TWINFO

Curtis, Golf, Jesse, Jon, Michael, Sam

Twinfo Info
(www.smileytweets.com)

Real-time twitter trending event tracker
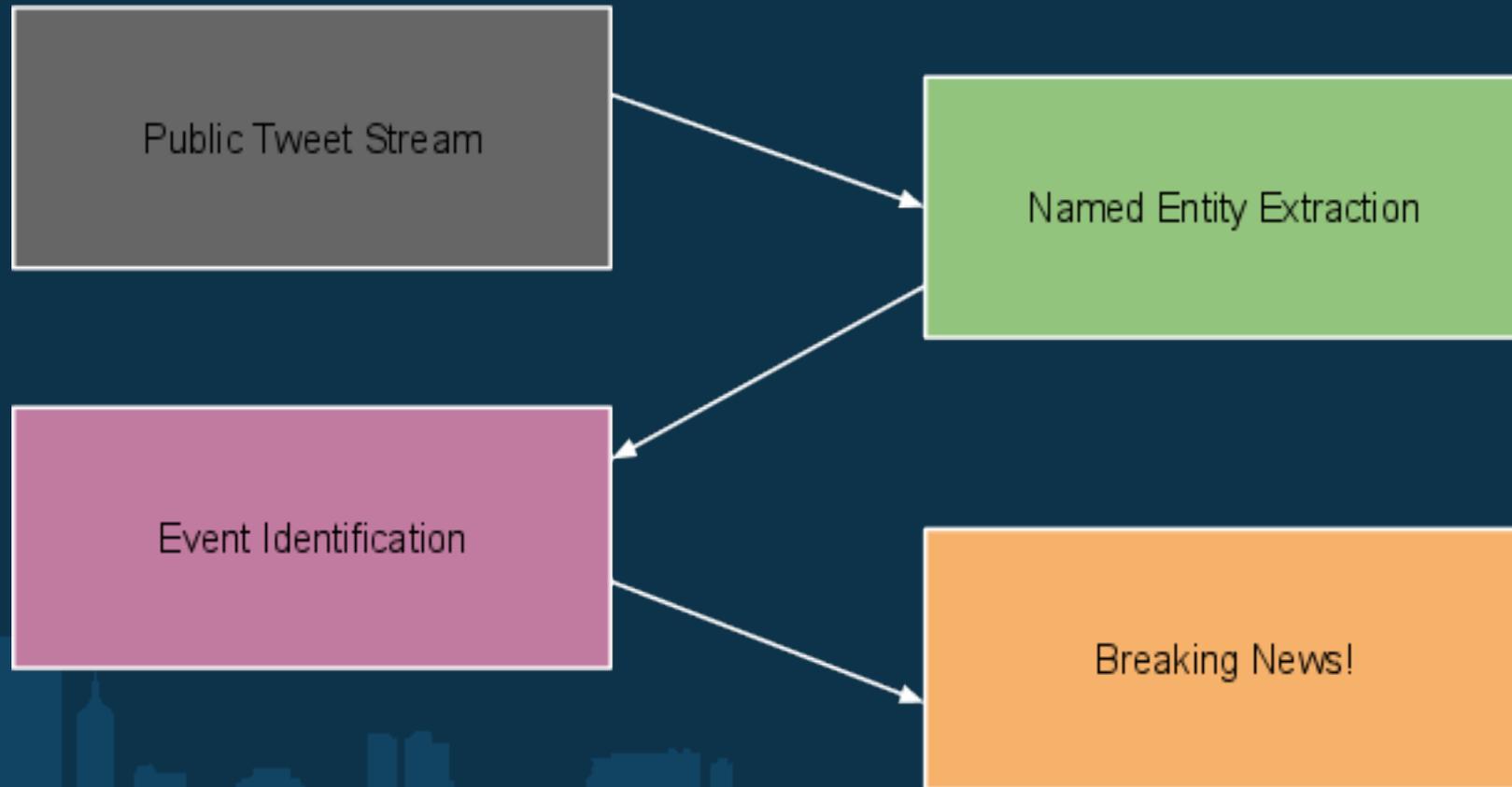
LIVE DEMO

Similar Sites

- Hand-picked trends, or poor quality

- Hand-written context, or none at all

- Pertinent information not displayed to user

twopopular

What the Trend™

TRENDISTIC see trends in twitter

Our Motivation

- Tap into the massive amount of information Twitter generates
- Use named entity recognition to generate rich features for analyzing what is being discussed
  - display content of interest to user
- Show real-world events as they happen
  - unfiltered, unedited, unbiased
  - catered to the ever-expanding twitter audience
  - let the news come to us

Changing Focus

- Initial Goal: Calculate and display sentiment of topics on twitter over time
  - Didn't generate interesting content at a high rate.

- Focus changed to event discovery
  - exciting results for content and event analysis

- Iterative back and front end redesigns
  - data driven decisions

What worked… and didn't

Worked:
- Named entity recognition
- Filter English tweets
- Catching trending topics
- Finding event context
- Categorizing topics

Didn't:
- Categorizing topics
- Adding ~1,000,000 rows to the db per day
- Hard to schedule meetings as a 6 man group

Suprises!

- Runnng 4 AWS machines for
  2 months is costly ~$500!
  - computational, financial limitations

- Roughly 99% of emoticon-bearing tweets are about Justin Bieber
  - alarmingly low, considering his cultural importance

- Event context was consistently precise and reliable
  - real-world events and trends caught quickly

What we learned

- Hard to build systems that do not turn off
  - Multiple processes running 24/7 over multiple machines

- NLP is difficult in the untamed frontiers of Twitter
  - Multiple languages, abbreviations, slang…

- Categorization is tricky in the real world
  - Categories overlap, not evenly distributed

Experiments and Validation

Finding trending info
- Tuning trending metric
- Filtering trends due to spam
  - Retweets

Validation
- Checking if trend happened using google
  - See how quickly trend was caught

Categorization
- Tried training on data from twitter and wikipedia
- Limiting over-zealous categories
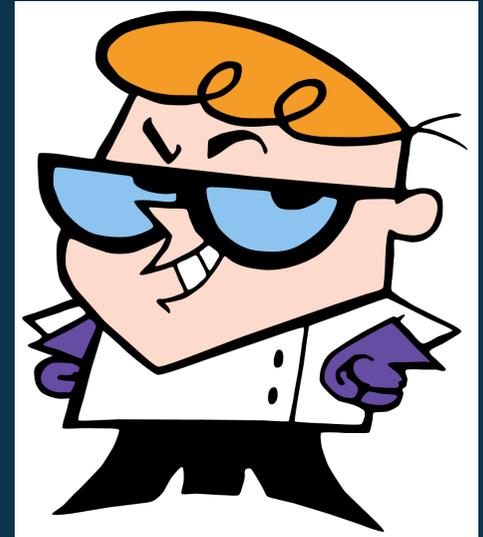
Division of Labor

- Curtis - Front End
- Golf - Front End
- Jesse - Language/Topic categorization
- Jon - Infrastructure, DB interfacing (back and front-end)
- Michael - AWS management, Training data accumulation
- Sam - PM, Named Entity Recognition, Sentiment analysis, Infrastructure, Context identification

Questions? Concerns?