Conference Email Extractor

Parth Ray, Yogitha Sunkara, Andrew Doud, Ben Dahlgren

Overview

———

- Conference Extractor
  - Extractoring conference email information
    - Conference Name
    - Location
    - Relevant Dates
      - Conference
      - Submission
      - Notification
- Website
  - Allows for input text to be parsed
  - React JS Frontend and Flask Python Backend

Event Name Extraction

- Regular expressions to remove junk
- Tokenize then vectorize sentences in an email with tfidf vectorizer
- Get the sentence most likely to contain the event name with a catboost classifier
- Regular expressions to get potential event names from the most likely sentence.
- Simple rule-based method to find most likely event name (based on term document frequencies)

Date Extraction

- Regular expressions to pattern match dates
- Send nearby tokens to a XGBoost classifier to classify each into date type (conference,submission,notification)
- Return the date with the highest probability for each date type
- Use dateparser package to standardize date format as datetime

Location Extraction

spaCy

- Initially using Spacy's NER
  - Tokens were mislabeled ORG instead of GRE
  - Difficult to differentiate city and countries
- Python geography package
  - Cross-checked acronyms with a list of common country names ex. US, UK, UAE

Technical Achievements

———

- Date Extraction
  - Date detection
  - Date type classification
- Location Extraction
  - Regex parsing
  - NER recognition for city and country extractor
- Conference Name Extraction
  - Ensemble of statistical and rule-based models that trains rapidly and is fairly robust

Results

———

- Date Extraction
    - 75% accuracy on all dates
    - Baseline: 16% (assuming all dates found)
- Location
    - Exact match – 37.5%
    - Partial Matches – 55.34%
    - Baseline: 0.09%
- Event Name Extraction
    - 91% accuracy for sentence classification
    - 3% Levenshtein edit distance ratio on average (not a representative performance metric)
    - Baseline: 0.5% (total # event name tokens / total # tokens)

Demo

https://docs.google.com/document/d/1WcugScoAvAnge4kft6GCeDHsrZJ3Q
MU06zZKDPsaqKM/edit?usp=sharing

Thank You
Any Questions?