UNIVERSITY OF CAMBRIDGE

Weak Supervised Learning on Ray using Snorkel

British Sentiment Analysis with LSTM Using Noisy Crowd

Sami (sa894) - R244 Large-Scale Data Processing and Optimisation

Snorkel – generating training label

- Generating training labels with weak-supervision
- Apply to unlabelled large quintinites of data
- Quantity over quality

Snorkel using Spark (Snark) [1]

- Quantity = scales. Scale = Large heavy systems. But how heavy?
- Load Data into SparkSQL and Dataframes
- Pre-process the data and convert them to Snorkel format
- Load batches of the snorkel data into Spark clusters
- Distribute the workload using Spark actors to apply labelling functions
- Spark is based on Scala.

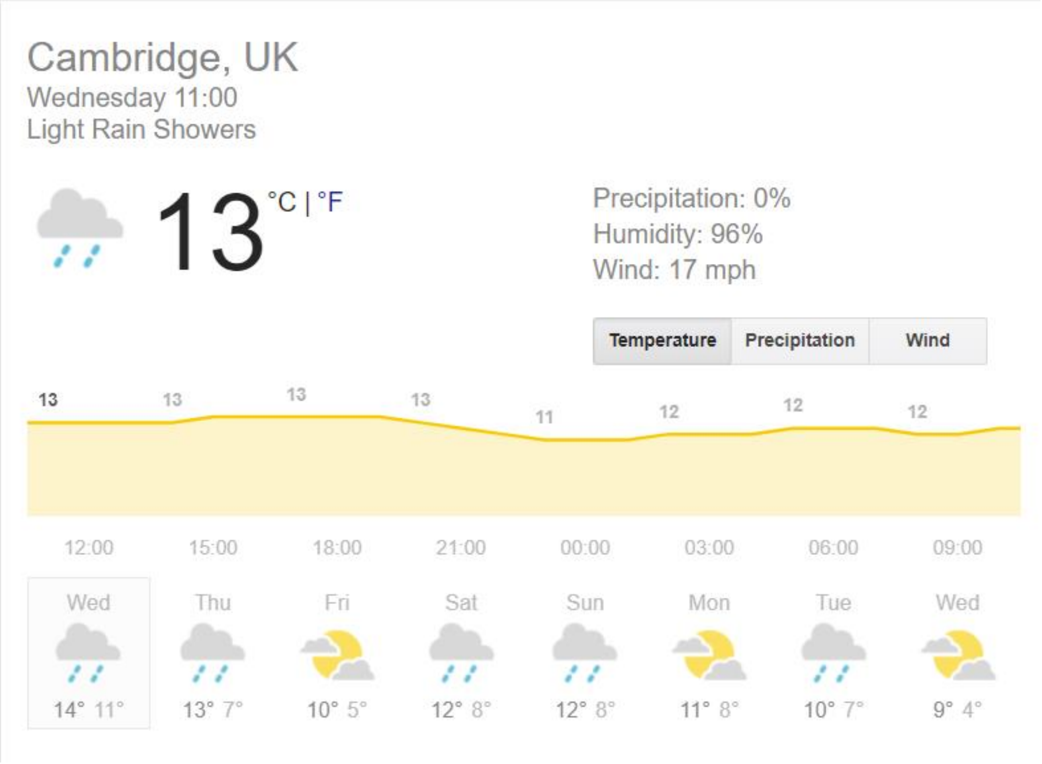
Ray – Making Snorkel spark

- Distributed execution backend optimised for ML tasks
- Lightweight actor model, ideal for iterative process
  - E.g. going through batches of data and apply labelling function.
- Extend the Snorkel codebase to allow an easy to use interface that uses Ray

The desired outcome

- Comparison between Ray and Spark as a backend in applying labelling functions.
- Labelling functions are short lived, Ray should in theory be better, but will it be?
- Open source the interface and making it available to public.

Demo - How’s the weather



Cambridge, UK
Wednesday 11:00
Light Rain Showers

13 °C | °F

Precipitation: 0%
Humidity: 96%
Wind: 17 mph

Temperature | Precipitation | Wind

Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed
14° 11°	13° 7°	10° 5°	12° 8°	12° 8°	11° 8°	10° 7°	9° 4°

The Telegraph

Home Video News World Sport Busines

Politics | Investigations | Obits | Education | Sci

HOME » NEWS » UK NEWS

Talking about weather is 'top British trait'

Talking about the weather, sarcasm and an abilty to queue have been found to be the traits which many people believe defines the British.

Sentiments Analysis - The British Version

- The most common asked question in the UK.
- Tweets sentiment analysis in regards to the weather: Positive, negative, neutral, not related, and unsure.
- Snorkel to resolve conflicts in a noisy crowdsourced dataset, from Crowdflower.
- Then use the denoised labels for an LSTM sentiments analysis
- The idea is to dogfood the interface
- Based on tutorial for learning Snorkel [0]

Progress - Somewhere

- ✓ Read the papers and related work
- ✓ Go through the tutorial for Ray and Snorkel
- ✓ Figure out the best level of abstraction (inspired by Snark)
- ✓ write those slides
- Next
  - Implement RayAnnotationDistribution
  - Run the tests provided in the Snorkel package
  - Evaluate the performance of the test
  - Evaluate the performance of a real application – Sentiment analysis
  - Write down the results