Entity-Level Sentiment

& Comparison Across Online Communities

McCoy Doherty & Nicolas Ortega

Background :: The Concepts

- We both had some exposure to
  - document sentiment -- from undergrad coursework [BSI]
  - entity extraction -- from SI 650 Information Retrieval + LHS 712 NLP
- We encountered an article expressing the idea of "entity-level sentiment"
  - All mentions combined, "what is the sentiment towards a specific thing"
- We wanted to compare how sentiment of an entity could vary between thematically similar, adjacent, or disjoint online communities

Background :: Reddit

- Reddit :: Platform hosting a diverse range of topic-specific online communities
- Some example “subreddit” categories ::
  - Hobbies & Recreation
    - r/Gaming, r/Hockey, r/SkateBoarding
  - Health & Health Conditions
    - r/Health, r/Fitness, r/Depression, r/Anxiety, r/EatingDisorders)
  - Drug Affinities
    - r/Drugs, r/DrugNerds, r/Stims, r/Shrooms, r/Trees, r/Psychonauts
  - Drug Moderation & Cessation
    - r/Leaves, r/Petioles, r/StopSpeeding, r/Recovery

Motivation

1) **Build an adaptable tool to compare entity-sentiment across subreddits**
    a) Specifically, we want to compare various entity sentiments between
        i) health-centric subreddit communities (r/anxiety, r/depression)
        ii) drug-centric subreddit communities (r/drug, r/stims, r/petioles, …)
2) **Explore what “cloud NLP services” have to offer**
    a) We’ve already “done entity-extraction & sentiment analysis before”
        i) Gensim, NLTK, SpaCy, TextBlob, etc
    b) We know IBM and Google advertise professional cloud-based API services for this
        i) Curious about performance of these commercial APIs

Goals, Intentions

- Goal 1 :: Build Flexible Tool to, given list of subreddits:
    - Harvest Data from Given Subreddit Communities
    - Analyze Entity-Sentiment through ~~IBM Watson or~~ Google Cloud NLP
    - Cache API-response analysis data paired with original document data
    - Analyze differences in cross-community entity-sentiment
- Goal 2 :: Compare sentiment of various terms of interest between
    - Health-Condition Subreddits (primarily r/anxiety, r/depression)
    - Substance Subreddits (r/LSD, r/stims, r/drugs, etc.)
    - Substance Moderation/Cessation Subreddits (r/addiction, r/petioles, r/leaves)

The Pipeline

Step 1: Acquisition :: getter([“uofm”, “anxiety”, “fasting”], n_posts=20)

- Requires Reddit API Key. Uses PRAW interface/wrapper.
- Generates data-frame of 20 posts from r/uofm, r/anxiety, r/fasting
- Each row has: ID, URL, subreddit of origin, title text, body text, and type(comment/post)

Step 2: Analyze Sentiment (API) :: analyze_entity_sentiment(document, encoding)

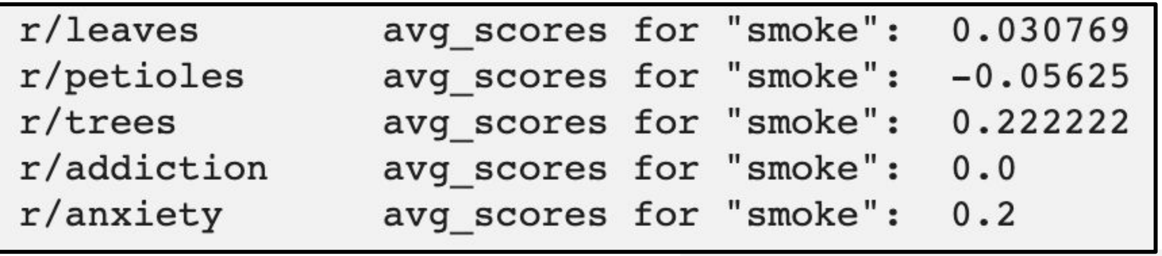
- Requires Google Cloud API Key.
- If document ID not in cache.keys(), API call for analysis & store the result

