DRAFT Text as Data for the Social Sciences: A Practioner's Guide

Leland Bybee Bryan Kelly

December 28, 2017

Contents

1 Introduction

Why you should care about text
- The challenges of text data
- The goal of this manual

2 Text Background

corpus
- document term matrix
- tf-idf/transformations

3 DiSTL

Intro
- Installation

```
pip install git+https://github.com/lbybee/DiSTL
```

3.1 Document Term Data-Frames

DTDF background
- building DTDF
- building DTDF from csv
- building DTDF from mongodb instance
- building options
- building workflow

3.2 Working with DTDFs

idf/tfidf
- freq terms
- word clouds
- similarity
- cosine similarity
- least squares similarity
- storing results