2019 Canadian election tweets
OSEMN methodology
Step 2: Scrub (clean) data
Cleanup plan for Sentiment 140 dataset

October 12, 2019

Abstract

The following document presents the cleanup plan to be performed on Sentiment 140 dataset.

1 Sentiment 140 cleanup plan

The following steps will be performed for cleaning the Sentiment 140 dataset:

1. Parse dates
2. Replace HTML character codes
3. Extract hashtags from tweets
4. Extract user handles from tweets
5. Remove links from tweets
6. Remove duplicates
7. Remove tweeter bots / users with high frequency of tweets