

# Next Word Prediction

Using n-gram(4) Models

Jay Yanamandala

12/5/2021

## Journey of the Shiny Predictor App

This presentation features the following

- Creating n-grams data set used to predict next word in the Shiny App
- Knowledge capture during the creation of data-set and Shiny App
- Next Word Prediction - why this specific approach

Links to the app and source code

- Source Code on GitHub
- Shiny App

## Data Wrangling and Tidy Data Set Creation

The raw data contains corpora in 4 different languages, for this project only en\_US locale files were downloaded.

Text mining was implemented using the **tidy** libraries. Due to huge size of twitter and blog files, used only 33% of their size in this project.

**Some details about the data:**

Stats	News	Blogs	Twitter
File.Size	196.2775	200.4242	159.3641
File.Length	77259	899288	2360148
Longest.Line	5760	40833	140

## Next Word Prediction Model

The predictive text model was built from a sample of 500k lines extracted from a random sample of a large corpus of blogs, news and twitter data (over 4 million lines).

The sample data was tokenized and cleaned with tidytext. Cleaning process included removing profane words, all non-ascii characters, and all words were lower-cased. The strings were then split into tokens (n-grams =4).

For the Shiny App, used quadgram (n-gram=4), basically a string of 4 words, datatable with frequencies of occurrence to predict next word.

## The Shiny App

Users enter text in the text box and the predicted words are posted below the text box

Users can also select the number of words they want to predict by inputting in the amount of words below the text box.

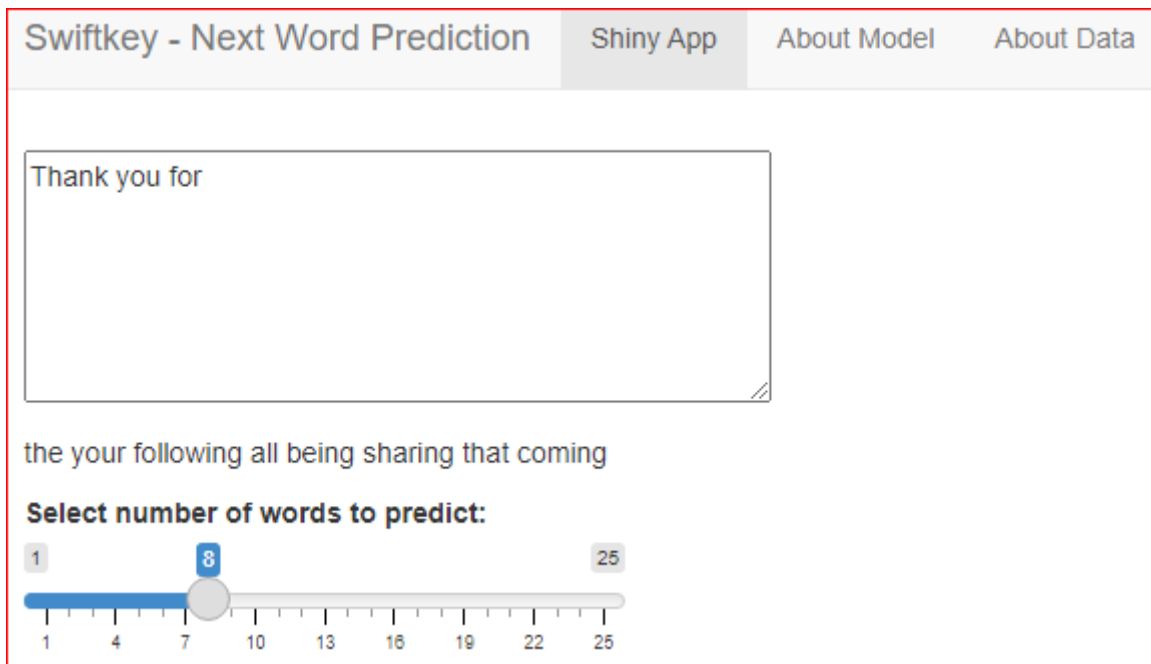
The screenshot shows a web application titled "Swiftkey - Next Word Prediction". It has three tabs: "Swiftkey - Next Word Prediction" (active), "Shiny App", "About Model", and "About Data". Below the tabs is a large text input box containing the text "Thank you for". Below the input box, the text "the your following all being sharing that coming" is displayed. Underneath this, there is a section titled "Select number of words to predict:" followed by a horizontal slider. The slider has a range from 1 to 25, with major tick marks at 1, 4, 7, 10, 13, 16, 19, 22, and 25. A blue circle marker is positioned at the value 8 on the slider.

Figure 1: Shiny App

## Summary

- Due to machine resource limitations, had to limit reading of blogs.txt, and twitter.txt files to 33% of their respective file sizes
- The Shiny App was built limiting n-grams data set to 500,000 rows
- Need more data to improve prediction accuracy and effectively suggest words that lead to higher usage by User

## References:

- Text Mining with R (Julia Silge & David Robinson)
- The Life-Changing Magic of Tidying Text | Julia Silge