**Instructional Day:** 20-22

**Topic Description:** In this lesson, computation with text is explored. A variety of filters and queries are used to create subsets of text data. Bar charts are used for graphical display.

**Objectives:**

The students will be able to:

* Read in a file containing text as data.
* Filter a text data set (remove punctuation, remove case, remove stop words, stemming).
* Create a bar chart as one method of analyzing text.
* Create and query subsets of a text data set.

**Outline of the Lesson:**

* Introduction to text data (25 minutes)
* Introduction to Text Activity (30 minutes)
* Basic Analytics (20 minutes)
* Computing with Text Activity Part I (30 minutes)
* Journal Entry (5 minutes)
* Focusing on the words (20 minutes)
* Computing with Text Activity Part II (30 minutes)
* Wrap up Question (5 minutes)

**Student Activities:**

* Participate in discussion of text data.
* Complete Introduction to Text Data Activity.
* Complete journal entry.
* Complete Parts I and II of the Computing with Text Activity and participate in the discussions associated with the activity.
* Provide responses to the wrap up question and participate in discussion.

**Teaching/Learning Strategies:**

* Introduction to text data (You can use Introduction to Text Data as a resource.)
  + Explain secondary uses of data.
  + Provide Twitter background.
    - Ask students what words or phrases people may use to describe the first warm day after winter.
  + Look at Jillamore pdf.
    - Have students plot her location on a map.
  + Look at Weather Underground site.
    - Have students search for Jillamore’s location.
  + Load the weather data file.
  + Navigate through the file.
    - Ask students to provide an explanation of what the variables are, etc.
  + Load the twitter data file. Navigate through the file.
    - Ask students to explain the variables, etc. Ask questions such as: What terms were searched for? Which had the most tweets?
    - Point out that this original file is very large, so to improve performance of the computers and make it easier for them to work with the file they will be using a subset. This is a good opportunity to remind students that computing is a powerful tool and allows working with large data sets, but is limited by things like processor speed and memory. (This can be linked back to the lessons in Unit 1.)
    - Demo how to create a subset of the data for a specific region of the country by picking latitude and longitude boundaries.
  + Load the twitter with date data file.
    - Point out the created date variable.
  + Have students complete the Introduction to Text Activity.
    - Discuss results.
    - Note that students may have some difficulty creating the subset expressions.
* Text Analysis
  + Part I—Basic Analytics
    - Text can be analyzed many different ways. Research areas like "stylometrics" attempt to say something quantitative about an author's work; e.g., by computing the average number of words per sentence or the average number of letters per word written by an author. Analytics can also be used to find patterns in other types of text. (You can use Computing with Text Background as a resource.) In the first part, the basics of counting words in a file and creating bar charts based on those counts will be addressed by working with the California subset of the twitter data file. Demonstrate how to do the following:
      * Load the CATwitter data file.
      * Look at the tweets.
      * Change the size of the column in order to view the entire tweet.
      * Scroll and look at the variety of tweets.
    - Text mining—analyzing word counts. Demonstrate how to do the following:
      * Turn the vector of tweets into a "corpus". A corpus is the term used to describe a collection of writings. This is necessary in order to do some more sophisticated analysis.
      * Demo creating the corpus.
      * This is a good opportunity to explain that the tweets are stored in an array (or vector) where the numbers in front indicate the place the tweet is in the vector. Arrays are an important concept in computer science. Storing items in an array allows us to access particular elements, search and sort.
      * Demo how to view the corpus and point out that each of the array elements of the corpus matches the corresponding tweet in the data file.
      * Create a frequency table that separates out each word and counts how many times it appears in all the tweets.
      * Ask questions such as: What is the word that appears least frequently? What is the word that appears most frequently?
      * Demo how to produce frequency tables that show only the most frequently appearing words and the different sorting options.
      * Demo how to produce a bar chart of frequently occurring words.
      * Journal Entry: What do you think would happen if you did all of these same things on the NJ subset?
    - Have students complete Part I of Computing with Text Activity.
      * Discuss results before going on to Part II.
  + Part II—Focusing on the Words
    - Demonstrate how to do the following:
      * Remove case. Make “Spring” and “spring” be the same thing by making everything lowercase. (Note: each new corpus that is created should be assigned a new name.)
      * Removing "stop" words. Some words like “a” and “the” are probably always going to appear frequently because they are common parts of speech. Those words can be removed to emphasize the other less common words. Demo the method for removing stop words.
      * Deleting punctuation. Notice that many of the captions include symbols other than numbers and letters. Demo the method for deleting punctuation.
      * Stemming. It might be useful to ignore the ending of words such as “s”, “ing”, etc. In other words, change words like “boats” and “boating” to just “boat”. This is called stemming. Demo stemming.
    - Have students complete Part II of Computing with Text Activity.
      * Discuss results.
* Wrap up question: What is the source of the words that will be analyzed for your final project?
  + Ask students to provide a response. Make sure they understand that the answer is any of the text they enter that is “free text”.

