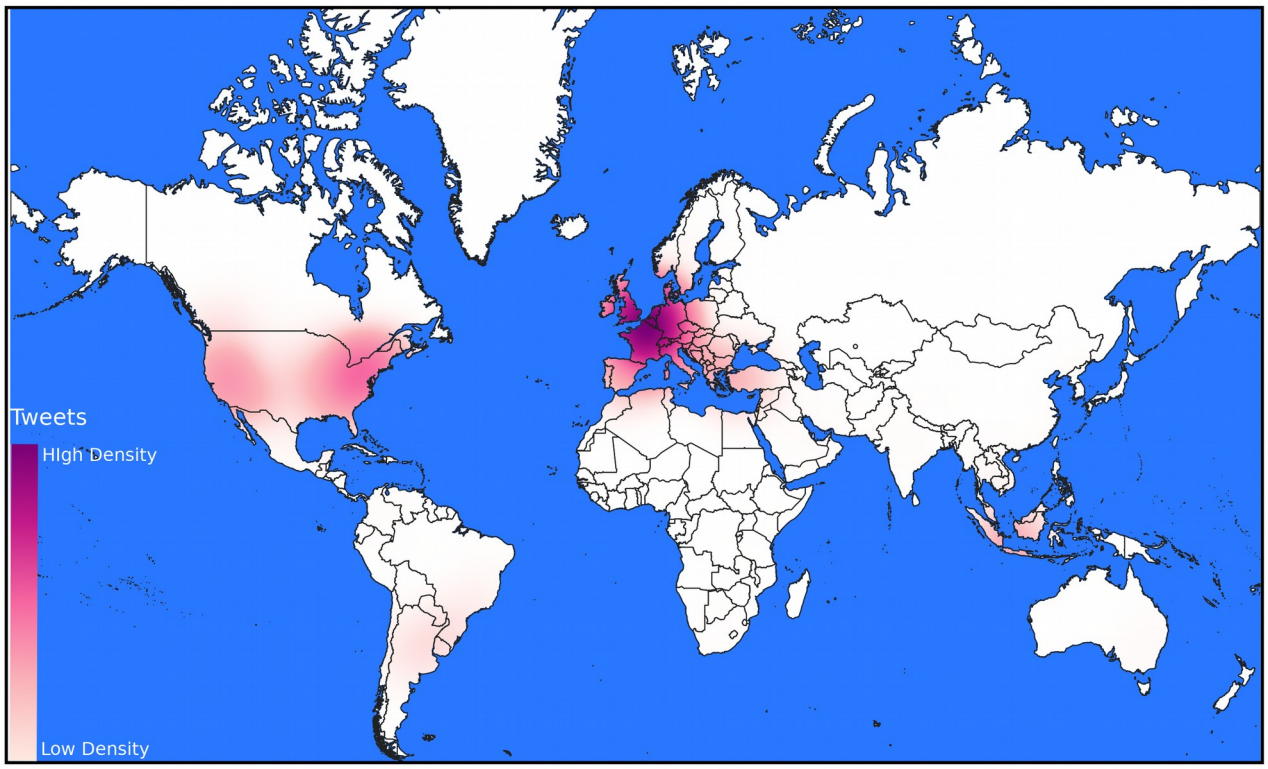

Figure 5: Map of Geotagged tweets of the data set

This data set has been queried showing a result of 275,697 total geotagged tweets, and approximately 25,543 tweets were located inside the Netherlands, with the bounding box approach the records were 31,082, the bounding box approach reduces the sample but is not entirely accurate. The prototype use the bounding box approach with a manual filter in GIS software; once applied the records collected by the boundary box exceeded by 18% outside the Netherlands, this depends on the population density outside the country but within the boundary box. In particular, this sub-task may be applied from the database with the use of a GIS extension in the DBMS or manually in GIS software. From the database, it is required to create vector field geometry for the data set, and intersect it with the geometry of a world boundaries vector to select only the tweets that are within the boundary of the selected area. The boundary approach was applied for this study case data set.

The sub-task of removing found six users hat repeated the exactly same tweet more than two times. This users were removed from the analysis, and the records eliminated (729 records). The cleaning sub-task removed all special characters from each record, the regular expression used for this sub-tasks was '[^A-Za-z0-9 @#]' which means all characters from A-Z, a-z, and 0-9 will stay on the sentence, also three special characters were added a space( ) character, at (@) and number sign(#). The purpose of leaving this characters was to count the hashtags with the number sign and remove user mentions with the symbol at from each record. Some of the findings in the cleaning sub-task is that even with the regular expression removing all special characters, some emoticons persist in the data set and they have to be removed with a special list in the code.

The language sub-task classification showed that 78.7% of the total geotagged data set were classified as Dutch language, 15.8% were classified in English and 5.5% were classified in another language. The geotagged records within the Boundary box, the majority were classified as Dutch (Figure 6).

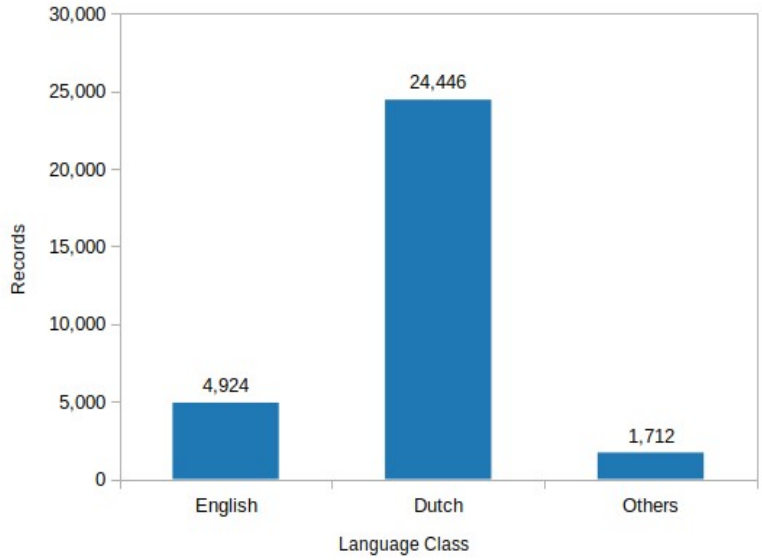

Figure 6: Language classification

The stop word list and the tokenization were applied without any complication. The Dutch stemmer fused some words such as 'teken' and 'teek' into 'tek', 'tekenbeten' and 'tekenbeet' into 'tekenbet', also the words 'wandeling' an 'wandelen' were fused in the term 'wandel', this is depicted in the Figure 7 and Table 10.The frequency and the co-occurrence analyses used the fused terms explained above.

The frequency of words sub-tasks (Figure 7) showed that words such as camping or play have a great influence over the data set, the more common the word it is, the more records will appear on the data set. Terms such as 'tekenbet' were mentioned only 21 times, and 'lyme' was mentioned 64 times.

