Text Classification with RTextTools

Odum Institute University of North Carolina

Loren Collingwood[1] (with Tim Jurka, Amber Boydstun, Emiliano Grossman, and Wouter van Atteveldt)

February 6, 2012

[1] Political Science, University of Washington

Workshop Road Map

- Brief lecture on content analysis, RTextTools and machine learning

Workshop Road Map

- Brief lecture on content analysis, RTextTools and machine learning
- Introduction to R

Workshop Road Map

- Brief lecture on content analysis, RTextTools and machine learning
- Introduction to R
- Classifying newspaper data – most basic example

Workshop Road Map

- Brief lecture on content analysis, RTextTools and machine learning
- Introduction to R
- Classifying newspaper data – most basic example
- Classifying congressional bill data

Workshop Road Map

- Brief lecture on content analysis, RTextTools and machine learning
- Introduction to R
- Classifying newspaper data – most basic example
- Classifying congressional bill data
- Classifying blog data

Workshop Road Map

- Brief lecture on content analysis, RTextTools and machine learning
- Introduction to R
- Classifying newspaper data – most basic example
- Classifying congressional bill data
- Classifying blog data
- Classifying survey open-ends data

Workshop Road Map

- Brief lecture on content analysis, RTextTools and machine learning
- Introduction to R
- Classifying newspaper data – most basic example
- Classifying congressional bill data
- Classifying blog data
- Classifying survey open-ends data

Content Analysis

- Study of recorded human communication

Content Analysis

- Study of recorded human communication
- Summary and quantitative analysis of communicated messages

Content Analysis

- Study of recorded human communication
- Summary and quantitative analysis of communicated messages
- Researcher looks for patterns/themes in text; develops "code frame" to categorize text

Content Analysis

- Study of recorded human communication
- Summary and quantitative analysis of communicated messages
- Researcher looks for patterns/themes in text; develops "code frame" to categorize text
- Essentially, variables are extracted from text

Content Analysis

- Study of recorded human communication
- Summary and quantitative analysis of communicated messages
- Researcher looks for patterns/themes in text; develops "code frame" to categorize text
- Essentially, variables are extracted from text
- Based on scientific method; establishes objectivity via inter-coder reliability

Pros and Cons of Content Analysis

- Very flexible

Pros and Cons of Content Analysis

- Very flexible
- Create all sorts of variables for data summarization

Pros and Cons of Content Analysis

- Very flexible
- Create all sorts of variables for data summarization
- Build theoretically motivated classification scheme (code frame)

Pros and Cons of Content Analysis

- Very flexible
- Create all sorts of variables for data summarization
- Build theoretically motivated classification scheme (code frame)
- Can apply to written language, speech, video

Pros and Cons of Content Analysis

- Very flexible
- Create all sorts of variables for data summarization
- Build theoretically motivated classification scheme (code frame)
- Can apply to written language, speech, video
- Manually intensive

Pros and Cons of Content Analysis

- Very flexible
- Create all sorts of variables for data summarization
- Build theoretically motivated classification scheme (code frame)
- Can apply to written language, speech, video
- Manually intensive
- Establishing inter-coder reliability takes time and serious attention to detail

Pros and Cons of Content Analysis

- Very flexible
- Create all sorts of variables for data summarization
- Build theoretically motivated classification scheme (code frame)
- Can apply to written language, speech, video
- Manually intensive
- Establishing inter-coder reliability takes time and serious attention to detail
- Can be expensive

Pros and Cons of Content Analysis

- Very flexible
- Create all sorts of variables for data summarization
- Build theoretically motivated classification scheme (code frame)
- Can apply to written language, speech, video
- Manually intensive
- Establishing inter-coder reliability takes time and serious attention to detail
- Can be expensive

What is RTextTools?

- R package for automating certain types of content analysis

What is RTextTools?

- R package for automating certain types of content analysis
- Uses supervised learning methods to automate text classification

What is RTextTools?

- R package for automating certain types of content analysis
- Uses supervised learning methods to automate text classification
- Uses many pre-existing text and machine learning R packages

What is RTextTools?