**Resources:**

* Introduction to Text Data
* Jillamore.pdf
* Introduction to Text Activity
* Computing with Text Background
* Computing with Text Activity
* Deducer Quick Start Guide

**Introduction to Text Data**

**\*\*Secondary uses of data**

Data that are publicly available on the web are subject to a host of secondary uses. As the consumer of these data, there are a few questions to ask: Who collected the data and why were they collected? When was the data collected and how old is the data? What was their original purpose? What are the strengths and limitations of these data for your problem? How are the data organized? How do you access them? Is there someone who you can ask for help if you have questions?

**\*\*Twitter and the Jillamore file**

Twitter is a micro-blogging site that handled 4 billion messages or tweets in the first three months of 2010 alone. As a social network, Twitter culls activities from millions of people and there have been several studies of what people are posting to Twitter.

It is possible to look to Twitter and its users for signs of spring. Somewhere in the daily observations of millions of people it should be possible to find comments about the changing season.

The easiest place to start is with "Spring is here". On April 5, 2010 at about 1 pm a search for the phrase "spring is here" was submitted. Jillamore.pdf is a screen shot of the search results.

The last tweet was from the user "jillamore" who comments that it is a beautiful day in her part of the country, with temperatures in the upper 70s. She declares "Spring is here!" A bit more about this person can be learned from the last page in the Jillamore.pdf file. She lives at 40.360171 latitude-74.079609 longitude. The point can be plotted on a map. (Enter latitude and longitude into Google Maps, for example.)

This point is Red Bank, New Jersey. To get a sense of what the weather has been like in her part of the world, a service like the Weather Underground can be examined.

[http://www.wunderground.com](http://www.wunderground.com/)

It allows searches for weather anywhere in the country. This site is interesting to us both because it is possible to see what the weather has been like for jillamore and also because of where the data to make this judgment comes from. The Weather Underground culls data from about 10,000 officially run weather stations (e.g., the National Oceanic and Atmospheric Administration or NOAA) and 8,500 that are privately run but subject to strict data quality controls. The idea that citizens would install sensors and volunteer their data is very much in the spirit of the phone applications being used.

Search the Weather Underground for Red Bank, NJ. It shows historical weather data for this city. A file starting from January 1, 2010 to April 5, 2010, when the tweet about spring was posted was created from this information.

Load the weather data file and look at it in the Data Viewer.

Each row represents a different day and the names of the different variables recorded for each day appear in the first row.

**\*\* Some historical data**

A group of researchers made hourly requests from the Twitter API using several different phrases in addition to "spring is here"—some relating to things turning green or trees beginning to bud. The researchers were also interested in the beginning of fall, so they also collected data on phrases like "fall is here" and comments about leaves turning colors.

Load the twitter data file and look at it in the Data Viewer.

Each row represents a single tweet. The variables include "created" which is a timestamp; Twitter "username" from the person who wrote the tweet; the "longitude" and "latitude" of their location (either their home or, if they are using a smart phone, the place where they typed in their tweet); which of the researcher's "search\_term"s the tweet matched; and then the "message" itself. The variable "search\_term" is a factor (categorical).

Load the twitter with date data file and look at it in the Data Viewer.

This file was created from the original twitter file and in addition to being a small subset; it includes a variable that indicates the created date in date format. (Note: to sort by date, the numerical created date needs to be used.) This file can be subsetted further by choosing latitude and longitude boundaries. For example, the New Jersey area would be approximately bounded by latitude between 38.5 and 41.5 and longitude between -75.5 and -73.5.

**Introduction to Text Activity**

Load the twitterwithdate data file and the weather data file into the Deducer Data Viewer.

1. Create a subset of the twitter data that includes only the tweets that contain "Spring is here".

* + - How many tweets contain “Spring is here”?
    - What other search terms could you include that might indicate spring?
    - Create a subset that includes “Spring is here” and at least one other search term. How many tweets were added?