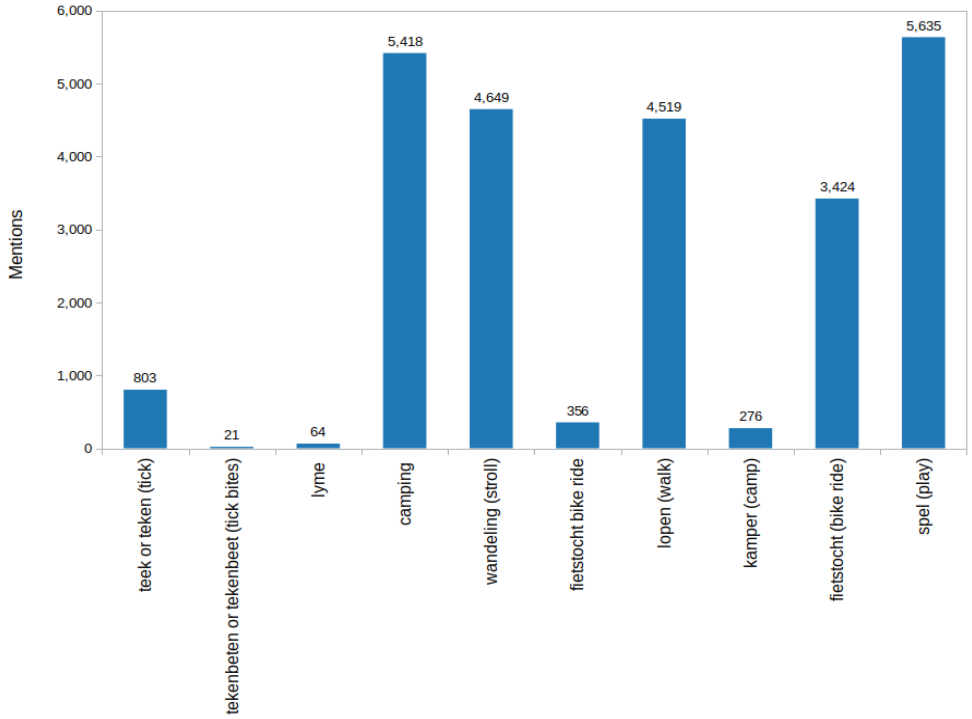

Figure 7: Initial searched word mentions in Dutch

The co-occurrence matrix (Table 10) showed a different behavior for each word, as an example the term 'tekenbeet' was mentioned 21 times, 19 times combined with the word 'hell', 16 with 'Family' and 11 with the name 'Manon', further research showed that there was a Lyme disease case with a girl named Manon and some media gave her support in some media like Twitter. Another specific case is given with a combination of words 'Lopen' and 'Voor.' This combination of words represents an event organized to fundraiser money for Lyme patients. The co-ocurrance allowed to find a particular combination of words that might have better chances to find records related to the questions topic, a bigram or trigram method approach may be an improvement for further analysis.

The terms 'tek' and 'tekenbet' had the lowest rate of mentions in the data set. For a data set collected during four years this might appear at a low rate of appearances. The terms 'tek' and 'tekenbet' seems to be related with the research question, and terms with several mentions such as 'speler' or 'camping' seems to be related to another topic. These assumptions were based on the co-occurrence matrix terms (Table 10), and the words associated with each term.

But most importantly, the use of common words had introduced noise to the data set. The words 'camping' and 'lyme' introduced several records outside the Netherlands boundary box, and the word 'camping' introduced noise off the context of the research question.

Matrix Co-ocurrence						
Word	**1rst Word**	**Ocurrences**	**2nd Word**	**Ocurrences**	**3rd Word**	**Ocurrences**
Teek or Teken	stat (state)	143	voor (in front of)	127	deze(this)	61
Tekenbeet or tekenbeten	hel (hell)	16	gezin (family)	16	Manon	11
Lyme	ziekt (sick)	41	dor (dry)	19	kost (costs)	15
Camping	voor (in front of)	174	weer (weather)	123	nar (jester)	108
Wandeling or Wandelen	heerlijk (lovely)	423	Lekker (yummy)	400	voor (in front of)	361
Fietstocht	organiser	55	hevo-fietstocht (hevo bike tour)	54	vrolijk (smiley)	54
Lopen	voor (in front of)	446	maar (but)	420	weer (weather)	369
Kamperen	voor (in front of)	25	weer (weather)	17	maar (but)	15
Fietsen	nar (jester)	379	voor (in front of)	319	dor (dry)	260
Spelen	we	575	voor (in front of)	534	Lekker (yummy)	327

Table 10: Twitter Initial searched words co-ocurrance matrix

For the visualization task a map displaying the density of the tweets in the period of 2015 - 2018 was developed in Qgis. The map displays the information within a determined boundary area. Figure 8, and shows the records within the boundaries of the Netherlands. On social media and specifically in Twitter the information tends to be in populated areas, the map confirms this was displaying more information near big cities like Amsterdam, Rotterdam. However some places such as “National Park De Loonse en Drunense Duinen” display a fair amount of records (2,270), these records may be related to outdoor activities like camping. Nevertheless 315 of these records correspond to a undetected bot. They were undetected by the prototype due each record contains a random URL and the prototype only recognize repeated records.

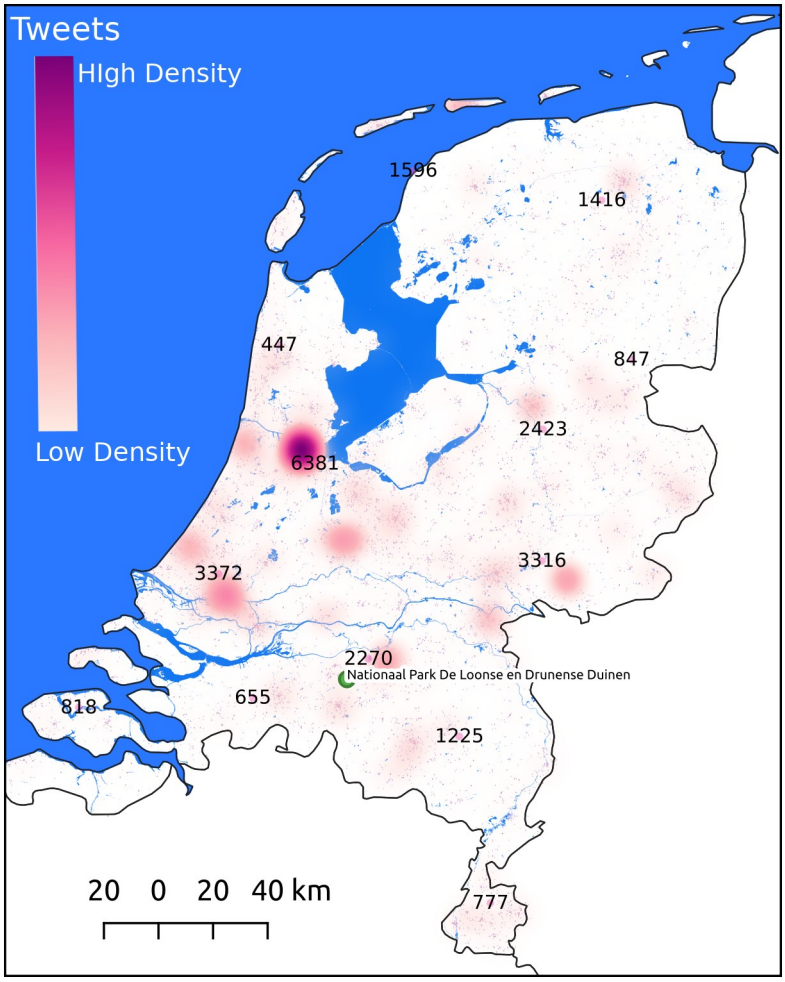

Figure 8: Map of geotagged tweets in the Netherlands 2015 - 2018

The prototype takes 9.21 minutes to accomplish all the tasks described in the prototype section. The total records in the sample scenario are 30,359, and the file size is 9.1 Mb. The local machine characteristics are four cores with an AMD A8-5545M processor with 8 Gb of memory.

5.2. Cloud Scenario with Complete Data set

The main findings on the complete data set were the reliability and fast processing of information in a framework such as Spark. The most frequent term in the analysis are the words 'camping' and 'spelen'. The geocoding retrieved several records, but after a manual inspection quite a few of them were incorrectly geocoded (explained below).

Just as a reminder for the following paragraphs, the prototype on the cloud implemented the following sub-tasks:

1) read all the records,

2) remove special characters (non alphanumeric),
3) remove specific elements of a tweet (URL's, retweets),
4) remove duplicated Tweets (12.2 millions),
5) load the gazetteer dropping duplicates,
6) tokenize all the elements,
7) remove the stop words in Dutch and English,
8) group and count all the words in the data set (frequency),
9) extract all unigrams and match them with the gazetteer (geocoding),
10) deliver two files with the most frequent words (JSON) and the geocoded tweets with latitude and longitude coordinates (CSV), and the necessary data to create the co-occurrence matrix (JSON).

The results of the frequency analysis is depicted in Figure 9, the terms 'camping' and 'spelen' are the most mentioned words in the data set. The term 'camping' is by far the most mentioned word, the term pop up in 59.1% of all records. Followed by the term 'spelen' with 9.3% and 'lopen' with 5.3%. The terms 'lyme'(5.1%) and 'teek'(4.5%) were in 4th and 5th position in the most mentioned ranking, and 'tekenbeet'(0.03%) resulted in the less mentioned term of the data set.

