Milestone Report

Noor Ahmed

Thursday, March 19, 2015

Introduction

Around the world, people are spending an increasing amount of time on their mobile devices for email, social networking, banking and a whole range of other activities. But typing on mobile devices can be a serious pain. SwiftKey, our corporate partner in this capstone, builds a smart keyboard that makes it easier for people to type on their mobile devices. One cornerstone of their smart keyboard is predictive text models.

Objective

The goal of this capstone project is to build a model that can predict the text accurately while user is typing and provide suggestions on the next related word(s) based on the context using prediction algorithm similar to those used by SwiftKey. As the usage is mostly through mobile devices which are resource constrained, the model should be effective in terms of performance and accuracy.

Planned Approach

The predictive model will consume a set of text data through multiple sources to learn various aspects of the language used (eg. style of written language, grammar and contexts) and then predict the next words appropriately while user is typing.

The following tasks have been planned as a guideline to achieve the above listed objective:

- **Task 0: Understanding the problem** *(Identify Data Sources, required Tools and Knowledge)*
- **Task 1: Data acquisition and cleaning** *(Build a Corpus from multiple data sources)*
- **Task 2: Exploratory analysis** *(Understand relationships in source data to build a model)*
- **Task 3: Modeling** *(Build first simple model using relationship between words)*
- **Task 4: Prediction** *(Build first predictive model)*
- **Task 5: Creative exploration** *(Explore to improve model accuracy and efficiency)*
- **Task 6: Build data product** *(Build a predictive text mining app using Shiny)*
- **Task 7: Presentation Slides** *(Model presentation)*

I will be using text mining libraries like 'tm' and 'RWeka' to clean the data and build the corpus. And then filter out profanity words from training data.

About Data

The corpus data used in this report is generated with a freely available data from HC corpora project. The corpora are collected from publicly available sources by a web crawler. The sources include news, blogs, and Twitter.

Listed below are the various data sources used to build the model:

- Capstone Dataset (source: Coursera website)

External Datasets

- Spell correction using aspell or Peter Norvig's method (source: norvig.com)
- Profanity words list for English (source: shutterstock)

Data Exploration

Capstone Dataset This Capstone dataset contains training data covering four different locales: US English (en-US), German (de-DE), Russian (ru-RU), and Finnish (fi-FI). However, only "en-US" locale dataset is being analyzed in this report.

Listed below are some summary stats (lines and word counts) of English dataset.

Blogs

- The data file **en_US.blogs.txt** contains text pulled from multiple internet blogs written by various authors with less degree of commanality in the written text.
- The data file contains a single blog document on each line.

Content Type	Filename	Wordcount	LineCount	Longest Line	Shortest Line
Blog entries	en_US.blogs.txt	38154238 words	899288	40835 characters	1 characters

News

- The data file **en_US.news.txt** contains news stories written by professional journalists.
- The data file contains a single news article on each line.

Content Type	Filename	Wordcount	LineCount	Longest Line	Shortest Line
Blog entries	en_US.news.txt	2693898 words	77259	5760 characters	2 characters

Twitter

- The data file **en_US.twitter.txt** contains text pulled from Twitter posted by diversified authors with less degree of commanality in the written text.
- The data file contains a single tweet on each line.

Content Type	Filename	Wordcount	LineCount	Longest Line	Shortest Line
Blog entries	en_US.twitter.txt	30218166 words	2360148	213 characters	2 characters

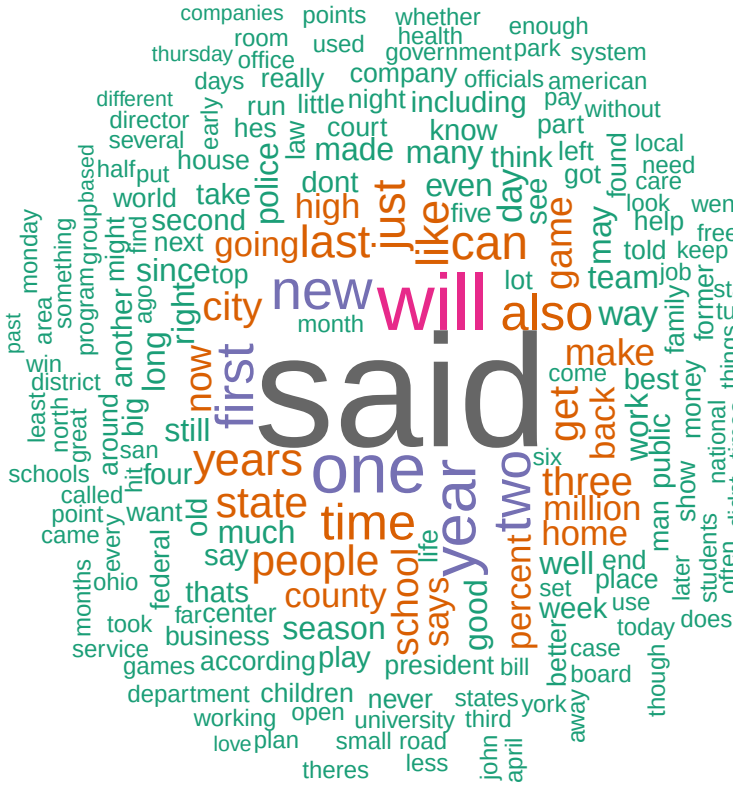
Visualizing individual datasets using wordclouds

A wordcloud (Tag cloud) is a visual representation for text data, typically used to depict keywords or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size

or color. This format is useful for quickly perceiving the most prominent terms and to determine its relative prominence.

Note: As the datasets are large, a small sample(10%) of the each dataset will be shown here for analysis.

Wordcloud of Blogs sample dataset



Wordcloud of News sample dataset

Wordcloud of Twitter sample dataset

Conclusion

- Capstone datasets contains text from diversified authors with less degree of commanality in the written text.
- Twitter feed contains majority of profane words compared to Blogs and News texts.

Overall, Blogs contribute 53.6882276 percent of words, twitter contributes 42.5210897 percent and remaining 3.7906827 percent is from news.