- R package for automating certain types of content analysis
- Uses supervised learning methods to automate text classification
- Uses many pre-existing text and machine learning R packages
- Built in text pre-processing and analytics

What is RTextTools?

- R package for automating certain types of content analysis
- Uses supervised learning methods to automate text classification
- Uses many pre-existing text and machine learning R packages
- Built in text pre-processing and analytics
- Fairly simple and intuitive to use, even for notice R users

What is RTextTools?

- R package for automating certain types of content analysis
- Uses supervised learning methods to automate text classification
- Uses many pre-existing text and machine learning R packages
- Built in text pre-processing and analytics
- Fairly simple and intuitive to use, even for notice R users
- Memory issues in R and text analysis in general

What is RTextTools?

- R package for automating certain types of content analysis
- Uses supervised learning methods to automate text classification
- Uses many pre-existing text and machine learning R packages
- Built in text pre-processing and analytics
- Fairly simple and intuitive to use, even for notice R users
- Memory issues in R and text analysis in general

Origins of the Project

- Policy Agendas Project

Origins of the Project

- Policy Agendas Project
- Congressional Bills Project

Origins of the Project

- Policy Agendas Project
- Congressional Bills Project
- Comparative Agendas Project

Origins of the Project

- Policy Agendas Project
- Congressional Bills Project
- Comparative Agendas Project
- TextTools

Origins of the Project

- Policy Agendas Project
- Congressional Bills Project
- Comparative Agendas Project
- TextTools
- Rtexttools

Origins of the Project

- Policy Agendas Project
- Congressional Bills Project
- Comparative Agendas Project
- TextTools
- Rtexttools
- RTextTools

Origins of the Project

- Policy Agendas Project
- Congressional Bills Project
- Comparative Agendas Project
- TextTools
- Rtexttools
- RTextTools

What is Machine Learning?

- Subfield of artificial intelligence

What is Machine Learning?

- Subfield of artificial intelligence
- Computer “learns” from empirical data

What is Machine Learning?

- Subfield of artificial intelligence
- Computer “learns” from empirical data
- Evolves behavior based on what is learned

What is Machine Learning?

- Subfield of artificial intelligence
- Computer “learns” from empirical data
- Evolves behavior based on what is learned
- Can make informed decision given new virgin data

What is Machine Learning?

- Subfield of artificial intelligence
- Computer "learns" from empirical data
- Evolves behavior based on what is learned
- Can make informed decision given new virgin data
- Basically... like regression except text are variables/data

What is Machine Learning?

- Subfield of artificial intelligence
- Computer "learns" from empirical data
- Evolves behavior based on what is learned
- Can make informed decision given new virgin data
- Basically... like regression except text are variables/data
- Supervised learning is a specific type of machine learning

What is Machine Learning?

- Subfield of artificial intelligence
- Computer "learns" from empirical data
- Evolves behavior based on what is learned
- Can make informed decision given new virgin data
- Basically... like regression except text are variables/data
- Supervised learning is a specific type of machine learning

How Does Supervised Learning Work?

- User presents classified data to software

How Does Supervised Learning Work?

- User presents classified data to software
- Learning algorithm creates a "behavioral model", and adjusts behavior given function parameters

How Does Supervised Learning Work?

- User presents classified data to software
- Learning algorithm creates a "behavioral model", and adjusts behavior given function parameters
- Software then classifies data the computer has never seen

How Does Supervised Learning Work?

- User presents classified data to software
- Learning algorithm creates a "behavioral model", and adjusts behavior given function parameters
- Software then classifies data the computer has never seen

When Should a Researcher Use It?

- You have a large corpus of text that your undergrad has already manually coded into pre-assigned topic labels

When Should a Researcher Use It?

- You have a large corpus of text that your undergrad has already manually coded into pre-assigned topic labels
- Then your undergrad quits, but you still have much more topic labeling to do

When Should a Researcher Use It?

- You have a large corpus of text that your undergrad has already manually coded into pre-assigned topic labels
- Then your undergrad quits, but you still have much more topic labeling to do
- You don't want to do the manual labeling because that is manually intensive

When Should a Researcher Use It?