2. Create a subset of all tweets from approximately the New Jersey area.

* + - Look at locations near where Jillamore lives. How many tweets are from that area? Do they match Jillamore’s description? What other ways could you use to verify this?
    - Plot the New Jersey subset on a map. Experiment with different point sizes and zoom levels. What inferences can you make from the plot?
    - Sort by created. Do the dates and search terms correspond correctly?
    - Who has the most tweets? How does that impact the total number of tweets from New Jersey? How does that impact the number of tweets that include “Spring is here”?

3. Create a subset of California.

* + - How many tweets are from that area?
    - Plot the California subset on a map.

4. What reasons can you think of to explain the difference in the number of tweets between New Jersey and California? How might you test your reasoning?

**Computing with Text Background**

A book usually has a fairly predictable structure. There are chapters which are made up of paragraphs which are made up of sentences which are made up of words. Research areas with names like "stylometrics" attempt to say something quantitative about an author's work. It is possible to compute the average number of words per sentence or the average number of letters per word written by an author. Some authors write in short, choppy sentences, while others craft sentences that are over a page long, adding phrase after phrase. Some authors choose simple vocabulary, while others prefer long, complex words. Statistics of this kind can not only point out interesting ways to think about the differences between authors, but they can even be used to help us figure out who wrote texts if their author is unknown or uncertain. One of the earliest analyses of this kind was of the famed Federalist Papers, a collection of documents describing the philosophy motivation behind our system of government. The papers are thought to be written by Alexander Hamilton, James Madison and/or John Jay. In the mid 1960s, a group of statisticians considered a number of novel statistics to differentiate the writing styles of the three men.

The counts of the different words in a document have also been used to characterize something about the document's subject.

The idea that the frequency with which words appear in a document might reflect something of its content has real-world applications. For example, the spam filter that intercepts junk e-mail is working on the frequency of words in each message. If a message makes too many references to "sales" or "won" or "Visa", there is a strong suspicion that the e-mail is spam.

The goal of this section is not to develop any of the topics above in any great depth. Instead, it will provide some basic tools for simple analysis on text.

**Computing with Text Activity**

**\*\*Part I—Basic Analytics**

Load the NJTwitter data file into theDeducer Data Viewer.

1. Create a corpus from NJTwitter.

2. View the corpus. Do the elements of the corpus match the messages in the NJTwitter data file?

3. Create a frequency table that counts how many times each word appears in all the tweets. What is the word that appears least frequently? What is the word that appears most frequently?

4. Create a frequency table that shows only the top 1% of frequently appearing words sorted by ascending frequency. What are the most frequently occurring words?

5. Create a bar chart of the most frequently occurring words. Does this match your frequency table?

6. Experiment with different percentages to determine the greatest percentage that allows you to read all of the words on the chart. Describe your process and the reasoning for your final answer.

7. Notice that there is a “spring” and a “spring!” (with an exclamation point). Do you think we should include those counts together or keep them separate? Why?

**\*\*Part II—Focusing on the Words**

Another data set was collected by using the API (Application Programming Interface) to conduct a search on Flickr for images that were tagged with the word "chill". The first 3,000 image captions in the list of search results were downloaded. Those captions containing words that were inappropriate were removed. In order to make it easier to work with the file, a subset was created.

Load the smallcaptions data file into the Deducer Data Viewer. Enlarge the column so that the entire caption can be viewed.

1. Change the list into a corpus. View several of the elements of the corpus. Do they match the file?

2. Create a frequency table for the entire list. What is the most frequently occurring word? Scroll to look through the entire file.

3. Run the frequency for only the top 10%. What does this do? Describe anything that you notice.

4. Run the frequency for the top 1%. Make a list of the words and their counts.

5. Create a bar chart for the top 1%. Describe what you see. Save the chart to a document.

6. Experiment with a few more % choices. Which gives the most information? Explain your choice.

7. Create a frequency table of the top 1% of words after making them all lower case and without punctuation. Make a list of the words and their counts.

8. Create a bar chart for the top 1%. Describe what you see. Save the chart to a document.

9. Experiment with a few more % choices. Which gives the most information? Explain your choice.

10. Create a frequency table of the previous corpus after deleting stop words. How did the file change?

11. Create a bar chart of the file without stop words. What are some of the words that disappeared from your bar chart? Why might it be useful to delete these stop words?

12. Create a frequency table of the previous corpus after deleting stems. How did the file change?

13. Create a bar chart of the file without stems. What are some of the words that disappeared from your bar chart? Why might it be useful to delete these stems?