Step 3: Cache Comparison :: term_check(term, subreddits, dropZeros)

- Reorganizes data → Combine Entity-Sentiment Scores Across Posts
- *[to-do :: Reduction Step! Group “Dream” and “Dreams” :: exact method pending]*
- Output entity-level sentiment scores for entity within given subreddit

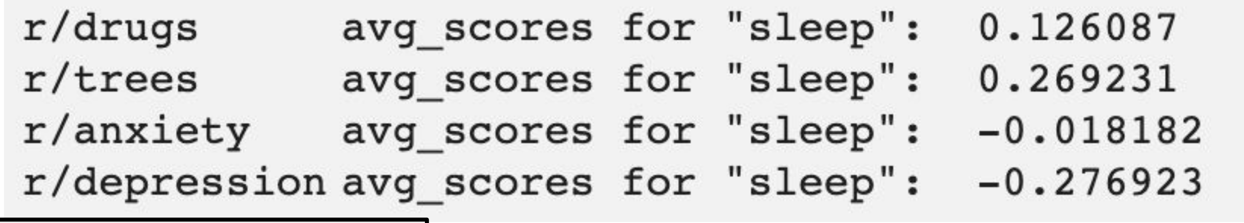
Some Initial Results

Disclaimer : currently disambiguating some API weirdness re: default value handling

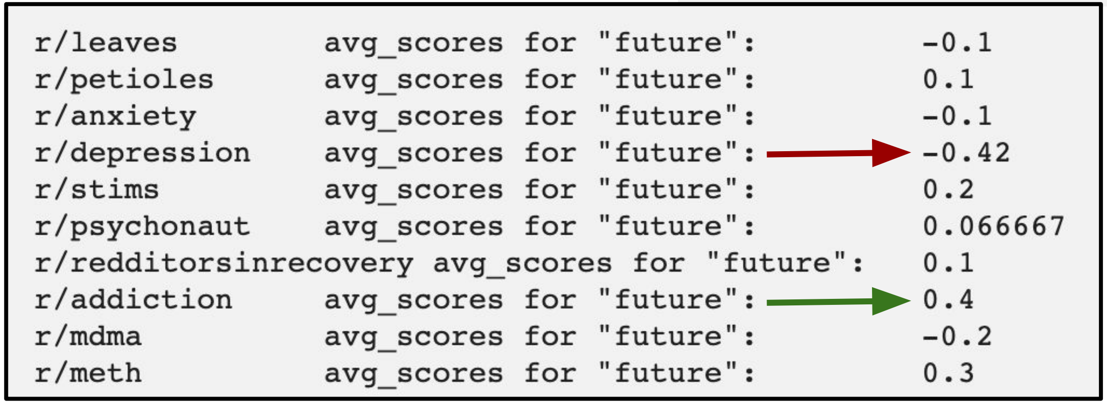


```
r/leaves          avg_scores for "smoke":  0.030769
r/petioles        avg_scores for "smoke": -0.05625
r/trees           avg_scores for "smoke":  0.222222
r/addiction       avg_scores for "smoke":  0.0
r/anxiety         avg_scores for "smoke":  0.2
```

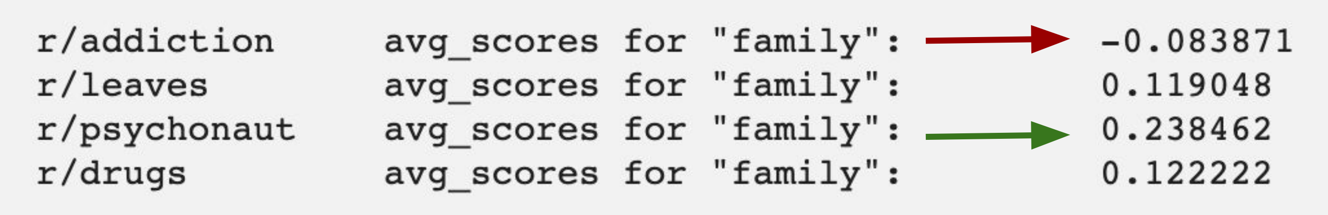
Scored from [-1.0 to +1.0]



