Brown Data Science - Datathon 2018

DISCOVERYENGINE

March 3-4 2018

Why a DiscoveryEngine?

DiscoveryEngine developed from a perceived need for new and quantifiable measures of quality in scientific output. Drawing from a unique definition of Discovery, DiscoveryEngine seeks to quantify this most fundamental of scientific values. Our objective is to create a multi-faceted understanding of Discovery and, by doing so, to orient opportunities in science toward this goal. For the first time, the measure of scientific value will be open to the opinions and experience of the entire scientific community. And for the first time we will have the means to understand where, how, and why we create new discoveries.

Measuring Discovery

The unique insight underlying DiscoveryEngine is that Discovery can be well understood - both intuitively and rigorously - as information that, when received and understood, changes ones understanding of the subject of the information. From this definition, we propose that *Discovery Value* as the extent to which that understanding has changed. The extent of the change is a qualitative experience that nevertheless can by quantified. DiscoveryEngine, therefore, is the platform that we use to poll individual scientists about the change in understanding that a particular manuscript invokes.

While *Discovery Value* is the primary focus of DiscoveryEngine, we also poll users across two further (and potentially dissociable) axes: *Actionability* and *Confidence. Actionability* is intended to capture the extent to which a paper proposes outcomes that are practical or directly beneficial to the field. *Confidence* is the measure of individual certitude in the results - how confident is the reader in what the paper proposes? In addition to the three

primary axes of measurement, we also poll the user for their individual expertise in the subject. We hypothesize that individual expertise is a normalizing factor for the other axes.

Taken together, we see these three axes as providing a measure of the quality and the type of contribution that a paper makes. We propose that the three primary axes can be used to define what we call *Discovery Space* with each paper occupying a unique location in this space based on its contribution.

Dataset

- The dataset is composed of 680 publications rated by 266 scientists
- The dataset comprises the following variables:
    - **id** = index for the data input
    - **rating** = a value assigned to a given publication on the scale from 0 to 5
    - **create_date** = data input timestamp
    - **question_id** = identifier for questions users answer when evaluating given publication
        - 1 = *Discovery Value*
        - 2 = *Actionability*
        - 3 = *Concreteness* (version_id = 1) / *Confidence* (version_id = 2)
        - 4 = *Expertise*
    - **version_id** = rating webpage version identifier
        - 1 = 1st version asking for publication evaluation along the dimensions of *Discovery Value*, *Actionability*, *Concreteness* and *Expertise*
        - 2 = 2nd version involving the following changes:
            - Questions 1-3 were rephrased
            - 3rd metric was renamed to *Confidence*
            - Rating slider initial position was changed from 2.5 (middle) to 0 (leftmost edge)
    - **user_id** = user identifier

    - **manuscript_id** = publication identifier
    - **manuscript_DOI** = Digital Object Identifier, a unique identification number for a publication, assigned by publishers, associated with dynamic metadata

Dataset exploration

So, what's there to analyze in these data? A lot: We have collected several hundred ratings, but we are only now beginning to try to figure out what they are showing us:

- You could be the first to map *Discovery Space*, to determine whether there are types of papers, defined by their separability across our ratings variables. Are there distinct ontologies of kinds of papers? Do papers with high utility have lower *Discovery Value*? Are there other papers that score high on both?
- You could be the first to determine if there are scientific personalities, expressed in how papers are judged. Are there types of raters that typically value what they read less than the norm, or more? Are there individuals who have completely distinctive ways of associating variables, different than the norm? This topic has particular importance, as we seek to understand both whether there are 'super raters' that are better at predicting the topics that will change the future: the first step is knowing whether there are different types.
- Do traditional metrics, such as the Relative Citation Ratio (RCR), correlate with our direct measurements of *Discovery Value* and other ratings? How do 'altmetrics' data, the number of reads/downloads/mentions a paper gets, predict its directly assessed value as a scientific contribution?

Some other questions to consider:

- To what extent are the primary axes of *Discovery Space* dissociable?
- What are the domains of *Discovery Space* and do papers tend to cluster in particular domains?
- Do DiscoveryEnging metrics covary with any particular DOI metadata?
- What are the most powerful ways of visualizing this type of data? Is there a way to create a simplified representation of a paper's position in *Discovery Space*?

- Do DiscoveryEngine metrics have any predictive value?
- Is the *Expertise* variable dissociable from other DE variables?
- Are there any outliers in the data (e.g. users, publications, etc)?
- At some point the survey questions were altered, creating a before/after split in the dataset (approx equal sized datasets, v1 and v2). Are there any differences between v1 and v2?

External information and data sources

- CiteSpace = Visualizing Patterns and Trends in Scientific Literature
- Crossref = interlinks the metadata for millions of items from a variety of content types, including journals, books, conference proceedings, working papers, technical reports, and data sets
    - Example of API data query by DOI: https://api.crossref.org/works/10.1038/s41598-017-02753-6
- iCite = a tool to access a dashboard of bibliometrics for papers associated with National Institute of Health portfolio
    - Contains number of articles, articles per year, citations per year, and Relative Citation Ratio
    - Example of API data query by PubMed ID (can be obtained using other APIs via DOI): https://icite.od.nih.gov/api/pubs/23456789
- Altmetric = tracks the attention surrounding the scholarly content
    - Example of API data query by DOI: https://api.altmetric.com/v1/doi/10.1038/480426a
- PubMed = a free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics
    - Example of API data query by PMID: *https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&id=25148042*
- Mendeley = reference manager and an academic social network containing the data on publication readership and field-labels
- PaperBuzz = track the online buzz around scholarly articles
    - Example of API data query by DOI: https://api.paperbuzz.org/v0/doi/10.1038/nature12373?email=test@example.com
    - Relies on CrossRef data
- Google Scholar = web search engine that indexes the full text or metadata of scholarly literature

- Contains individual paper citations and author specific metrics (h-index, i10-index)