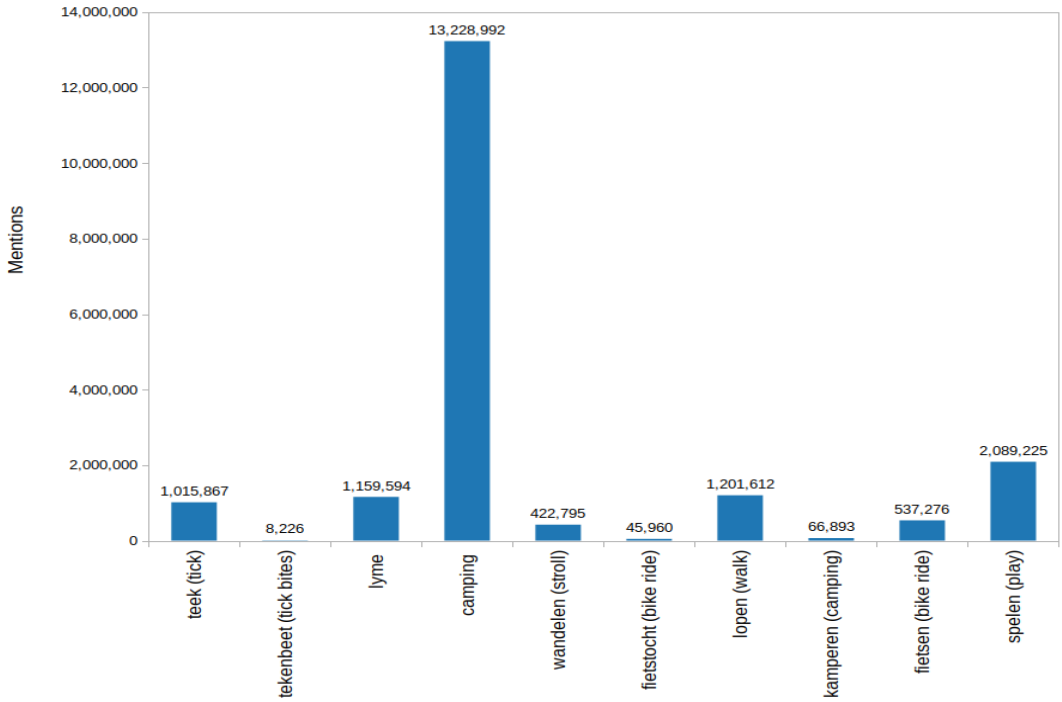

Figure 9: Complete data set most mentioned terms

Two frequency analyses were developed too observe the most common word associated with 'camping' and a combined analysis of 'tekenbeten' with 'tekenbeet'. Figure 10 showed the ten most common words with the term 'camping'. These words seem that they do not have any relationship with the research question.

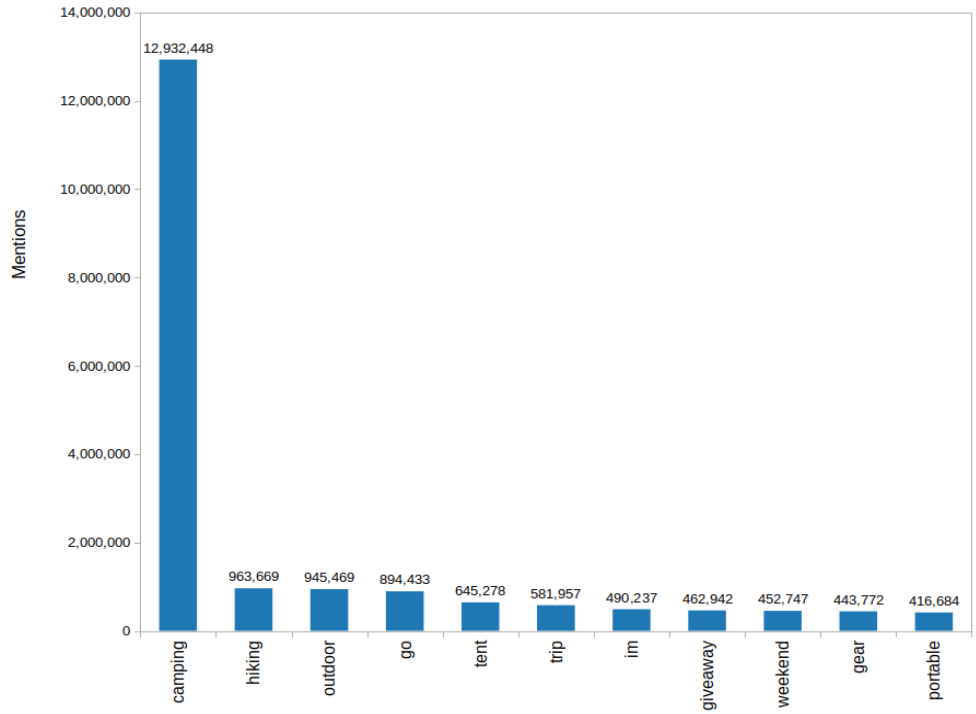

Figure 10: Filter with the most mentioned words with 'camping'

The Figure 11 shows that the less mentioned words in the data set 'tekenbeten' and 'tekenbeet'. However, most of the words in Figure 11 seem to have some relationship with the research question. These words might represent the first insights to evaluate word combinations for future analysis. The words associated with 'tekenbeten' seems to be related with the Lyme disease but also with the associated costs of this disease and the status of a tick bite.

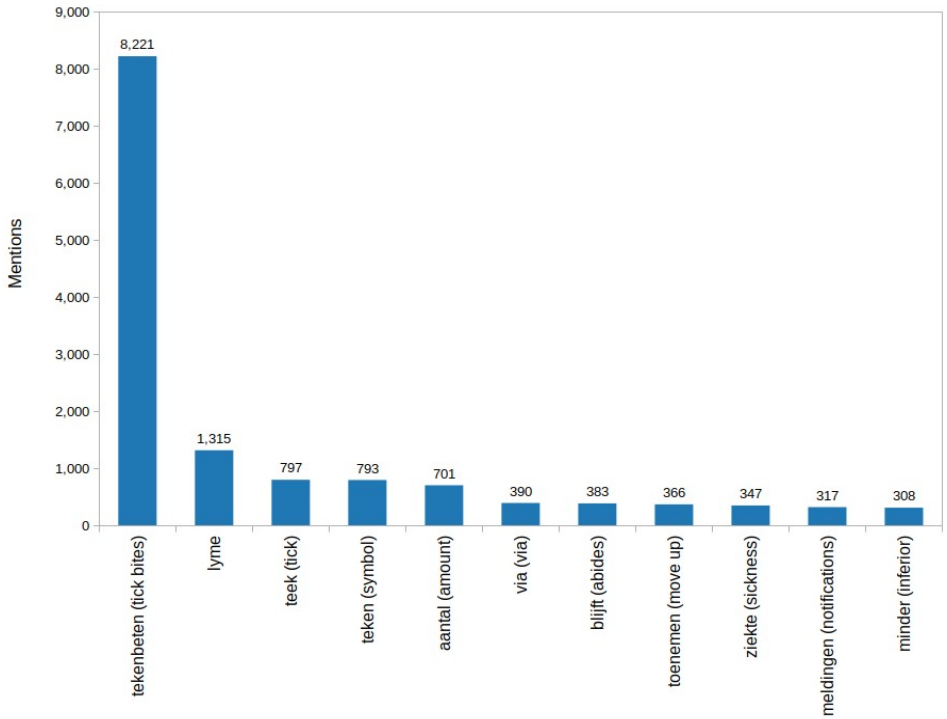

Figure 11: Filter with the most mentioned words with 'tekenbeten' and 'tekenbeet'

The Table 11 presents the co-occurrence matrix of the complete data set. The combinations of words that were mentioned most times are 'camping' with 'hiking' and 'spelen' with 'olympische'. Both combinations are not related to the research question. The combinations that are related with the research question are 'teken' with 'petitie' and 'lyme' with 'fietsen'. Surprisingly Lyme is related with an outdoor activity, which can serve as a clue for further research. It is important to mention that 'camping', 'wandeling', 'fietstocht', 'lopen', 'kamperen', 'wandelen', 'fietsen', and 'spelen' do not contain any words related to the research question. At the same time, words like 'teek', 'teken', 'tekenbeet', 'tekenbeten', and 'Lyme' have less occurrences, but the words seems to be related to the research question.