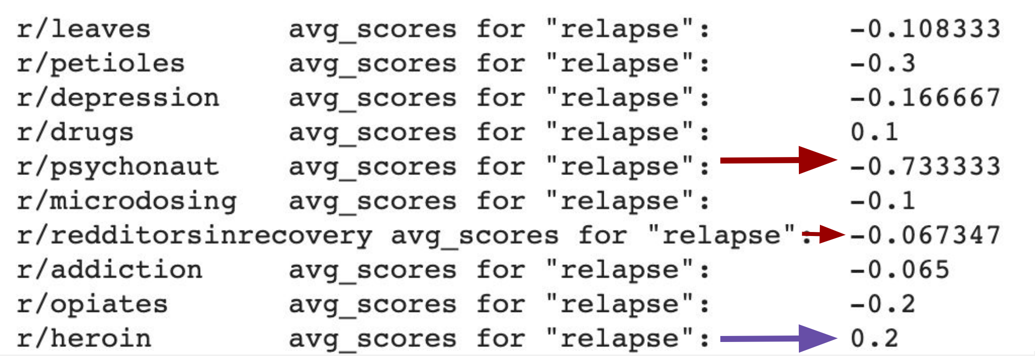
```
r/drugs      avg_scores for "sleep":  0.126087
r/trees      avg_scores for "sleep":  0.269231
r/anxiety    avg_scores for "sleep": -0.018182
r/depression avg_scores for "sleep": -0.276923
```



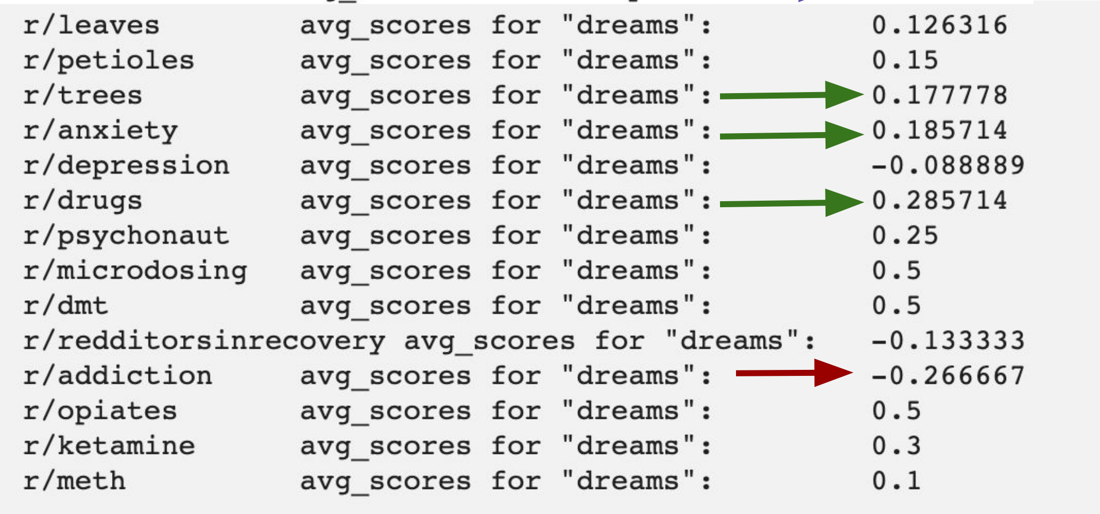
```
r/leaves            avg_scores for "future":        -0.1
r/petioles          avg_scores for "future":        0.1
r/anxiety           avg_scores for "future":        -0.1
r/depression        avg_scores for "future":        -0.42
r/stims             avg_scores for "future":        0.2
r/psychonaut        avg_scores for "future":        0.066667
r/redditorsinrecovery avg_scores for "future":      0.1
r/addiction         avg_scores for "future":        0.4
r/mdma              avg_scores for "future":        -0.2
r/meth              avg_scores for "future":        0.3
```



