Final Paper Checklist

**Name:** Andrew Larkin

**Title of Paper:** Evaluating Twitter for Nature-Related Public Health Research

**Hypothesis:** Twitter can be used to capture temporal trends in social, exercise, and psychological dimensions of public perceptions and use of urban nature reported on Twitter.

**Data:** 72,007 green space-related tweets collected from January 1st to December 1st, 2018 in the Portland, OR metropolitan area

**Data Collection:** Search-based data collection from Twitter data stream using the Tweepy module in Python 2.7.

**Advanced Analytics Methods:** term frequency–inverse document frequency was used to identify exclusion terms and major social themes. Part of speech tagging and stemming was used to identify characteristics associated with keywords. Sentiment analysis was used to predict weekly sentiment. Moving averages and correlation matrices were used to analyze sensitivity in temporal resolution and identify temporal trends.

**Why will someone want to read your paper?** This paper demonstrates the utility of Twitter for environmental health research, and identifies several limitations and subsequent future directions to improve the described methodology. The target audience consists of social media analysts interested in public health and public health researchers interested in social media.

For your submission ensure you include:

-

This checklist

-

**The paper in Word or LaTex format (so it can be edited later)**

Larkin\_Andrew\_Final\_v2.docx

**All data used for the paper**

TweetDataset.csv – dataset containing 72000 tweets after applying exclusion criteria (data before applying exclusion criteria is more than 300GB).

PartOfSpeechExample.txt – example intermediate result after applying Stanford NLP Part of Speech Tagger

WordExclusion.xlsx – initial words used for screening out unrelated tweets

TF\_IDR\_Results.xlsx – social themes identified by TF\_IDR scoring

WordFrequencyAndSentiment.xls – Results of the word frequency and sentiment analysis, along with corresponding graphs

**Any code/scripts used for the paper**

calcWeekly\_tfidf.py – partition tweets into weekly groups and calculate TF\_IDF values

calcWordAssocations.py – calculate word frequencies based on part of speech syntax restriction

GreenSpaceTwitterSearch.py – download data from Twitter stream and export to csv.

initialScreen.ipynb – apply exclusion criteria to downloaded data

**Any additional comments for the professors:**

In the code used to partition Tweets into week groups, I mistakenly grouped Tweets into 8 day rather than 7 day time intervals. I caught this error in time to update graphs and methodologies in the paper to document this error, but don’t have enough time to rerun the analysis for 7 day rather than an 8 day time interval.