Matrix Co-ocurrence						
Word	1rst Word	Ocurrences	2nd Word	Ocurrences	3rd Word	Ocurrences
Teek	week	2,587	teken	1,868	lyme	1,809
Teken	petitie (petition)	96,931	staat (state)	50,591	goed (good)	26,400
Tekenbeet	lyme	984	teek	533	teken	471
Tekenbeten	aantal (number)	781	lyme	411	via	392
Lyme	fietsen	71,466	disease	57,380	lymedisease	30,855
Camping	hiking	962,440	outdoor	941,854	go	894,135
Wandeling	mooie (beautiful)	13,291	tijdens (while)	12,188	weer (weather)	11,824
Fietstocht	langs (along)	1,405	zondag (sunday)	499	zaterdag (saturday)	472
Lopen	wel (well)	74,791	weer (weather)	68,534	gaan (to go)	66,998
Kamperen	kampeernieuws (camping news)	11,041	camping	4,487	gaan (to go)	3,967
Wandelen	gaan (to go)	17,384	weer (weather)	16,251	lekker (tasty)	15,824
Fietsen	weer (weather)	32,396	ga (go)	27,000	wel (well)	25,053
Spelen	olympische (olympic)	101,790	wel (well)	90,513	weer (weather)	89,329

Table 11: Complete data set co-occurrence matrix

The unigram matching geocoder registered 478,691 records. This represents less than 3% of the complete data set. Still these records need to be inspected regarding their correctness, accuracy and precision. Most of them were captured in cities such as Amsterdam (34,607) or Rotterdam (21,301). Higher spatial resolution might be needed for the research question at hand. The unigram method for geocoding helped us to visualize the trend of mentioned places in a data set. However, to improve the spatial resolution, another method should be used. In Figure 12 can be visualized the unigram geocoder results.

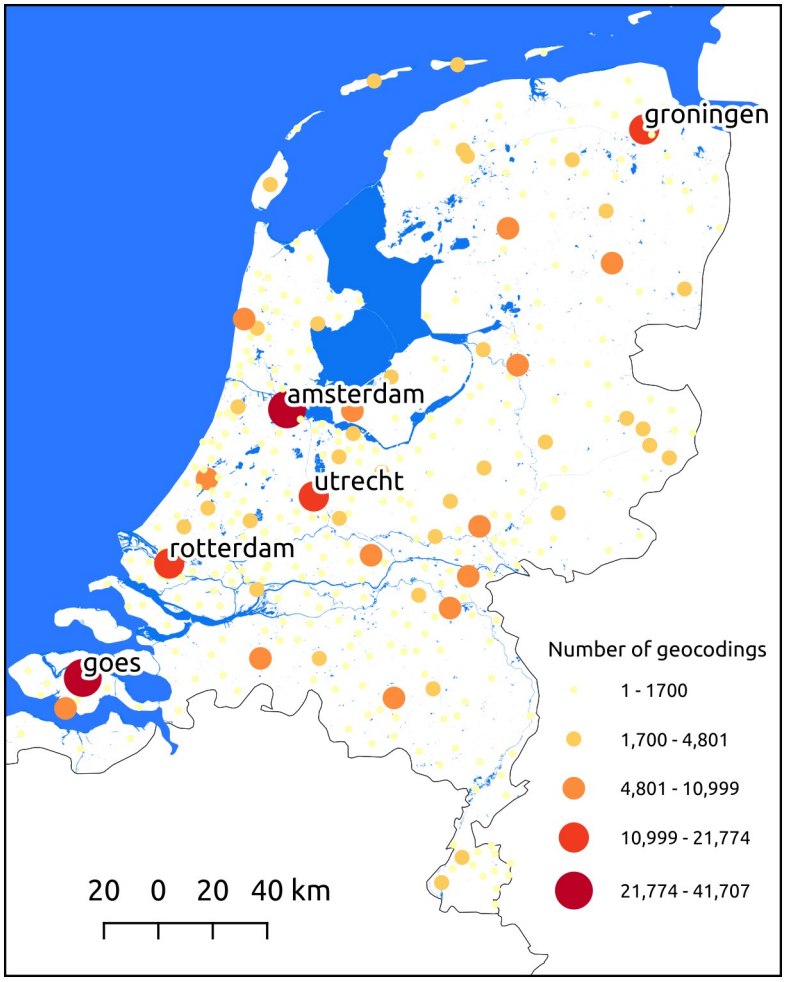

Figure 12: Geocoded records by township and provinces

The geocoding was applied with only township and provinces administrate places, with 450 towns and provinces in total. This approach located 478,692 records within the Tweets context. This unigram approach has some complications, one of such problems can be distinguished in southwestern of the Netherlands. The Figure 12 represent all the tweets that were geolocated by the unigram geocoder, because the language detector is not applied in the prototype, some of the terms such as 'goes' matched with cities like Goes. Figure 12 display all the geocodings that the prototype processed. The places that display names are cities with more than 18,000 geocoded records. For instance, the city of Goes contains 41,707 records, but most of the geocodings of this city are errors. After a review, almost all the records of this city are related with the English word 'goes ' instead of the city Goes, this kind of problem is depicted in Figure 12. This specific error can be solved once the language detection is implemented on the prototype of the complete data set.

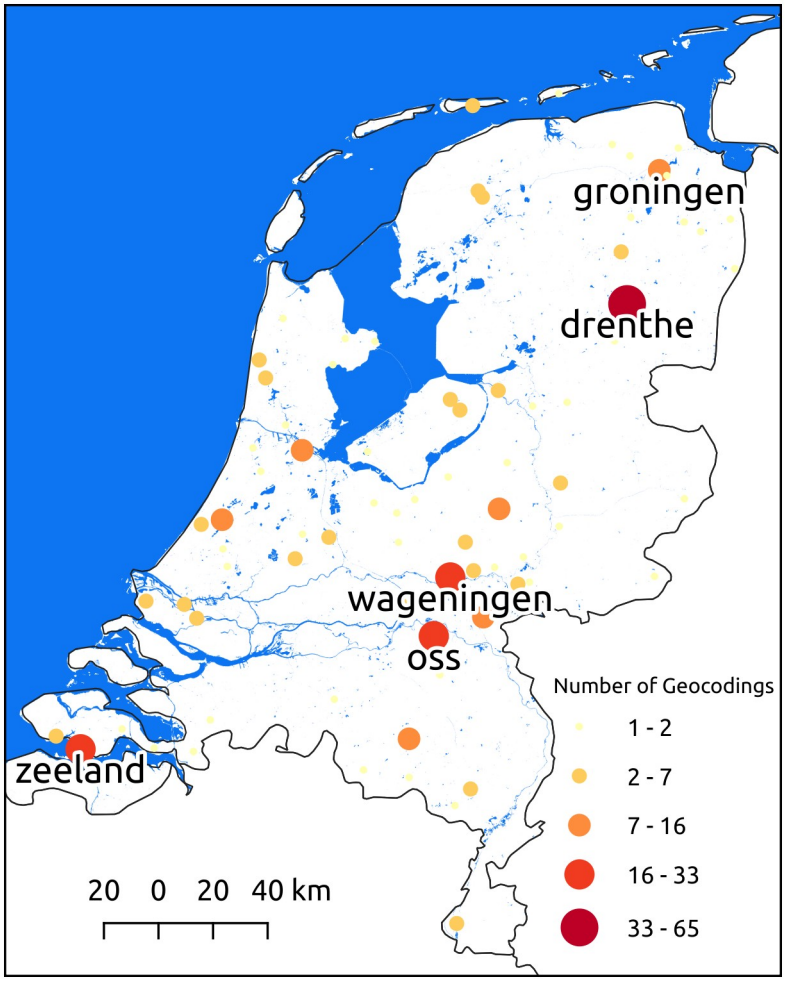

Figure 13: Geocoded records of term 'tekenbeten' by township and provinces

Figure 13 displays the geocoded records that contained the term 'tekenbeten'. Compared to Figure 12 records are more scarce, and they concentrate on different cities. There are only 380 records, and with manual inspection, most of them contain preventive advice, and news related to tick bites. In this inspection only one tweet resulted with information associated with a tick bite event. Nevertheless, some of the geocoded tweets showed information with upcoming tick risks in some specific cities.