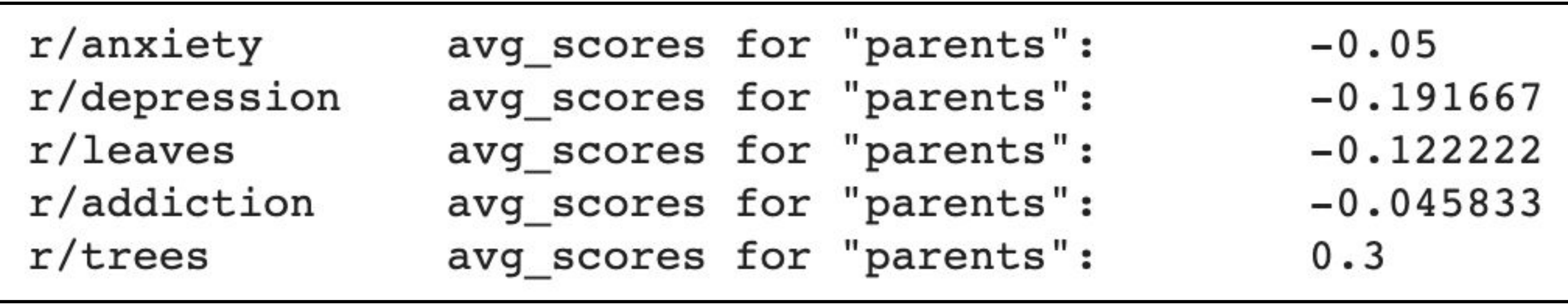
```
r/addiction      avg_scores for "family":      -0.083871
r/leaves         avg_scores for "family":       0.119048
r/psychonaut     avg_scores for "family":       0.238462
r/drugs          avg_scores for "family":       0.122222
```



