CDIPS Data Science Workshop
Progress Report and Plan for Mentors
7/19/14
Cory Wyatt, Pranidhi Sood, Thomas Stevens

(1) Summary of progress to-date
- Sample statistics on categorization of ads.
    What numerical features were descriptive of the classification?
        price, closing time
- Understood benchmark avito.ru python script.
    - What learning algorithms were used?
        Logistic regression
    - What features were selected for the analysis, how were they converted into a usable form?
        Only the ad description field
        Sparse array of word counts
    - How does sklearn work with large datasets?
        joblib, SGD

(2) Action Items: between Saturday and Tuesday (when Check-In forms are due).
- Begin to employ logistic regression and expand the features and classifiers considered:
    -- additional features to consider: price, “other” categories, mixed Russian/English character words
    -- Identify the features that are most informative or that influence the ability of the algorithm to classify ads
- Benchmark CPU time for a run of SGD-logistic regression on the full training dataset
- Identify potential alternative approaches

(3) Check-points and Delivery Dates
    7/22 -- (see above)
    7/24 -- Results of new feature sets with same learning algorithms.
    7/29 -- New models explored. Features finalized.
    7/31 -- Presentation Mock-up. Identify areas to fine-tune or boost performance.

(4) Dates/Times Best-to-meet