The complete data set (172.1 Gb) processing in EMR with Spark last 37 minutes, the architecture was composed of 10 nodes with 16 cores and 64 Gb per node. Approximately the costs of processing the data to of obtain the co-occurrence matrix were around $1.48 USD with ten m4.4xlarge instances.

6. DISCUSSION

This chapter focus on the workflow scenarios, the scenarios relation with the outcomes, and the limitations in the workflow design and in the prototype.

One purpose of the workflow was to test the selected tasks and sub-tasks to process social media. The general workflow was used to select the architecture and the implemented workflow that applies general concepts of social media analysis and processing. However, for an advanced user, the general workflow might be uninteresting and these decisions might be irrelevant. The advanced user might be interested in the implemented workflow which is more specific and can complement another workflows. Nevertheless, other users will find this work informative, and with some modifications, such as paths and libraries on the prototype, the work can be easily reproduced with a raw Twitter data set. The results chapter demonstrated that the main tasks could be implemented in both scenarios and return basic results such as frequency of words, co-occurrence matrix, and geovisualization. The challenge remains in the selection of the sub-tasks and methods to improve the outcomes for further analysis.

The sub-tasks proposed on this study are considered a basic analysis in social media terms. There are techniques that are more sophisticated and additional methods that can be incorporated to modify the prototype, and from here, a question is generated; are these sub-tasks enough to answer a research question? The results showed that there is no sufficient evidence to link the research question with the data set. The application of prototype combined with the noisy records did not provide enough information to relate the data set with a tick bite events. Furthermore, when filtering the records with only the term 'tekenbeten' the records did not showed a significant amount of records with tick bite events. However, some of the sub-tasks were applied with simple techniques; an improvement on these techniques may generate a positive answer for the study case.

The sub-task removing of the prototype can be improved depending on the situation and the skills of each user. There exist more sophisticated approaches that go deeper on locating and remove specific users considering bots (Wetstone et al., 2017). Certainly, these users influence the data set, a different technique may deliver results that are more accurate. In relation with the NLP sub-tasks, the language tools used on the prototype provide percentages of accuracy in each classification. The accuracy can be improved by capping these percentages and only take the sentences or words with a very high percentage of accuracy. This may have one effect, reduce the records with the risk of losing valuable information but assuring the language consistency. The language sub-task in the cloud was not applied, due to the complexity of installing and training the language detector on pyspark within EMR. It is desirable to apply this sub-task inside the EMR cluster and categorize each tweet, by applying this procedure the geocoding will improve by eliminating such records with the term 'goes' or similar cases. Due the language detection was not

applied in the complete scenario, the results between scenarios are not coherent. By adding this sub-task both scenarios results could be comparable.

The lemmatizer in dutch was not added on the prototype this open the possibility of improvement, the only tool available to lemmatize Dutch words is the frog tool[16] which requires technical knowledge in libraries manual installation and dependencies. In the local machine sample scenario, all the suggested sub-tasks were applied except for the topic modeling and the lemmatizer. Mentioned before the unigram geocoder is a limited option to provide spatial information, for sure the results were affected due to this limited geocoding implementation, this opens the possibility to improve the geocoding with methods that reported more accuracy than the unigrams (Melo & Martins, 2017).

On the web, there exist sites dedicated to monitor tick bites and Lyme disease and contain reports provided by users. One example is tekenradar, which collects information related to Lyme disease and tick bites direct from the users. Comparing the results of the geocoded 'tekenbeten' of the last four years and the reports of the last two weeks of the tekenradar only one place match, Drenthe (Figure 14). Further exploration is required by comparing similar dates of both maps and revisiting the text context of the 'tekenbeten' Tweets. This comparison shows a weak relation between real reports and social media information.

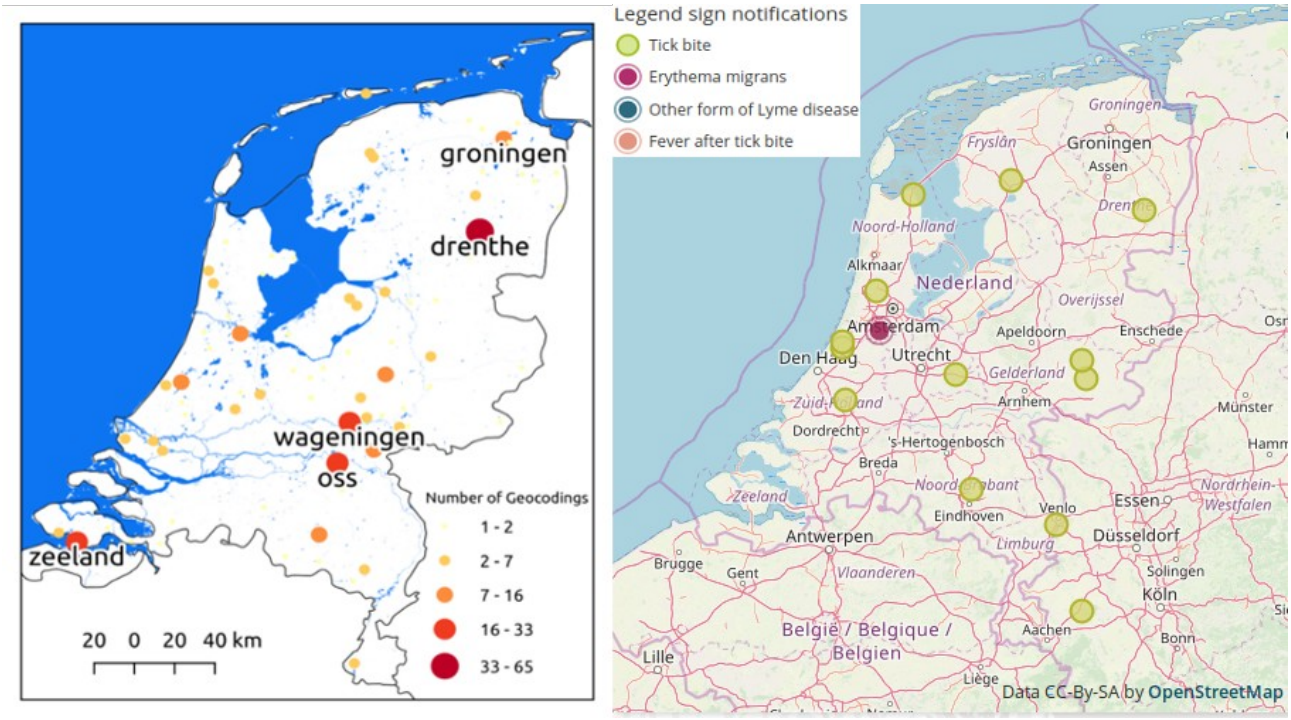

Figure 14: Comparing the results of 'tekenbeten' map and Tekenradar

The results showed that the data set contains noisy words that can be combined with another word that represents well the research question. For example, combinations such as 'camping' with 'teek' are a possible way to filter the complete data set., Perhaps this analysis should be added to the workflow as an

16 https://languagemachines.github.io/frog/

additional sub-task. Each Tweet contains more than 50 fields with different types of information; this study only used information within the text, users, and geolocation. Undoubtedly, there is more valuable information in each tweet such as the date, the popularity of the tweet, or replies. By adding such information the analysis may be enriched.

One of the objectives of this research was to include geoprocessing as part of the workflow. In a way some geoprocessing has been incorporated with the geocoding and the analysis of the geotagged data in the cloud scenario. However, frameworks such as Geospark or Spatial Hadoop were not incorporated. These frameworks would allow the implementation of powerful geoprocessing tools such as spatial filtering or join queries(Hagedorn & Ilmenau, 2017).

Technical materials such as the code and the step by step guide ensure the reproducibility of this study. The code and a step by step guide, are published on the web[17]in a GitHub repository. Because of the restricted access to the database, the reproducibility can be achieved to only those who can access that database. With the use of the step by step guide, the prototype code can be adjusted in python SQL and CLI. At this point, the usage of the prototype can be reproduced but with the limitations mentioned above.

