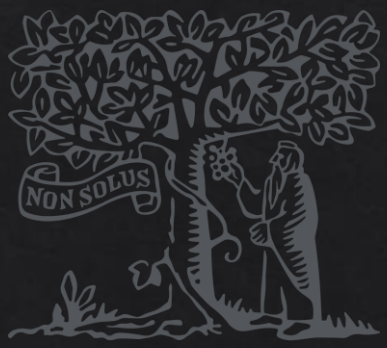
Discovering Data Sets Through Machine Learning: An Ensemble Approach to Uncovering the Prevalence of Government-Funded Data Sets

Ryan Hausen & Hosein Azarbonyad

JOHNS HOPKINS
INSTITUTE *for*
DATA-INTENSIVE
ENGINEERING & SCIENCE



ELSEVIER

Problem Setup

- Documents and labels from Coleridge Show US the Data competition: Coleridge Initiative - Show US the Data | Kaggle
  - Over 14k documents with known data sets
  - Data sets are defined, but location within the document are not given
- We examine the first, second, and third place submissions the competition
  - Each submission took a unique approach to discovering data sets
  - We further develop and refine these approaches

What does a data set reference look like?

Data Set Format

- Data Set Name

Examples

- Rural or urban residence was defined using the US Department of Agriculture's (USDA) 2003 Rural-Urban Continuum Codes.
- The county-level irrigated statistics provided by the USDA Census of Agriculture provided the spatial area target for the MIrAD-US model.
- Perhaps the best known is the USGS North American Breeding Bird Survey (http://www.mbr-pwrc.usgs.gov/), a continuing effort of about 40 years' duration.

What does a data set reference look like?

Data Set Format

- Data Set Name
- ["/']Data Set Name[,]["/']

Examples

- Data for total confirmed cases per capita were derived from the 'Our World in Data', which is a platform for global data on a broad range of conditions.
- According to the National Science Foundation "Survey of Earned Doctorates," 40% of doctoral degrees earned in political science in 2009 went to women.

What does a data set reference look like?

Data Set Format

- Data Set Name
- ["/']Data Set Name[,]["/']
- Data Set Name (DSN)

Examples

- We will draw an example from the Baltimore Longitudinal Study of Aging (BLSA), which is an ongoing research project of the National Institute on Aging of the NIH.
- Our analysis of the 2012 Program for the International Assessment of Adult Competencies (PIAAC) data indicated that, for ages 55-65, rates of AET participation are similar for both men and women.

What does a data set reference look like?

Data Set Format

- Data Set Name
- ["/']Data Set Name[,]["/']
- Data Set Name (DSN)
- Data Set (Alt-DSN) Name

Examples

- The Beginning Postsecondary Students (BPS) Longitudinal Study, in which you have been a continuing participant, is one of the major surveys used to provide this information.
- The NSF-NIH Survey of Graduate Students and Postdoctorates in Science and Engineering (GSS), which collects these data, was refined in 2007 to improve reporting.

What does a data set reference look like?

Data Set Format

- Data Set Name
- ["/']Data Set Name[,]["/']
- Data Set Name (DSN)
- Data Set (Alt-DSN) Name

Examples

- The Baccalaureate and Beyond Longitudinal Study (B&B) will provide information concerning education and work experiences after the bachelor's degree.
- A notable example of a project incorporating these techniques is the National Oceanic and Atmospheric Administration's (NOAA) Coastal Change Analysis Program (C-CAP).
- For understanding the SCS SST variability, the NOAA daily Optimum Interpolation Sea Surface Temperature (AVHRR OISSTv2) is used in this study.

What does a data set reference look like?

Data Set Format

- Data Set Name
- ["/']Data Set Name[,]["/']
- Data Set Name (DSN)
- Data Set (Alt-DSN) Name
- DSN

Examples

- These are derived from a stochastic set of wind fields, calculated using a model forced by historical observations of TCs from the IBTrACS v02r01 dataset (Knapp et al., 2010).
- Figure 3 shows examples of surges from simulations used to create the Reference Set of storms along the Mississippi coast using the SLOSH model.

