Universidade do Minho

Gestão de Grandes Conjuntos de Dados

Ciência de Dados

Customer Support on Twitter Proposal

Authors:	**Number:**
Manuel Monteiro	PG37158
Tiago Alves	A78218
Vitor Peixoto	A79175

31st March 2019

Domain Description

The Customer Support on Twitter includes over 3 million tweets and replies to aid innovation in natural language understanding and conversational models, and for study of modern customer support practices and impact.

This dataset offers a large corpus of modern English (mostly) conversations between consumers and customer support agents on Twitter, and has three important advantages over other conversational text datasets:

- Focused - Consumers contact customer support to have a specific problem solved, and the manifold of problems to be discussed is relatively small, especially compared to unconstrained conversational datasets like the reddit Corpus.
- Natural - Consumers in this dataset come from a much broader segment than those in the Ubuntu Dialogue Corpus and have much more natural and recent use of typed text than the Cornell Movie Dialogs Corpus.
- Succinct - Twitter's brevity causes more natural responses from support agents (rather than scripted), and to-the-point descriptions of problems and solutions. Also, its convenient in allowing for a relatively low message limit size for recurrent nets.

Dataset description

Each row on this dataset is a tweet. Every conversation included has at least one request from a consumer and at least one response from a company.

Variables:

- ***tweet_id*** - A unique ID for the Tweet. Referenced by response_tweet_id and in_response_to_tweet_id.
- ***author_id*** - A unique user ID. The '@s' in the dataset have been replaced with their associated anonymized user ID.
- ***inbound*** - Whether the tweet is "inbound" to a company doing customer support on Twitter. This feature is useful when re-organizing data for training conversational models.
- ***created_at*** - Date and time when the tweet was sent.
- ***text*** - Tweet content. Sensitive information like phone numbers and email addresses are replaced with mask values like __email__.
- ***response_tweet_id*** - IDs of tweets that are responses to this tweet, comma-separated.

- ***in_response_to_tweet_id*** - ID of the tweet this tweet is in response to, if any.

Motivation

A dataset of costumer service based on a social network can answer an enormous amount of questions, specially a dataset of such size and breadth. Therefore, the main motivation for this data analysis is the improvement of customer service. This improvement can be achieved by multiple questions to the dataset:

- Can we predict company responses?
- How quickly do the best companies respond?
- What day and time is it more likely for me to get an answer from a company?
- How does tone affect the customer support conversation?
- Can we help companies identify new problems?

This questions help to achieve a better customer support to the clients, by studying which type of approach to make, but also tells the customers which brand will give them a better support and care in any case of need. Logically, happier customers will buy more products and that translates into more revenue.

Architecture of Proposed Solution

This is a big data set, with over 3 million entries, so it's required the use of some specific technologies so we can have a good and efficient analysis.

Having the dataset ready to be read, the steps coming after that are:

- Data collection
- Data storage
- Data processing
- Data analytics

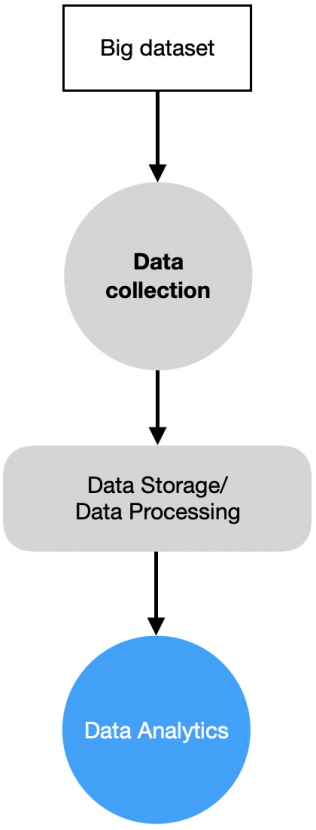

Figure 1: Process architecture

The projection of a possible architecture for the data flow is important to develop reliable and scalable data pipelines. The stack of the architecture involves profound knowledge of every layer of the architecture, so, we will design a more abstract architecture to use as a guideline for the data flow development.

First of all, the cluster planning. Cloud services will probably be our choice since we can move things around, setting up the machine size to fit the dataset needs.

To handle and process the streams of data we more likely will be using Hadoop. It can handle huge volumes and varieties of data in a relatively affordable and flexible way. Hadoop also features a distributed processing framework, MapReduce, to distribute tasks across clusters of nodes, so that large volumes of data can be processed very quickly.