The prototype is not completely automated, some parameters of the code need manual inputs, the programming of the prototype provides the processing and analysis but there might be more efficient forms to code this prototype. Nevertheless, more powerful libraries[18] that require training are available, to modify the prototype with this type of libraries might useful for further analyses.

In this research the usage of the cloud infrastructure as a service was cheap and fast. The architecture on the cloud was used rationally, with gradual testing to save resources. However, the requirements to modify the prototype in the cloud are complex because any modification requires knowledge of the cloud environment and Spark configuration. Thus, cloud scenario may not be simple to replicate as any amendment to the code or a different format of the data set may result in several modifications on the EMR and Spark configurations and in the prototype code.

EMR offers the possibility to install Spark or Hadopp or both. For this research the Spark framework was selected based on the possibility to work with both frameworks in the same environments. While Hadoop is restricted by the use of HDFS and applies the Map/Reduce algorithm, Spark allows to process the information in-memory, also applies the Map/Reduce model and works with a HDFS file system. These allows flexibility to the system and to a user to use both. The main difference is based on the Map/Reduce model which is strictly disk-based while Spark uses memory trough RDD's and DAG to process the information. Spark was designed to enhance Hadoop, not to replace it (Verma et al., 2016).

The EMR adds a level of complexity to work due to the additional settings and steps for the configuration. In this case, the benefits from the AWS architecture and EMR are shown in the scalability.

17 https://github.com/mrfiesta/ProcessAnalyze_Tweets_Local_and_EMRSpark
18 https://fasttext.cc/ or https://nlp.johnsnowlabs.com/quickstart.html

EMR allowed increasing or decreasing a cluster size with a few lines in CLI. This is an advantage with respect to the selection or architecture that the user can pick according to his necessities and save time and resources. Nevertheless, an architecture outside EMR may provide more control to the resources management and be cheaper in the long term, but it requires a deep knowledge on clustering, parallel computing, Spark configurations, and functional programming. Also, it is valuable to mention that this type of service is provided by other companies that may have better and friendlier services.

As mention previously the reason for selecting AWS was an administrative reason, but there exist other options similar to EMR that can be explored such as Dataproc of Google, Microsoft Azure HDInsight or Databricks. The workflow may be implemented with another platform service provider. As an example of the possible advantages with different providers:

1. the configuration may be simpler in Databricks,
2. Dataproc seems to be relatively cheaper per instance,
3. HDInsight has a connection with excel as its front-end analysis tool.

The selection should be based on the ease of use, performance, cost and compatibility , but mainly on the requirements of each study (Tereshko, 2017).

7. CONCLUSIONS

Following the conceptual workflows, a prototype was developed, and tested with a data set. The workflow design tried to cover a general perspective of processing and geoprocessing social media that was based on previous workflows and provided three angles. The first angle is the one to pick an architecture decision with the usage of the general workflow. The second is the exploratory option for a newcomer. And the third, it can be used by a person who requires more resources to analyze his research question. The results provide insights for further research related with this specific study case. In addition, it is proved that the cloud processed the information with a good timing and few resources. However, the workflow needs to be tested with more research question and data sets, and modifications might be needed depending on each case study. Finally, the prototype code and the step by step guide were designed so that anyone with basic notions of programming in python, SQL, and AWS CLI can run and replicate the prototype.

7.1. Research Questions Answered

The first objective was to specify principal tasks needed to transform social media into geoinformation and incorporate them in a workflow. The specific answers to the related research questions are as follows:

1. *Which tasks and techniques are necessary to incorporate geosocial media, geoprocessing and a cloud environment in the same workflow?*

   The selection of the tasks and techniques are based on a literature review. The tasks and techniques are related to the environment of work (local or cloud). Defined tasks for this workflow are data management, pre-processing, NLP, and analysis. These tasks were designed to handle the social media information; the geoprocessing is incorporated in the sub-tasks of geocoding and visualization.

2. *How to operationalize the workflow integrating the required tasks and techniques?*

   The workflow was operationalized by implementing the workflow on a prototype.. The prototype is not fully automated and requires some changes to be reproduced, but the code is public in GitHub and anyone can adjust it and do his own needs.

This objective analyzes the relationship between the architecture characteristics and the social media data, from this objective the scenarios were defined. The following answers are related with scenarios, technologies, and the pros and cons between scenarios.

1. *Which scenarios can be defined based on the stored data, geoprocessing tasks, and system infrastructure?*

   Based on the data set, architecture and spatial data two scenarios were defined considering the user resources. One scenario involves a local machine and targeting users with little or few experiences in processing social media. The second scenario is on a cloud environment and focuses on advanced users that might want to explore the cloud environment with a big data framework. The geoprocessing sub-tasks were adapted on both scenarios based on the characteristics of the geotagged information on social media.

2. *Which and why different type of technologies are required for each scenario?*

   The technologies are linked with the architecture and the type of processing. On the prototype, the technologies that were used in the local machine scenario are: a) DBMS (PostgreSQL) for managing the data set, b)python programming for processing and analyses, with c) libraries such as Pandas designed to handle a data frame processing, nltk used to process the text of each tweet, and folium to visualize the spatial information. For the complete data set scenario the technologies that were adopted are: cloud computing (AWS), parallel computing, and in-memory processing (Spark). In the Pyspark framework the technologies were to create a context in a local machine connected to a cluster (SparkContext), or to arrange the data set as a framework with SQL functionality (SQLContext), and to add NLP tasks into the framework (ml.feature).

3. *What are the advantages and disadvantages between the selected scenarios?*

   The advantages of the sampled scenario are: 1) the availability of spatial information, 2) the easiness to implement it on a local machine with only basic programming knowledge, 3) the opportunity to explore the data set with limited resources. The main disadvantage is that the sample do not represent the complete data set. The advantages of the complete data set scenario are: 1) the aggregated value of using the complete data set on the analysis, 2) the possibility to implement a framework such as Spark and process the scenario data set in a few minutes, 3) the scalability that this type of framework provides. The disadvantages are the following: 1) there is a high chance that the scenario requires additional resources, 2) specialized technologies are required to process and manage massive amounts of data.

This objective is related with the implementation of the workflow in a prototype with a study case data set. The answer is provided below:

1. *Which type of limitations from the case study data set and the proposed scenarios will affect the prototype?*

   The first limitation for the prototype is the noisy records collected, is essential to filter this records to continue the analysis. Another limitation is to process the complete data set, this require extra resources, depending on the processing time and the size of the data set. In the cloud, to add extra libraries in the EMR cloud is a complex task. The prototype does not include the lemmatizer, and LDA on both scenarios. These sub-tasks require training and extra libraries in the EMR cluster. Specifically, the extra installation on the cluster becomes a technical limitation.

The following contains three answers to questions on the reproducibility and performance of the prototype.

1. *How do the characteristics of input data and scenarios affect the reproducibility of the workflow and the re-usability of the prototype?*

   The input data can affect the reproducibility of the workflow when the data set does not contain any geotagged records or when the data contain errors from the original data set or any other source. The workflow does not have a sub-task that considers how to clean errors. The prototype is affected also by errors, the data set size, the format of the data (csv, JSON, txt), the modifications on the data set variables names (column names), the updates on EMR programming languages (e.g. Python 4), and the quality of the step by step guide.

2. *Which techniques or benchmarks are the most feasible to evaluate the performance of the prototype?*

   Three benchmarks can be pointed from this study that are related to quality, quantity and time. In the quality field, the geotagged scenario contains a good quality of spatial information, but it lacks quantity. The complete scenario contains all the records and the co-occurrence matrix seems to have a good quality of data, but the spatial information is not accurate. Regarding time, both scenarios fulfill their tasks and sub-tasks in an acceptable time. However, with increased the data, the local machine scenario may fail while the EMR cloud environment is scalable and the time does not become a problem.

3. *How the results from the social media spatial analysis can be used for further research?*

The results from this research provide an insight to the complete data set and should be used to filter and clean the data set from noisy words. Once filtered, it is recommended to rerun the frequency and co-occurrence analysis and compare the results. Finally, a following task should be to apply the topic modeling analysis and check a possible link with the research question.

