






#HASHTAG: A Twitter Curated Music Recommender



Harrison Tsai, Adam Wang, David Wu

#hashtag is a music recommendation system built from Twitter tweets from the Coachella Music Festival. Social media data is growing at an incredibly fast rate everyday; yet despite this wealth, it is difficult to create insights into people's preferences as they write, tweet, or post about their everyday lives. We believe that extracting data from the raw text of social media can give valuable insights into individual preferences for music service industry.

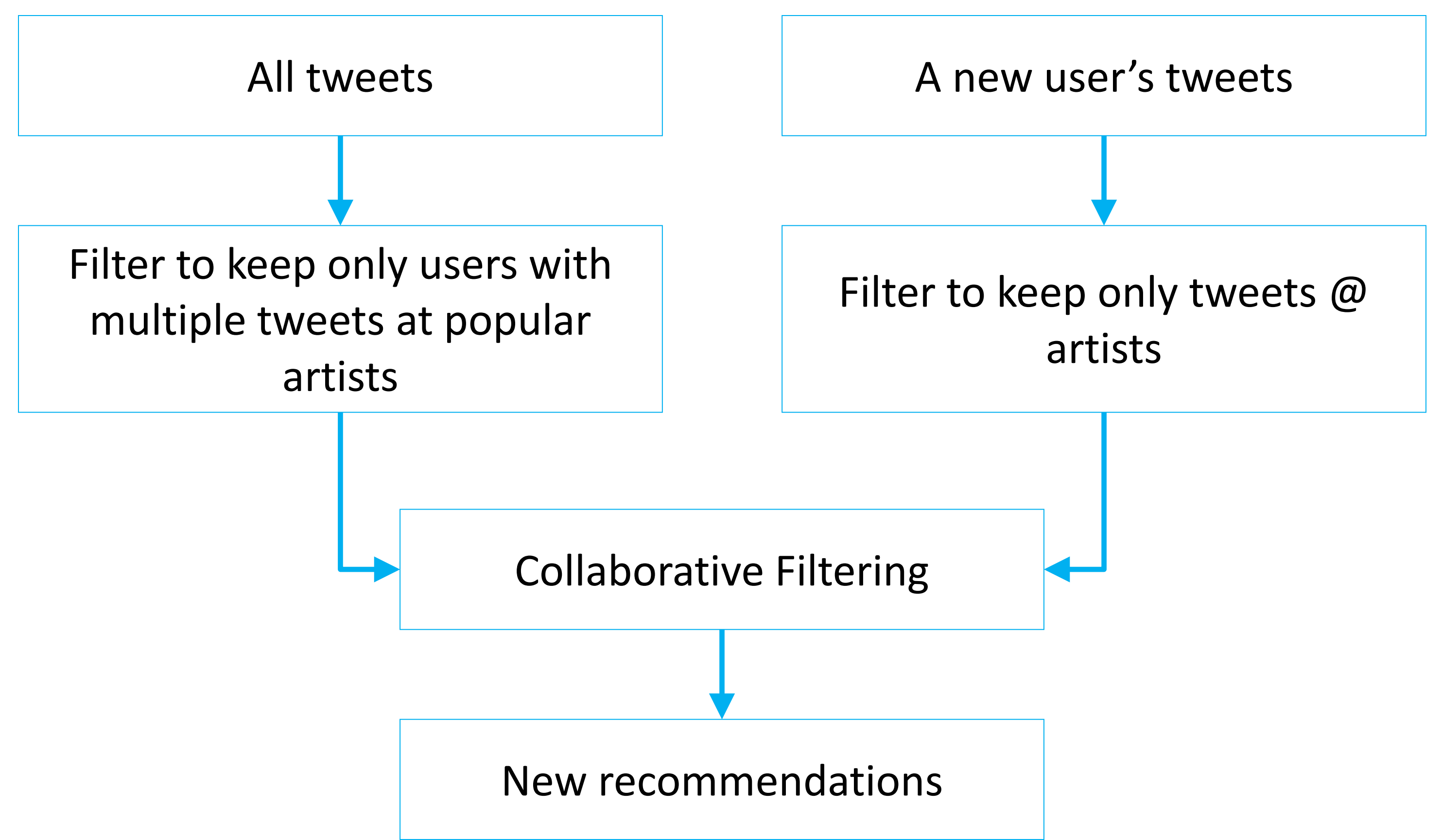
Related Systems

 Spotify  PANDORA <small>internet radio</small>	 #HASHTAG
Binary Rating (Like or Dislike)	Extract Granularities Within Ratings
Active Participation	Implicit Participation
Data Limited to User Inputs	Data Limited to Raw Text

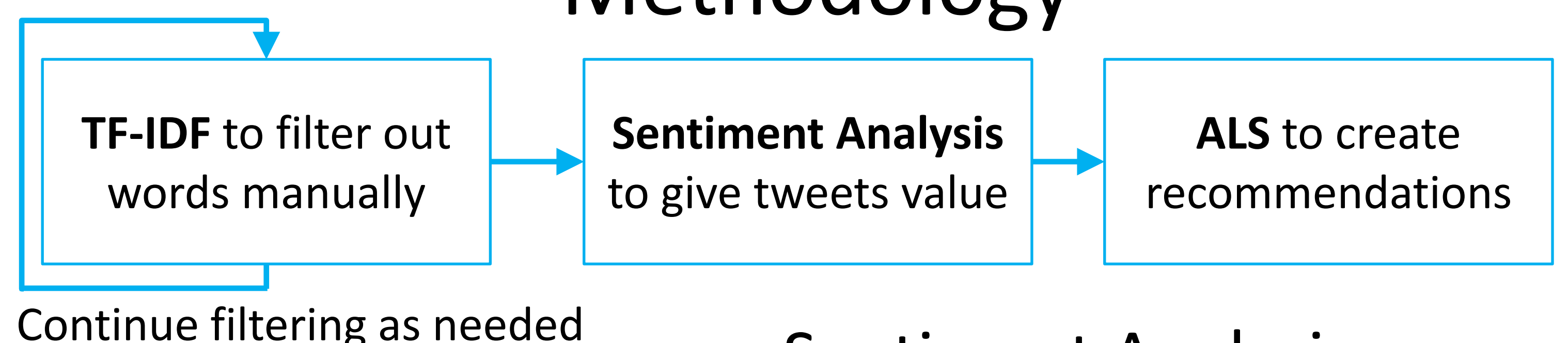
Open Issues: Processing raw text is incredibly difficult. Our bag-of-words model may not capture various aspects of natural language such as the context of words. Future implementations require additional natural language processing research to correct underlying assumptions and improve our collaborative filtering/ALS analysis

Future Work: Yelp restaurant recommendations, Amazon product recommendation, etc.

The Data Product



Methodology



Tools used

- **Pandas:** Data structures and data analysis tools for the Python programming language
- **Spark:** "Fast and general engine for large-scale data processing"
- **D3.js:** JavaScript library for manipulating documents based on data
- **ScraperWiki:** Scraper tool that enables quick collection of data and cleaning
- **Twython:** "Actively maintained, pure Python wrapper for the Twitter API. Supports both normal and streaming Twitter APIs"
- **Scikit-learn:** Machine learning library in python that provides simple and efficient tools for data mining and data analysis
- **Natural Language Toolkit (NLTK):** "A leading platform for building Python programs to work with human language data."
- **Wordle.net:** A tool to generate word clouds
- **Datamaps:** Customizable SVG map visualizations for the web in a single Javascript file using D3.js
- All other analysis was built using the standard Python library and shell scripts (primarily bash)

Problems/Difficulties → Solutions:

- non-ASCII characters → Manual filtering
- Noisy data → Consistent stop-words update
- No training data → Manual labeling
- Frequent false negatives → Ignored
- Memory errors → Created more filters

Datasets

- **Tweets with "#coachella":**
 - **Source:** Twitter API
 - **Size:** 241,026 tweets over the course of the 3 weeks
 - **Schema:** id_str,tweet_url,created_at,text,lang,retweet_count,screen_name,hashtags,query,url,user_mention,media,lat,lng,in_reply_to_screen_name,in_reply_to_status_id
- **Artist names and twitter handles:**
 - **Source:** <http://www.coachella.com/lineup/>
 - **Size:** ~200
 - **Schema:** {name: handle} and {handle: name} in JSON format
 - We manually added surprise guests

Sentiment Analysis