- You have a large corpus of text that your undergrad has already manually coded into pre-assigned topic labels
- Then your undergrad quits, but you still have much more topic labeling to do
- You don't want to do the manual labeling because that is manually intensive
- Supervised learning automates the labeling of a large portion of remaining text documents

When Should a Researcher Use It?

- You have a large corpus of text that your undergrad has already manually coded into pre-assigned topic labels
- Then your undergrad quits, but you still have much more topic labeling to do
- You don't want to do the manual labeling because that is manually intensive
- Supervised learning automates the labeling of a large portion of remaining text documents
- But you are likely to still have to manually label some of the documents (active learning)

When Should a Researcher Use It?

- You have a large corpus of text that your undergrad has already manually coded into pre-assigned topic labels
- Then your undergrad quits, but you still have much more topic labeling to do
- You don't want to do the manual labeling because that is manually intensive
- Supervised learning automates the labeling of a large portion of remaining text documents
- But you are likely to still have to manually label some of the documents (active learning)

What Do You Need?

- An Excel (or other) file with manually coded data

What Do You Need?

- An Excel (or other) file with manually coded data
- A substantial number ($>$3000 documents) of manually labeled documents

What Do You Need?

- An Excel (or other) file with manually coded data
- A substantial number ($>$3000 documents) of manually labeled documents
- One (or more) column(s) for text data

What Do You Need?

- An Excel (or other) file with manually coded data
- A substantial number ($>$3000 documents) of manually labeled documents
- One (or more) column(s) for text data
- One column for topic label

What Do You Need?

- An Excel (or other) file with manually coded data
- A substantial number ($>$3000 documents) of manually labeled documents
- One (or more) column(s) for text data
- One column for topic label

What the Data May Look Like

Article_ID	Date	Title	Subject	Topic Code
41246	1-Jan-96	Nation's Small	Jails overwhel	12
41257	2-Jan-96	FEDERAL IMPA	Federal budge	20
41268	3-Jan-96	Long, Costly P	Contenders fo	20
41279	4-Jan-96	Top Leader of	Bosnian Serb l	19

Basic Workflow

- Import your hand-coded data into R

Basic Workflow

- Import your hand-coded data into R
- Remove "noise" from your data, and create a text corpus the computer can interpret

Basic Workflow

- Import your hand-coded data into R
- Remove "noise" from your data, and create a text corpus the computer can interpret
- Use algorithm(s) to train a model

Basic Workflow

- Import your hand-coded data into R
- Remove "noise" from your data, and create a text corpus the computer can interpret
- Use algorithm(s) to train a model
- Test on reference out-of-sample data; establish accuracy criteria

Basic Workflow

- Import your hand-coded data into R
- Remove "noise" from your data, and create a text corpus the computer can interpret
- Use algorithm(s) to train a model
- Test on reference out-of-sample data; establish accuracy criteria
- Use model to classify virgin data

Basic Workflow

- Import your hand-coded data into R
- Remove "noise" from your data, and create a text corpus the computer can interpret
- Use algorithm(s) to train a model
- Test on reference out-of-sample data; establish accuracy criteria
- Use model to classify virgin data
- Manually label data that do not meet accuracy criteria

Basic Workflow

- Import your hand-coded data into R
- Remove "noise" from your data, and create a text corpus the computer can interpret
- Use algorithm(s) to train a model
- Test on reference out-of-sample data; establish accuracy criteria
- Use model to classify virgin data
- Manually label data that do not meet accuracy criteria

Main Functions

- create_matrix

Main Functions

- create_matrix
- create_corpus

Main Functions

- create_matrix
- create_corpus
- train_model or train_models

Main Functions

- create_matrix
- create_corpus
- train_model or train_models
- classify_model or classify_models

Main Functions

- create_matrix
- create_corpus
- train_model or train_models
- classify_model or classify_models
- create_analytics

Main Functions

- create_matrix
- create_corpus
- train_model or train_models
- classify_model or classify_models
- create_analytics
- Today and tomorrow we will walk through several examples using these and other functions and bits of code

Main Functions

- create_matrix
- create_corpus
- train_model or train_models
- classify_model or classify_models
- create_analytics
- Today and tomorrow we will walk through several examples using these and other functions and bits of code

Thank You

- Any questions contact Loren Collingwood lorenc2@uw.edu