7.2. Further Work

The workflow needs to be tested by a newcomer and advanced users. They will be the judges on the usability of the workflow and most importantly the missing tasks and sub-tasks. Also, adding an implementation of how to select a cloud provider will increase the usefulness of the workflow.

Some sub-tasks of the prototype in the local machine scenario need to be improved. In the cloud, the language detection may be implemented within the cluster. The lemmatizer and topic modeling were not performed in both scenarios. Further work in the EMR is to install and run external Python libraries into the cluster and incorporate them along with Pyspark.

Based on the results, removing records may clean the data set from noisy words. For example, co-occurrence matrices may filter the data set with specific combinations of non-relevant words such as ‘camping’ and ‘spelen’. The combination of certain words such as ‘lyme’ and ‘tekenbeten’ may result in a useful combination. This type of co-occurrence filter may lead to successfully track tick bite events.

A part of the prototype was focused on geocoding with unigrams. The complete data set scenario may be improved with a more consistent approach that yields better results than the unigrams such as approaches based on K-trees, logistic regressions, or deep neural networks. Once implemented, the geotagged data set can be used to locate specific events in space and time. An additional stage may be focused on the implementation of isolated clustering events by space and time, one option is to implement a Density-Based Spatial Clustering (DBSCAN) which can be controlled by time and space simulating an event bubble. Isolate and research spatial activities such as the geotagged records from “National Park De Loonse en Drunense Duinen” may give clues with respect of outdoor activities and tick bite events, hence it is recommended to analyze isolated events.

LIST OF REFERENCES

Alarabi, L., Mokbel, M. F., & Musleh, M. (2018). ST-Hadoop: a MapReduce framework for spatio-temporal data. *GeoInformatica*, *22*(4), 785–813. https://doi.org/10.1007/s10707-018-0325-6

Al-rfou, R., & Perozzi, B. (2013). Polyglot-distrubuted word representation for multilingual NLP (pp. 183–192).

Alarabi, L., Mokbel, M. F., & Musleh, M. (2018). ST-Hadoop: a MapReduce framework for spatio-temporal data. *GeoInformatica*, *22*(4), 785–813. https://doi.org/10.1007/s10707-018-0325-6

Altintas, I., Block, J., De Callafon, R., Crawl, D., Cowart, C., Gupta, A., ... Smarr, L. (2015). Towards an integrated cyberinfrastructure for scalable data-driven monitoring, dynamic prediction and resilience of wildfires. *Procedia Computer Science*, *51*(1), 1633–1642. https://doi.org/10.1016/j.procs.2015.05.296

Amazon. (2018). Cluster Configuration Guidelines and Best Practices - Amazon EMR. Retrieved December 26, 2018, from https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-instances-guidelines.html

Armstrong, M. P., Wang, S., & Zhang, Z. (2018). The Internet of Things and fast data streams: prospects for geospatial data science in emerging information ecosystems. *Cartography and Geographic Information Science*, *46*(1), 39–56. https://doi.org/10.1080/15230406.2018.1503973

Batrinca, B., & Treleaven, P. C. (2014). Social media analytics: a survey of techniques, tools and platforms. *AI and Society*, *30*(1), 89–116. https://doi.org/10.1007/s00146-014-0549-4

Blanford, J. I., Huang, Z., Savelyev, A., & MacEachren, A. M. (2015). Geo-located tweets. Enhancing mobility maps and capturing cross-border movement. *PLoS ONE*, *10*(6), 1–16. https://doi.org/10.1371/journal.pone.0129202

Blei, D., Jordan, M., & Ng, A. Y. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

Chang, Z. E., & Li, S. (2013). Geo-social model: A conceptual framework for real-time geocollaboration. *Transactions in GIS*, *17*(2), 182–205. https://doi.org/10.1111/j.1467-9671.2012.01352.x

Chris Holmes. (2018). Cloud Native Geoprocessing Part 1: The Basics – Planet Stories – Medium. Retrieved July 30, 2018, from https://medium.com/planet-stories/cloud-native-geoprocessing-part-1-the-basics-9670280772c8

Croitoru, A., Crooks, A., Radzikowski, J., & Stefanidis, A. (2013). Geosocial gauge: A system prototype for knowledge discovery from social media. *International Journal of Geographical Information Science*, *27*(12), 2483–2508. https://doi.org/10.1080/13658816.2013.825724

Edd, M. O., & Rn, S. Y. (2015). American Journal of Infection Control What can we learn about the Ebola outbreak from tweets? *American Journal of Infection Control*, *43*(6), 563–571. https://doi.org/10.1016/j.ajic.2015.02.023

Eldawy, A., & Mokbel, M. F. (2015). SpatialHadoop: A MapReduce framework for spatial data. *Proceedings - International Conference on Data Engineering*, *2015–May*, 1352–1363. https://doi.org/10.1109/ICDE.2015.7113382

ESRI. (n.d.). What is geoprocessing?—ArcGIS Pro | ArcGIS Desktop. Retrieved September 20, 2018, from http://pro.arcgis.com/en/pro-app/help/analysis/geoprocessing/basics/what-is-geoprocessing-.htm

Frias-Martinez, V., Soto, V., Hohwald, H., & Frias-Martinez, E. (2012). Characterizing urban landscapes using geolocated tweets. *Proceedings - 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012*, 239–248. https://doi.org/10.1109/SocialCom-PASSAT.2012.19

Garg, S. K., Versteeg, S., & Buyya, R. (2013). A framework for ranking of cloud computing services. *Future Generation Computer Systems*, *29*(4), 1012–1023. https://doi.org/10.1016/j.future.2012.06.006

Google. (2019). Get Started | Geocoding API | Google Developers. Retrieved February 5, 2019, from https://developers.google.com/maps/documentation/geocoding/start

Goonetilleke, O., Sellis, T., Zhang, X., & Sathe, S. (2014). Twitter analytics. *ACM SIGKDD Explorations Newsletter*, *16*(1), 11–20. https://doi.org/10.1145/2674026.2674029

Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S., & Ball, J. (2010). Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1925), 3875–3889. https://doi.org/10.1098/rsta.2010.0149

Gurunath, R., & Kumar, R. A. (2015). Saas Explosion Leading To a New Phase of a Learning Management System. *International Journal of Current Research and Review*, *7*(22), 62–66.

Hagedorn, S., & Ilmenau, T. U. (2017). Big Spatial Data Processing Frameworks: Feature and Performance Evaluation. *Proceedings of the 20th International Conference on Extending Database Technology (EDBT)*, 490–493. https://doi.org/10.5441/002/edbt.2017.52

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, *41*(3), 260–271. https://doi.org/10.1080/15230406.2014.890072

Hettne, K., Wolstencroft, K., Belhajjame, K., Goble, C., Mina, E., Dharuri, H., ... Roos, M. (2012). Best practices for workflow design: How to prevent workflow decay. *CEUR Workshop Proceedings*, *952*.

Houde, S., & Hill, C. (1997). *Chapter 16 What do prototypes prototype? Handbook of Human Computer Interaction*. https://doi.org/10.1016/B978-044481862-1.50082-0

INAP. (2016). Cloud 101: The Differences Between Four Types of Cloud Environments - INAP. Retrieved February 3, 2019, from https://www.inap.com/blog/cloud-101-environments/

Integrify. (2017). Workflow Management Benefits. Retrieved December 31, 2018, from https://www.integrify.com/workflow-management-benefits/

Jurafsky, D., & Martin, J. (2007). *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*. https://doi.org/10.1162/089120100750105975

Kirschnick, J., Alcaraz Calero, J. M., Wilcock, L., & Edwards, N. (2010). Toward an architecture for the automated provisioning of cloud services. *IEEE Communications Magazine*, *48*(12), 124–131. https://doi.org/10.1109/MCOM.2010.5673082

Lansley, G., & Longley, P. (2016). The geography of Twitter topics in London. *Computers, Environment and Urban Systems*, *58*, 85–96. https://doi.org/10.1016/j.compenvurbsys.2016.04.002

Larkin Andrew. (2018). Disadvantages of Cloud Computing - Cloud Academy Blog. Retrieved September 21, 2018, from https://cloudacademy.com/blog/disadvantages-of-cloud-computing/