Approach 1: String Matching

Submitted Approach

Given a collection of known data sets and a document to search:

1. Convert the data sets to lower case
2. Convert the document to lower case
3. For each data set, use substring matching to see if it is in the document

Advantages:

- Fast

Disadvantages:

- Must know all data sets and permutations
- Changing the case can lead to unwanted behavior in acronyms that are homographs (e.g., SLOSH, HERD → slosh, herd)

Updated Approach

- Given a collection of known data sets and a document to search:
  - Convert the data sets into a regular expression:
    - Each word is searched for both upper and lower-case versions
    - All acronyms are searched only in upper-case form

Advantages:

- Fast, but not as fast as simple substring search
- Does not confuse acronyms with common homographs

Disadvantages :

- Must know all data sets and permutations

Data Set Name ✓ ["/']Data Set Name[,]["/'] ✓ Data Set Name (DSN) ✓ Data Set (Alt-DSN) Name ✓ DSN ✓

Data Set Name ✓ ["/']Data Set Name[,]["/'] ✓ Data Set Name (DSN) ✓ Data Set (Alt-DSN) Name ✓ DSN ✓

Approach 2: Entity Classification

Submitted Approach

Given a document:

1. Find all instances of the pattern *Data Set Name (DSN)* in the text, these are entities.
2. For each entity, use a SciBERT-based classifier to classify the entity as a data set entity or not.

Advantages:

- Model learns what to separate data set like entities from non-data set like entities

Disadvantages:

- Only captures data sets that follow the name and abbreviation pattern

Updated Approach

- Given a document:
  1. Extract *entities* using a regular expression (see article).
  2. For each entity, use a SciBERT-based classifier to classify the entity as a data set entity or not.

Advantages:

- Model learns what to separate data set like entities from non-data set like entities

Disadvantages :

- Only captures data sets that follow the regular expression pattern

Data Set Name ✗ ["/']Data Set Name[,]["/'] ✗ Data Set Name (DSN) ✓ Data Set (Alt-DSN) Name ✓ DSN ✗

Data Set Name ✓ ["/']Data Set Name[,]["/'] ✗ Data Set Name (DSN) ✓ Data Set (Alt-DSN) Name ✓ DSN ✗

Approach 3: Token Classification

Submitted Approach

Train a masked language-based model:

1. Find all sentences containing a data set reference.
2. For each sentence, mask out the data set and train a model to classify if the masked entity is likely a data set reference
3. Store masked classification embeddings for inference

Advantages:

- Model learns how data sets are described

Disadvantages:

- Masked embedding vectors are randomly sampled for inference. Sampling from that distribution can produce inconsistent results

Updated Approach

- Train a Named Entity Recognition (NER) model
  1. Train a NER model to predict whether or not each word is likely to be a dataset

Advantages:

- Learns to distinguish data sets at the word-level

Disadvantages :

- Predictions at the word-level could have odd false positives (e.g., “in Science”, “Study”)

Data Set Name ✓ ["/']Data Set Name[,]["/’] ✓ Data Set Name (DSN) ✓ Data Set (Alt-DSN) Name ✓ DSN ✓

Data Set Name ✓ ["/']Data Set Name[,]["/’] ✓ Data Set Name (DSN) ✓ Data Set (Alt-DSN) Name ✓ DSN ✓

Where do we go from here?

- Openness is key
- Code ✓:
  - GitHub repository: https://purl.archive.org/democratizing-data/code
  - Develop new methods and compare them with methods in our work
- Data ✓:
  - Kaggle data: https://www.kaggle.com/c/coleridgeinitiative-show-us-the-data/
  - Dataset Mentions Detections Dataset: https://doi.org/10.1162/tacl_a_00592
- Benchmarking:
  - The initial Kaggle competition worked well for this.
  - Need a community-based approach:
    - Adding more documents to the data set
    - Improving current data set quality