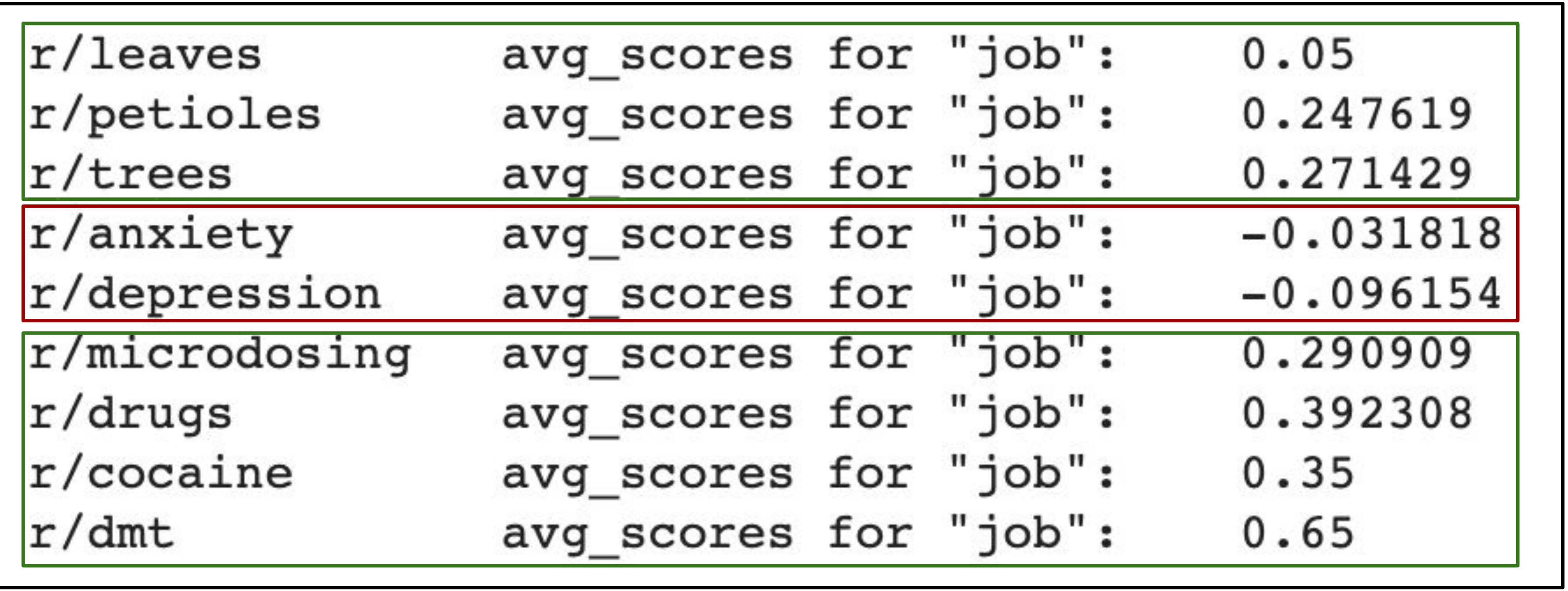
```
r/leaves            avg_scores for "relapse":       -0.108333
r/petioles          avg_scores for "relapse":       -0.3
r/depression        avg_scores for "relapse":       -0.166667
r/drugs             avg_scores for "relapse":       0.1
r/psychonaut        avg_scores for "relapse":       -0.733333
r/microdosing       avg_scores for "relapse":       -0.1
r/redditorsinrecovery avg_scores for "relapse":     -0.067347
r/addiction         avg_scores for "relapse":       -0.065
r/opiates           avg_scores for "relapse":       -0.2
r/heroin            avg_scores for "relapse":       0.2
```



```
r/leaves            avg_scores for "dreams":        0.126316
r/petioles          avg_scores for "dreams":        0.15
r/trees             avg_scores for "dreams":        0.177778
r/anxiety           avg_scores for "dreams":        0.185714
r/depression        avg_scores for "dreams":        -0.088889
r/drugs             avg_scores for "dreams":        0.285714
r/psychonaut        avg_scores for "dreams":        0.25
r/microdosing       avg_scores for "dreams":        0.5
r/dmt               avg_scores for "dreams":        0.5
r/redditorsinrecovery avg_scores for "dreams":      -0.133333
r/addiction         avg_scores for "dreams":        -0.266667
r/opiates           avg_scores for "dreams":        0.5
r/ketamine          avg_scores for "dreams":        0.3
r/meth              avg_scores for "dreams":        0.1
```



```
r/anxiety          avg_scores for "parents":        -0.05
r/depression       avg_scores for "parents":        -0.191667
r/leaves           avg_scores for "parents":        -0.122222
r/addiction        avg_scores for "parents":        -0.045833
r/trees            avg_scores for "parents":        0.3
```



```
r/leaves          avg_scores for "job":     0.05
r/petioles        avg_scores for "job":     0.247619
r/trees           avg_scores for "job":     0.271429
r/anxiety         avg_scores for "job":     -0.031818
r/depression      avg_scores for "job":     -0.096154
r/microdosing     avg_scores for "job":     0.290909
r/drugs           avg_scores for "job":     0.392308
r/cocaine         avg_scores for "job":     0.35
r/dmt             avg_scores for "job":     0.65
```

Caveats & Potential Further Exploration

- As a commercial product, the Google's Entity Sentiment API is "a black box"
  - This creates uncertainty around explainability
  - Per documentation examples, data not pre-processed, possibly worthwhile though?
- Data Excluded :: Post-Title ++ "See More" Range ++ Sub-Comments
  - All can definitely be added in future pipeline revisions as function parameters
  - Left out due to perception of negligible difference / low ROI if time invested into adding them
- Not all entities had sentiment
  - API also could have missed entities occurring in documents
  - API sometimes returned entities without sentiment attributed
  - Manual annotation to verify entity extraction and sentiments cut due to time constraints

Caveats & Potential Further Exploration

- Alternative Sentiment-Metrics Available
  - aggregate-sentiment-of-subreddit-documents, document sentiment, mention sentiment
- Antecedence Ambiguity :: What is “it”
  - API seemed unable to disambiguate antecedent tokens like [it/this/that/he/her/they]
- Variation Reduction
  - Need to collapse name-variants into cohesive clusters
    - “habit” // “habits” // “my habit” // “my new habit”
  - Many methods exist, many tradeoffs to evaluate, no single perfect answer