

**Text Analytics and Natural Language Processing (NLP) with R**

**Course Number:** RPROG-114
**Duration:** 3 days

**Overview**

Accelebrate's Natural Language Processing (NLP) with R training course teaches attendees how to use R programming to explore and analyze text data.  This class comprehensively covers methods for ingesting text data from a variety of sources such as plain text files, pdfs, or the web, and then processing that data using the latest natural language processing and deep learning techniques.

**Prerequisites**

Students must have completed Accelebrate's [Intro to R Programming training](file:////training/r-programming-introduction) or have the equivalent experience. Students should have a working knowledge of the R language, RStudio, and the dplyr/tidyverse packages.

**Materials**

All R Programming training students receive a copy of O’Reilly's Text Mining with R and related courseware.

**Software Needed on Each Student PC**

* A recent release of R 4.x
* IDE or text editor of your choice (RStudio recommended)

**Objectives**

Students will be able to

* Import text data from a variety of source formats
* Tokenize text data to meaningful units
* Wrangle text data using specific textual functions
* Compute aggregating measures on tokenized data
* Translate between text data formats
* Complete a sentiment analysis
* Perform document classification
* Perform topic modeling
* Build a simple neural network appropriate for NLP modeling

**Outline**

* Working with unstructured text data
	+ string methods
	+ regex
	+ reading in text files
	+ review of base (R/Python)
* Importing
	+ parsing data from a text file
	+ importing it into a tidy structure
	+ parsing data from a pdf
		- From a “pile of pdfs”
	+ scraping data from the web
	+ Discussion of other methods
		- OCR
		- Handwriting recognition
* Managing Text Data 1
	+ a tidy text format
	+ Overview of text data formats
		- tidy text
		- token list
		- Bag of words
		- document term matrix or document frequency matrix (dfm/dt)
		- corpus
		- docvars
	+ associated formats
		- stop words
		- Sentiment lexica
		- word vectors / models
* Managing Text Data 2
	+ tokenizing text
	+ units of tokenization
		- tokens
		- lemma
		- stems
		- n-grams
		- sentences
		- Tweets
	+ Tf-idf
	+ Log-odds (tidylo)
* Sentiment Analysis
	+ Sentiment lexica
	+ Sentiment analysis with inner\_join
	+ Analyzing by other units
	+ Valence shifting
	+ VADER
* Document Classification
	+ Text similarity - stringiest
		- Cosine
		- Edit distance
	+ Machine Learning for document classification
		- Naive Bayes model
* Topic Modeling / Document Clustering
	+ LDA
	+ stm
* Text and Deep Learning
	+ Deep learning introduction
	+ Architecture of neural networks
	+ Tensorflow + keras
	+ Word vectors
		- word2vec
		- Text2vec
		- GloVe
		- Spacy
	+ Combining Deep Learning and NLP
		- CNN
		- RNN
		- LSTM
	+ Named Entity Recognition (NER)
	+ Part of Speech tagging (POS)
	+ Dependency Parsing
* Conclusion