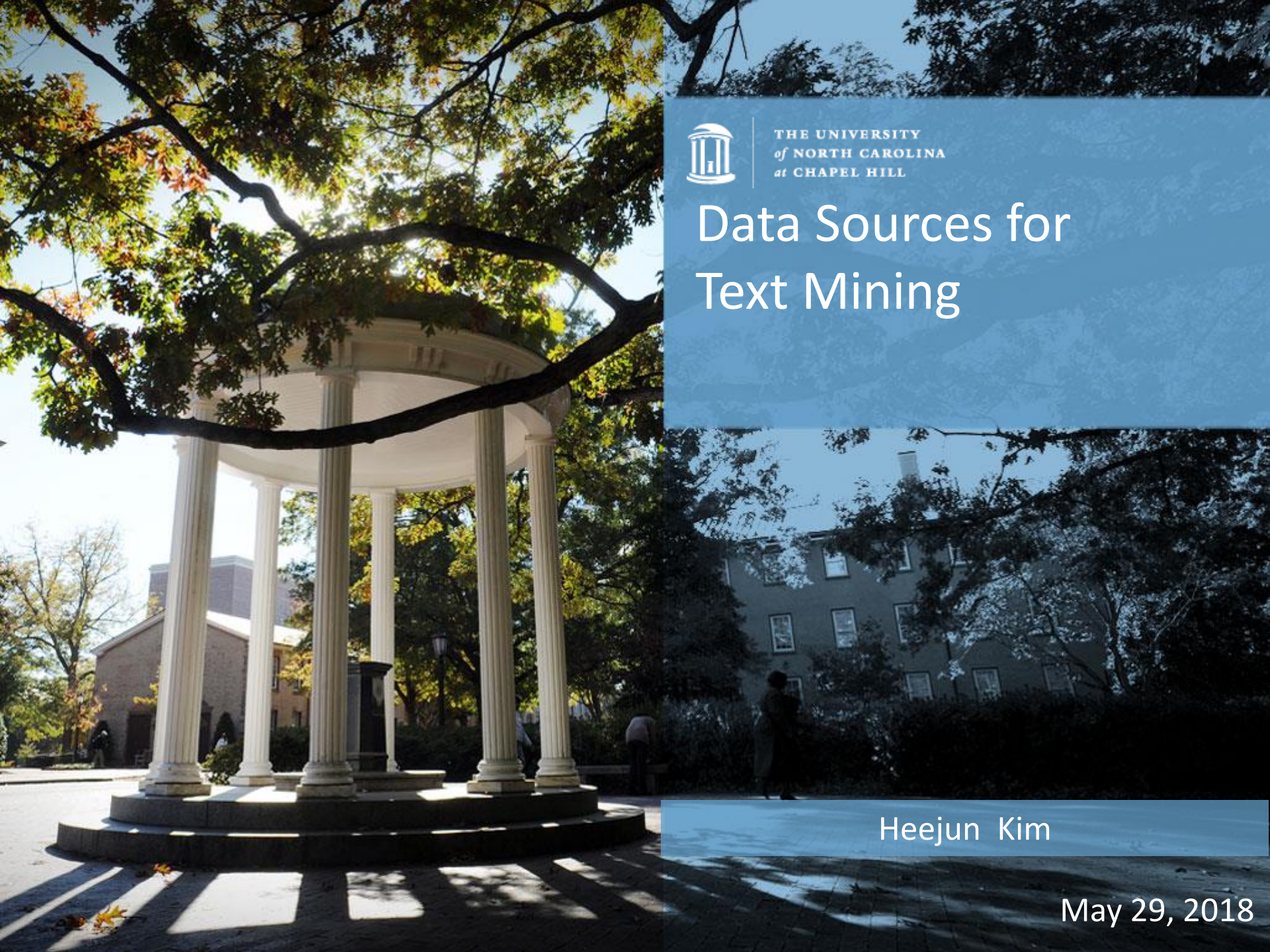

Data Sources for Text Mining

Heejun Kim

May 29, 2018

Data sources for text mining

- Twitter API: https://dev.twitter.com/streaming/overview
- Yelp Challenge Dataset : https://www.yelp.com/dataset_challenge
- Kaggle: https://www.kaggle.com/datasets
- AWS Public data sets: https://aws.amazon.com/datasets/
- UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets.html
- Yahoo Research Data: http://webscope.sandbox.yahoo.com/

Data sources for text mining

- CrowdFlower: https://www.crowdflower.com/data-for-everyone/
- Awesome Public Data: https://github.com/caesar0301/awesome-public-datasets
- Stanford Large Network Dataset Collection: https://snap.stanford.edu/data/
- Cornell movie review data: https://www.cs.cornell.edu/people/pabo/movie-review-data/
- Reddit: https://www.reddit.com/dev/api/

Twitter API

- Provides programmatic access to read and write Twitter data with 3rd party program
- Twitter Search API: uses queries to retrieve recent Tweets (up to 7 days). More strict rate limits (180 requests/queries per 15 minutes)
- Twitter Streaming API: retrieves Tweets from real-time streams. Less strict rate limits (about 1% of Tweets)
- Pro: diverse/recent topic, live voice, and a lot of related studies
- Con: label needed, short content, and a lot of link
- Need to apply for an OAuth access token

Yelp Challenge Dataset

- A huge review data of local businesses (e.g., restaurant and grocery)
- Pro: clean data, rich data on restaurant review (2.7M), availability of social network data (687K users), opportunity for $5,000 cash prize, and some related studies
- Con: domain specific and parsing needed
- Have to sign on Yelp's dataset terms of use

Principle for Selecting Data

- Find a good balance between your goal and reality (e.g., topic and availability of data)
- Exploring pre-processed data can save a lot of time and potentially meet your goal (do not try to create labels by yourself)
- Find data related to a topic you like or are familiar with
- Consider your programming skill (many open source NLP tools are available in Java and Python)

Any questions?