Li, S., Dragicevic, S., Antón, F., Sester, M., Winter, S., Coltekin, A., ... Cheng, T. (2015). ISPRS Journal of Photogrammetry and Remote Sensing Geospatial big data handling theory and methods : A review and research challenges. *Isprs Journal of Photogrammetry and Remote Sensing*, *115*, 119–133. https://doi.org/10.1016/j.isprsjprs.2015.10.012

Li, Z., Yang, C., Liu, K., Hu, F., & Jin, B. (2016). Automatic Scaling Hadoop in the Cloud for Efficient Process of Big Geospatial Data. *ISPRS International Journal of Geo-Information*, *5*(10), 173. https://doi.org/10.3390/ijgi5100173

M. Jimenez, J., R. Diaz, J., Lloret, J., & Romero, O. (2018). MHCP: Multimedia Hybrid Cloud Computing Protocol and Architecture for Mobile Devices. *IEEE Network*, (February), 106–112. https://doi.org/10.1109/MNET.2018.1300246

Maynard, D., Bontcheva, K., & Rout, D. (2012). Challenges in developing opinion mining tools for social media. *LREC 2012 Workshop @NLP Can u Tag #usergeneratedcontent*, 8. Retrieved from http://gate.ac.uk/sale/lrec2012/ugc-workshop/opinion-mining-extended.pdf

Mazhar Rathore, M., Ahmad, A., Paul, A., Hong, W.-H., & Seo, H. (2017). Advanced computing model for geosocial media using big data analytics. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-017-4644-7

McGough, S. F., Brownstein, J. S., Hawkins, J. B., & Santillana, M. (2017). Forecasting Zika Incidence in the 2016 Latin America Outbreak Combining Traditional Disease Surveillance with Search, Social Media, and News Report Data. *PLoS Neglected Tropical Diseases*, *11*(1), 1–15. https://doi.org/10.1371/journal.pntd.0005295

McPhillips, T., Bowers, S., Zinn, D., & Ludäscher, B. (2009). Scientific workflow design for mere mortals. *Future Generation Computer Systems*, *25*(5), 541–551. https://doi.org/10.1016/j.future.2008.06.013

Mell, P., & Grance, T. (2011). The NIST Definition of Cloud Computing Recommendations of the National Institute of Standards and Technology. *Nist Special Publication*, *145*, 7. https://doi.org/10.1136/emj.2010.096966

Melo, F., & Martins, B. (2017). Automated Geocoding of Textual Documents: A Survey of Current Approaches. *Transactions in GIS*, *21*(1), 3–38. https://doi.org/10.1111/tgis.12212

Moreno-Vozmediano, R., Montero, R. S., & Llorente, I. M. (2012). Iaas cloud architecture: From virtualized datacenters to federated cloud infrastructures. *Computer*, *45*(12), 65–72. https://doi.org/10.1109/MC.2012.76

Neves, P. C., Schmerl, B., Camara, J., & Bernardino, J. (2016). Big data in cloud computing: Features and issues. *IoTBD 2016 - Proceedings of the International Conference on Internet of Things and Big Data*.

Ostermann, F. O., García-chapeton, G. A., Kraak, M., & Zurita-milla, R. (2018). *Towards a crowdsourced supervision of the analysis of user-generated geographic content : Engaging citizens in discovering urban places*.

Ostermann, F. O., & Granell, C. (2017). Advancing Science with VGI: Reproducibility and Replicability of Recent Studies using VGI. *Transactions in GIS*, *21*(2), 224–237. https://doi.org/10.1111/tgis.12195

Preotiuc-Pietro, D., Samangooei, S., Cohn, T., Gibbins, N., & Niranjan, M. (2012). Trendminer: an architecture for real time analysis of social media text. *AAAI Technical Report*, 38–42. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4739

Ramos, J. A., Mary, R., Kery, B., Rosenthal, S., & Dey, A. (2017). Sampling Techniques to Improve Big Data Exploration. *IEEE Symposium on Large Data Analysis and Visualization (LDAV)*, 26–35.

Rao, T. R., Mitra, P., Bhatt, R., & Goswami, A. (2018). The big data system, components, tools, and technologies: a survey. *Knowledge and Information Systems*, 1–81. https://doi.org/10.1007/s10115-018-1248-0

Reynold, X. (2014). Apache Spark the fastest open source engine for sorting a petabyte - The Databricks Blog. Retrieved December 11, 2018, from https://databricks.com/blog/2014/10/10/spark-petabyte-sort.html

Roca, S. F., & Cited, R. (2001). System and method for synchronizing, time across a computer cluster. https://doi.org/10.1038/incomms1464

Rountree, D., & Castrillo, I. (2014). *Basics of cloud computing: understanding the fundamentals of cloud computing in Theory and Practice*. Elsevier Syngress.

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, *39*(October 2017), 156–168. https://doi.org/10.1016/j.ijinfomgt.2017.12.002

Suma, S., Mehmood, R., & Albeshri, A. (2018). *Smart Societies, Infrastructure, Technologies and Applications* (Vol. 224). Springer International Publishing. https://doi.org/10.1007/978-3-319-94180-6

Sun, S., Luo, C., & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information Fusion*, *36*, 10–25. https://doi.org/10.1016/j.inffus.2016.10.004

Tereshko, T. (2017). Why Dataproc — Google's managed Hadoop and Spark offering is a game changer. Retrieved February 2, 2019, from https://hackernoon.com/why-dataproc-googles-managed-hadoop-and-spark-offering-is-a-game-changer-9f0ed183fda3

Tozzi, C. (2018). 4 Big Data Infrastructure Pain Points and How to Solve Them - Syncsort Blog. Retrieved February 9, 2019, from https://blog.syncsort.com/2018/11/big-data/4-big-data-infrastructure-points-solve/

Verma, A., Hussain, A., & Jain, N. (2016). Big Data Management Processing with Hadoop MapReduce and Spark Technology: A Comparison. In *Comptes rendus hebdomadaires des seances de l"Academie des sciences. Serie D: Sciences naturelles*. IEEE. https://doi.org/10.1109/CDAN.2016.7570891

Wachowicz, M., Arteaga, M. D., Cha, S., & Bourgeois, Y. (2016). Developing a streaming data processing workflow for querying space–time activities from geotagged tweets. *Computers, Environment and Urban Systems*, *59*, 256–268. https://doi.org/10.1016/j.compenvurbsys.2015.12.001

Wetstone, J. H., Edu, W., & Nayyar, S. R. (2017). I Spot a Bot: Building a binary classifier to detect bots on Twitter. *Cs 229 Final Project Report*, (DECEMBER), 1–6.

Yang, C., Yu, M., Hu, F., Jiang, Y., & Li, Y. (2017). Utilizing Cloud Computing to address big geospatial data challenges. *Computers, Environment and Urban Systems*, *61*, 120–128. https://doi.org/10.1016/j.compenvurbsys.2016.10.010

Yang, J.-A., Tsou, M.-H., Jung, C.-T., Allen, C., Spitzberg, B. H., Gawron, J. M., & Han, S.-Y. (2016). Social media analytics and research testbed (SMART): Exploring spatiotemporal patterns of human dynamics with geo-targeted social media messages. *Big Data & Society*, *3*(1), 205395171665291. https://doi.org/10.1177/2053951716652914

Yu, J., Jinxuan, W., & Mohamed, S. (2015). GeoSpark: A Cluster Computing Framework for Processing Large-Scale Spatial Data. *SIGSPATIAL International Conference on Advances in Geographic Inormation Systems*, (3), 4–7. https://doi.org/10.1145/2820783.2820860

Yue, P., Zhang, C., Zhang, M., Zhai, X., & Jiang, L. (2015). An SDI Approach for Big Data Analytics: The Case on Sensor Web Event Detection and Geoprocessing Workflow. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *8*(10), 4720–4728. https://doi.org/10.1109/JSTARS.2015.2494610

Zhang, M., Bu, X., & Yue, P. (2017). GeoJModelBuilder: an open source geoprocessing workflow tool. *Open Geospatial Data, Software and Standards*, *2*(1), 8. https://doi.org/10.1186/s40965-017-0022-7