Time Series Analysis of Twitter Data

Matthew Tiger

December 2, 2015

1. Introduction

With the advent of social media, the way in which society communicates has evolved. Popular social media networks document these new forms of communications and as a result, a rich set of data surrounding these interactions emerges. In particular, Twitter remains one of the most popular social media platforms to date. Twitter was first launched in July 2006 as a social networking service designed to allow its users to communicate via short 140-character messages called "tweets". These tweets are sometimes affixed by the sender with a meta-label called a "hashtag", denoted by a string leading with the # symbol, that is meant to categorize the information contained in the message. As of May 2015, Twitter's active user base numbers 302 million users sending these categorized messages every second.

In this report, we will analyze data pertaining to a popular television show collected from Twitter's streaming API over the course of three weeks. This analysis will consist of measuring the number of tweets that contain a certain hashtag sent in a given hour over this timeframe. We will then fit a time series model to these measurements and provide a forecasting model of the next week's projected data.

2. Data

In an effort to help researchers glean insight from tweets, Twitter offers a streaming API that streams all tweets that contain a certain string. We leverage this service to gather data surrounding the popular television show *The Walking Dead* and then measure the number of tweets that occur each hour over a time span.

Using the programming language Python and the library `tweepy`, a popular wrapper for the Twitter streaming API, we collected all tweets containing the hashtag "#thewalkingdead" for three weeks from 2015-11-07 21:00 EST to 2015-11-23 21:00 EST. As this television show airs Sundays at 21:00 EST, of particular importance is the tweets occurring around this time. We therefore restrict our measurements to 24 hours prior to and 24 hours after the television show's airing for the above three week time frame.

These measurements give rise to the time series presented in Figure 1. From this plot it is clear that, due to the weekly episodic nature of the television show, there is a seasonal component to this time series. It also appears that there is a trend component. Thus, in order to fit a stationary time series to the underlying data, we will need to first apply transformations to the data.

The details behind the process of fitting a stationary time series model to this data is handled in Section 3.

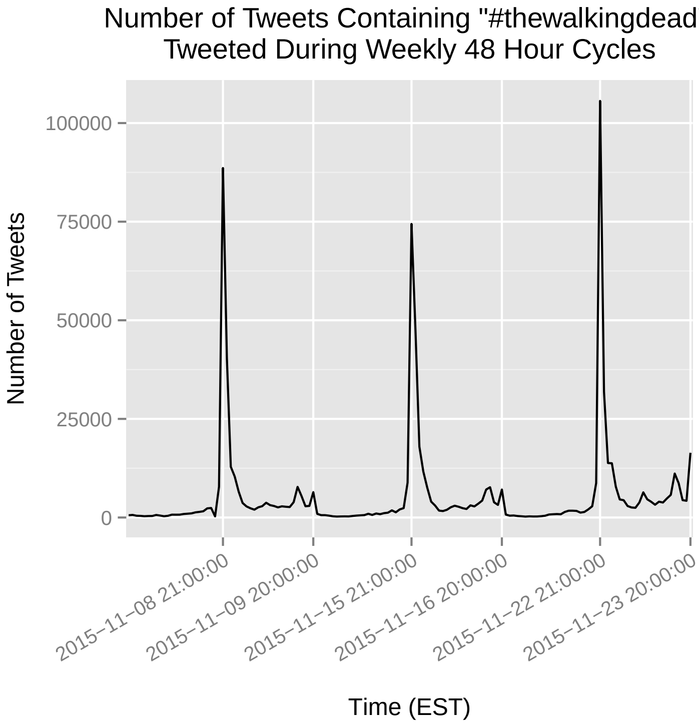

Figure 1: Time series plot of number of tweets containing the hashtag #thewalkingdead over three 48 hour cycles.

3. Model Fitting

4. Forecasting

5. Conclusion