Lecture 6: pandas, Data organization

LING 1340/2340: Data Science for Linguists

Jevon Heath

Objectives

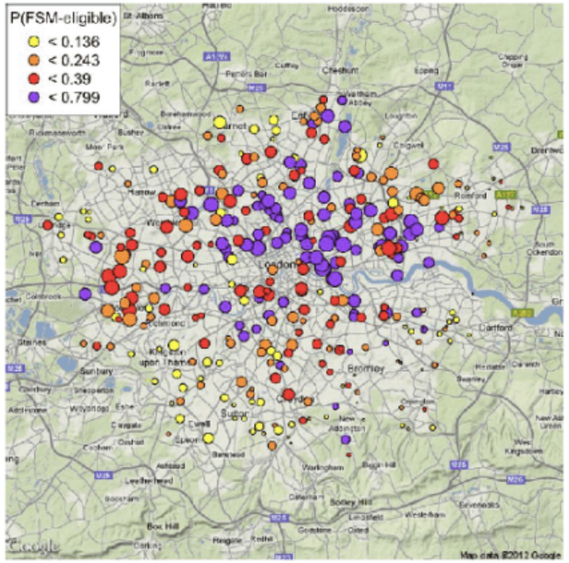
- To-do4 review: study notes in JNB
  - What did you learn from each other?
- `pandas` with linguistic data
- Data structuring and evaluation
- Tools:
  - Jupyter Notebook

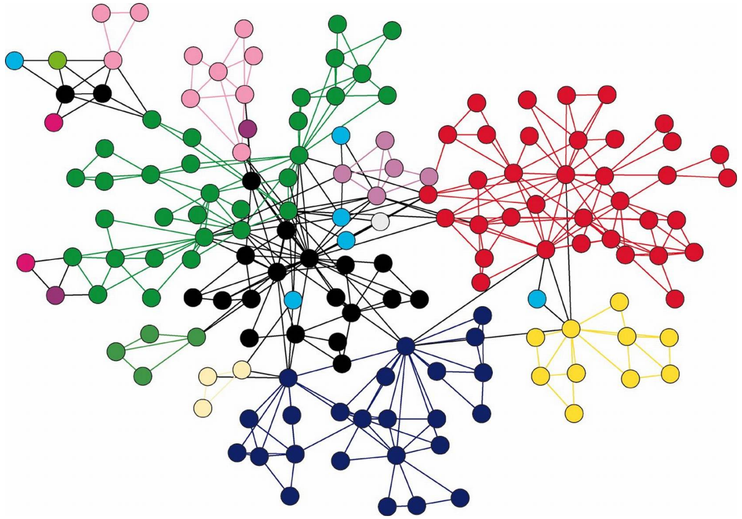
pandas practice with lexical decision times

- In Class-Exercise-Repo, `activity3/` folder:
  - You will find two files:
    - `visualize_english_BLANK.ipynb`
    - `english.csv`
  - Make a copy of the notebook file, per usu.
  - Anything you notice up front?

Data structures for statistical analysis

- Types of data:
  - nominal vs. numeric data
    - Numeric: continuous vs. discrete data
      - Continuous: interval vs. ratio data
    - nominal: categorical, binary, and ordered data
- Shapes of data:
  - rectangular
  - time series
  - spatial
  - graph





Working with rectangular data

- Rows and columns
  - Rows: observations/cases/records
  - Columns: variables/factors/features
- Why is `pandas` useful here?

EDA: understanding your data

- EDA = Exploratory Data Analysis
  - How much is there?
  - What is the magnitude (location)?
  - How much variation is there?
  - How is the data distributed?
  - How are different factors related?

How much data is there?

- ⟦This is not a property of data *per se*, but of data sets⟧
- Number of observations
  - total `df.info()`
  - per (relevant) category `df.value_counts()`
- Number of (relevant) factors `df.shape()`

What is the magnitude?

- Are we talking 0.5 or 5000 (or 500000000000000000000)?
- Estimates of location:
  - Mean `df.mean()`
    - Weighted mean `df.average(weights = df[‘Weight’])`
    - Trimmed mean `stats.trim_mean(df, proportiontocut)`
  - Median `df.median()`
- Outliers

How much variation is there?

- Is 10 a *very large value* or a *very small value*?
- Estimates of dispersion:
  - Deviation – not robust to outliers
    - Standard deviation `np.std(df)`
    - Variance `np.var(df)`
  - Median `df.median()`
  - MAD (median absolute deviation) `stats.median_absolute_deviation(df)`
  - IQR (interquartile range) `stats.iqr()`

Wrapping up

- Project ideas for Tuesday!
  - Look over last year's projects:

https://github.com/Data-Science-for-Linguists-2019