Common-Crawl analysis plan

# Overview

The general idea is to use a three step approach to finding the matches between foundations based on two key-word-based dimensions: Goals and Values. Step 1 is to find all documents within common crawl that contain a foundation name (the sample). Step 2 is to find the metric scores based on co-occurrence within a document for each foundation with each of the keywords. Step 3 is to create an adjacency matrix of foundations (and words) based on the correlation between their keyword scores. Based on the resulting tables, Ivar and Bob will create niche/network analysis and especially their visualizations. For this, the following input is required: The match tables containing the metric scores of each foundation with each keyword (1 table with only goal-keywords, 1 table with only value-keywords, 1 table with both keywords), the adjacency tables between foundations of each list (so goal, value and goal+value) and the adjacency table between words (agains, goals, value and both).

IF POSSIBLE it would be interesting to include 5 applicant organizations in the search to demonstrate their matching with foundations:

Arthritis Foundation: <http://www.arthritis.org/chapters/northeastern-new-york>

Academy of Friends: <http://academyoffriends.org>

Worcester County Horticultural Society: <http://www.towerhillbg.org/>

Jewish National Fund: <http://www.jnf.org/>

Academy Art Museum: <http://www.art-academy.org/>

# Measurement Concepts

The idea is to use three distinct measurements and compare the resulting matches. Each measure takes foundation name A and sequentially compares it with keyword B drawn from a list of keywords, essentially a kind of proximity or mutual information measure. The metric is computed for each name-keyword pair, yielding a list of keyword-value pairs for each foundation. The lists of different foundations or applicants and a given foundation can then be compared via simple correlation between their lists (as all foundations or organizations get a score for keywords from the same list). One is a ‘custom’ measurement selected because of it’s external validity (asymmetric weighing of name and keyword(x)). The second is an advanced measure previously employed with success, but still vulnerable to measurement errors. The third is the ‘safe bet’ measurement, which functions as a low-risk control. Having three measurements will enable us to compare the resulting matches and thereby have stronger support for the applicability of this common-crawl based approach to solving matching problems in these markets (and potential other applications of the type).

## Metric 1: Custom (asymmetric, range [0,1])

This function for each organizations-keyword pair, compares the number of co-occurance documents with different weighing for organization names and keywords. Using the formula:

custom

where is the number of documents with co-occurrences of the foundation/organization name and the keyword, the number of documents with occurrences of the foundation/organization name alone (excluding co-occurrence), the number of documents with occurrences of the keyword alone (without co-occurences with the foundation/organization) and is the total number of documents including at least one foundation name (so the total of our sample excluding documents with either the foundation name, keyword or both).

## Metric 2: Mozley, Margalef (symmetric, range [0, ∞] )

This function looks at each organization-keyword pair, and counts the number of co-occurrences in terms of shared documents, and yields an unbounded similarity coefficient.

MM =

where is the number of documents with co-occurrences of the foundation/organization name and the keyword, the number of documents with occurrences of the foundation/organization name alone (excluding co-occurrence), the number of documents with occurrences of the keyword alone (without co-occurences with the foundation/organization) and is the total number of documents including at least one foundation name (so the total of our sample excluding documents with either the foundation name, keyword or both).

## Metric 3: Jaccard coefficient (symmetric, range [0,1])

This function simply compares the number of documents with a co-occurrence versus the total number of documents contain either the foundation name or the keyword.

Jaccard =

where is the number of documents with co-occurrences of the foundation/organization name and the keyword, the number of documents with occurrences of the foundation/organization name alone (excluding co-occurrence), the number of documents with occurrences of the keyword alone (without co-occurences with the foundation/organization) and is the total number of documents including at least one foundation name (so the total of our sample excluding documents with either the foundation name, keyword or both).

*a*

*c*

*b*

*d*