DLATK: Differential Language Analysis ToolKit

H. Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Johannes C. Eichstaedt, Lyle Ungar

| Stony Brook University | University of Pennsylvania | University of Washington | Qtty

Key Functionality 1: Multiple Levels of Analysis
DLATK allows one to work with a single corpus at multiple levels of analysis (document, user, date, community). At each level one can incorporate extra-linguistic information:

- **Document**: time, location, likes
- **User**: demographics, medical records, questionnaire responses
- **Community**: Census or CDC data

Key Functionality 2: Extra-linguistic information
DLATK enables incorporation of “extra-linguistic” or human/community-level attributes (e.g., examples of such information: age, gender, personality, health, income, education-level).

- Differential language analysis can utilize as either an ‘outcome’ or ‘control’ to reveal distinguishing language.
- Prediction can incorporate alongside linguistic features and has functionality to handle the heterogeneity of including both linguistic and human/community features.

Key Functionality 3: Integration of Popular Packages
- **Python**: numpy, scikit-learn, statsmodels, pandas
- **NLP**: Stanford parser, TweetNLP, NLTK
- **Other**: Mallet, IBM wordcloud
- **Install**: pip, conda, GitHub

Analysis Pipelines
- Feature Extraction (n-grams, part of speech, topics / lexica)
- Correlation (Differential Language Analysis)
- Prediction and Classification
- Dimensionality reduction and clustering
- Mediation
- Wordcloud visualization

Differential Language Analysis (DLA): the identification of linguistic features which either (a) independently explain the most variance for *continuous outcomes* or (b) are individually most predictive of discrete *outcomes* [1].

- **(a)** Prototypical use of DLATK is to perform DLA
- **(b)** Goal is to produce language that is most related to or independently discriminant of outcomes
- **(c)** Univariate, per-feature fashion or with a limited set of control variables
- **(d)** Corrects for multiple comparisons using the Benjamini-Hochberg correction

**Figure 1.** 3-grams significantly correlated with (a) age (positive; higher age), (b) age (negative; lower age), (c) educator occupation and (d) technology occupation. This was run over the Blog Authorship Corpus [2] packaged with DLATK. Here color represents the word’s frequency in the corpus (grey to red for infrequent to frequent) and size represents correlation strength.

**References**