1 Data Preparation Software

The following is a list of a few major data preparation software options. Many computer scientists simply write code to do this processing in either Python or Perl which are relatively easy to learn for political scientists. For the more casual user or in the beginning, these software options can handle the standard data formatting requests.

1.1 tm() Package in R

The tm() package provides a library of text preparation and analysis tools that are designed to smoothly integrate with other tools. The major data preparation features include:

1. Data Import from Text Documents and PDFs
2. Convert to Plain Text
3. Eliminate White Space
4. Convert to Lower Case
5. Remove Stopwords
6. Stem Words in Multiple Languages
7. Create Term-Document Matrix

The package also provides analysis tools which include:

1. Frequent Term Location
2. Correlation Analysis
3. Dictionary Coding
4. Meta-data Management

*Handout prepared for the Tools for Text Workshop at the University of Washington. If you have anything that you feel should be on this sheet, please send me an email at bstewart@fas.harvard.edu. I’d love to know about it.
The most striking advantage of the package though is its integration with RWeka, the R front-end for the popular java-based Weka data mining library, and OpenNLP, a popular library for natural language processing tasks. Between these two libraries you have access to every standard tool in the data mining and natural language processing toolbox. These resources can be explored in the following sources:


- http://tm.r-forge.r-project.org/


1.2 Tools from Will Lowe

Will Lowe has a whole sequence of tools available at http://williamlowe.net/software/ which are useful for processing texts. YKConverter converts texts from a variety of formats and converts them to UTF-8 which is the most common encoding used for automated content analysis of all types. He also offers JFreq which performs multi-lingual stemming, stop-word removal and term-document matrix creation. These are probably the most stable, easy-to-use open source tools for performing these basic processing tasks.

1.3 Commercial Software

I personally use several pieces of professional software for basic data preparation tasks. These are useful if you have free access to them, but the extent of your individual project will determine whether or not they are worth paying for.

- Adobe Acrobat (the full version, not the reader): Acrobat can convert your PDFs into plain UTF-8 .txt files rather quickly as well as perform top of the line OCR. If you are having trouble with formatting issues using other tools you may consider trying Acrobat.

- Textpipe Pro This software suite for text processing makes it easy to write filters for just about anything. I typically use it to strip documents of html tags or other similar formatting tasks. It is very powerful although I can’t say that I have explored even a fraction of its potential.

- UltraEdit This is a powerful Window’s based tool-editor. Its two biggest advantages are that it can load extremely large .txt files (I’ve loaded files up to 2-3GB) and it has a column editor that allows you to copy and paste on a vertical axis. This can be very useful for eliminating parts of formatting that are irregular in content (so you can’t use find and replace) but regularly spaced.
• **BBEdit** This is a very powerful text editor for the Mac. I don’t generally use Macs but when I need to I use BBEdit.

• **Excel** Microsoft’s spreadsheet program can do a lot of text formatting and management of term-document matrices with a little bit of practice. You need at least 2007 though so you don’t run into size limitations.

# 2 Recommended Reference Texts for Models

Below are some writings that you should be aware of along with short summaries.

## 2.1 Canonical Texts


This textbook covers the major topics of machine learning that characterize both supervised and unsupervised document classification. He also covers variational inference and other techniques for estimating complicated Bayesian hierarchical models.


This canonical text is also available for free online. It is an extremely attractive account of machine learning clearly written from the statistical perspective. The coverage is somewhat biased to the areas in which the authors have previously worked but there is some fantastic material. It is particularly strong on regularization techniques which are very useful in models with many parameters.


This textbook covers many of the topics in supervised and unsupervised learning with particular applications to text. The chapters on clustering models and labeling clusters are brief but useful. This also covers many useful distance functions for quantitatively representing documents. The entire book is available for free in chapter PDFs on the book’s website.


This is the canonical textbook of computational linguistics. It is a bit dated at this point and I tend to prefer the newer *Introduction to Information Retrieval* but out of convention this is the much more common citation. It has also has a great deal more information on natural language processing and linguistics; however, these points are often not as interesting to political science users.

This is the canonical text book on natural language processing. If you are doing something more than bag of words representation, this is a great text for learning the necessary computer science and linguistics background. It also has a great section in the front on regular expressions.

### 2.2 Additional Useful Texts

If you are working with Bayesian hierarchical models you may find Andrew Gelman and Jennifer Hill's book *Data Analysis Using Regression and Hierarchical/Multi-level Models* useful. Gelman's textbook with several authors *Bayesian Data Analysis* is the canonical text on models of this form.

If you are getting into Python, you might consider the instruction text book *An Introduction to Computer Science Using Python* from Practical Programming. This text is a simple introduction to the core concepts of computer science that many of us who never took formal computer science missed. When you are ready you can move on to *Natural Language Processing Using Python* from O'Reilly which covers more advanced techniques in NLP.

### 3 Political Science Literature Review

The literature on document clustering in political science is extremely new and undeveloped. In my estimation the most important pieces are:


We might note that there are other relevant works. Some articles have used clustering techniques for entirely different purposes (Phil Schrodt has some work on forecasting using clustering techniques). Other papers have used clustering as a side element purely in an applied setting (such as my 2009 paper with Yuri Zhukov).
4 Resources and Packages in R

R has a wide variety of packages that implement clustering algorithms. Here are a few key resources:

• **hclust** This package contains a wide array of distance metrics and tools for hierarchical clustering. The website Quick-R has a nice tutorial [http://www.statmethods.net/advstats/cluster.html](http://www.statmethods.net/advstats/cluster.html)

• **lda** This package implements Latent Dirichlet Allocation and a variety of similar models.

• **RWeka** As mentioned above this package provides access to a wide variety of clustering algorithms and can be used in tandem with the `tm()` package

• **MCMCPack** This library contains useful tools for Bayesian statistical computing that may be helpful in building more complex models.

• **Matrix** This library offers a sparse matrix formulation that may prove useful when handling very large data.

5 Additional Articles Worth Reading

You may also find the following articles worth reading:


