Predictive text Shiny App

Synopsis

Mobile devices like smartphones and tablets are already part of daily life of many people but typing on mobile keyboards can be sometimes frustrating. The goal of this project is to develop a predictive text model like SwiftKey and integrate it on a Shiny App available online to make typing easy and life better for the users.

The key idea of the solution developed here is to apply a well-known best practice of user-experience design to predictive text models: context matters. In other words, the Shiny App will adapt not only to the user as the current smart devices usually do but also to the context in order to better target the prediction and improve typing experience.

Details

The Data

The data for the project is a corpus of three documents containing not formatted English text. This text have been crawled from three different public web sources (Twitter, blogs and news articles) and saved in three different files. The files can be downloaded from here (https://d396qusza40orc.cloudfront.net/dsscapstone/dataset/Coursera-SwiftKey.zip).

Basic summaries

The summaries below show the number of lines, the number of words and some text of example for each of the three files.

- Twitter file

```
## [1] "Lines count: 2360148"
## [1] "Words count: 30359852"
## [1] ""
## [1] "How are you? Btw thanks for the RT. You gonna be in DC anytime soon? Love to see you. Been way, way too long."
## [2] "When you meet someone special... you'll know. Your heart will beat more rapidly and you'll smile for no reason."
```

- Blog file

```
## [1] "Lines count: 899288"
## [1] "Words count: 37334114"
## [1] ""
## [1] "Chad has been awesome with the kids and holding down the fort while I work later than usual! The kids have been busy togeth
er playing Skylander on the XBox together, after Kyan cashed in his $$$ from his piggy bank. He wanted that game so bad and used hi
s gift card from his birthday he has been saving and the money to get it (he never taps into that thing either, that is how we know
he wanted it so bad). We made him count all of his money to make sure that he had enough! It was very cute to watch his reaction wh
en he realized he did! He also does a very good job of letting Lola feel like she is playing too, by letting her switch out the cha
racters! She loves it almost as much as him."
## [2] "With graduation season right around the corner, Nancy has whipped up a fun set to help you out with not only your graduatio
n cards and gifts, but any occasion that brings on a change in one's life. I stamped the images in Memento Tuxedo Black and cut the
m out with circle Nestabilities. I embossed the kraft and red cardstock with TE's new Stars Impressions Plate, which is double side
d and gives you 2 fantastic patterns. You can see how to use the Impressions Plates in this tutorial Taylor created. Just one pass
through your die cut machine using the Embossing Pad Kit is all you need to do - super easy!"
```

- News file

```
## [1] "Lines count: 1010242"
## [1] "Words count: 34365936"
## [1] ""
## [1] "WSU's plans quickly became a hot topic on local online sites. Though most people applauded plans for the new biomedical cen
ter, many deplored the potential loss of the building."

## [2] "The Alaimo Group of Mount Holly was up for a contract last fall to evaluate and suggest improvements to Trenton Water Works
. But campaign finance records released this week show the two employees donated a total of $4,500 to the political action committe
e (PAC) Partners for Progress in early June. Partners for Progress reported it gave more than $10,000 in both direct and in-kind co
ntributions to Mayor Tony Mack in the two weeks leading up to his victory in the mayoral runoff election June 15."
```

Text processing

A random sample has been extracted from each file to make exploratory analysis. After sampling, the text has been cleaned using open source libraries (R packages). Punctuation, numbers, "stop words" (http://en.wikipedia.org/wiki/Stop_words), "bad words" (words with offensive and profane meaning, see profanity filter (http://en.wikipedia.org/wiki/Wordfilter#Removal_of_vulgar_language)) and extra white spaces have been removed. The text has been divided in sentences and words.

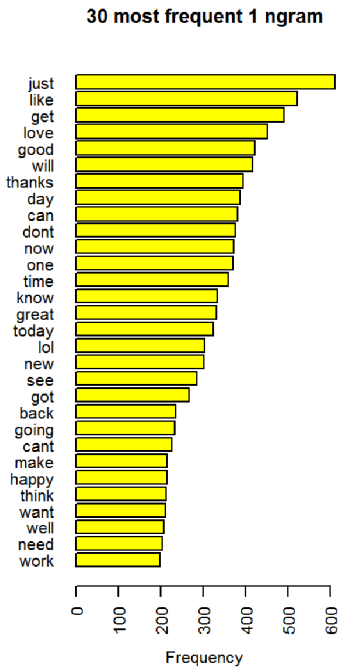
Exploratory analysis

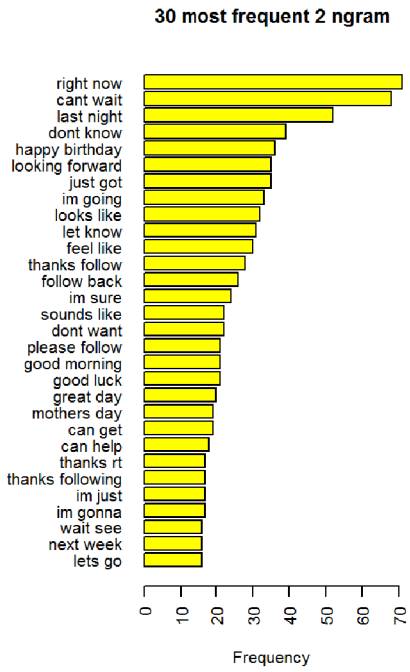
It has been expected there are obvious linguistic differences between the files because their text has been written in completely different contexts: Twitter more informal with more slang and abbreviations, blogs and news more formal with longer sentences and less spelling errors.

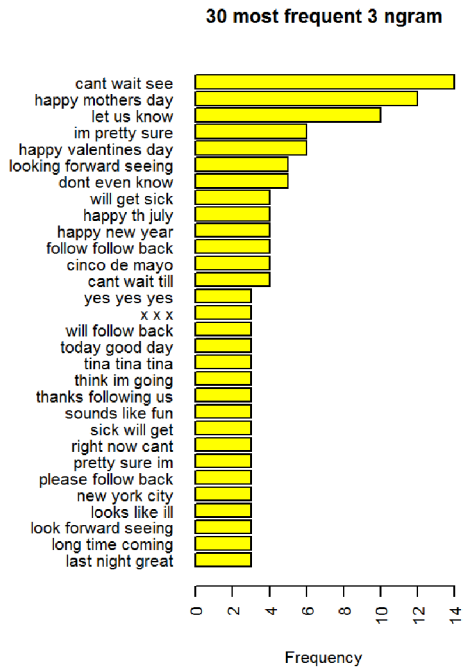
N-gram distributions

From the plots below, it is possible to see the frequency of the most frequent sequences of 1, 2 and 3 words extracted from the three files, called n-grams (for more details about n-gram, see n-gram (http://en.wikipedia.org/wiki/N-gram)). The distributions are different for the three files and the difference is bigger the higher is the number of words in the sequence. The relative summary statistics for n-grams are available in the appendix (Summary statistics).

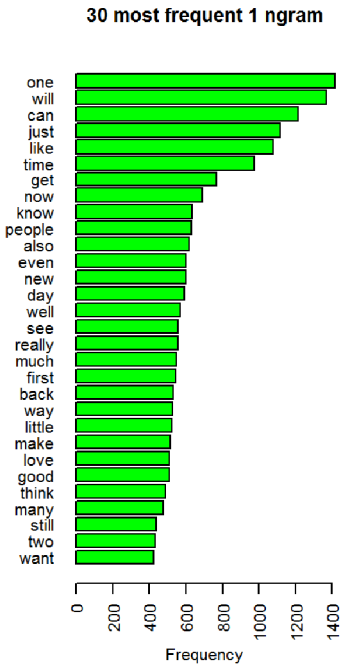
Twitter

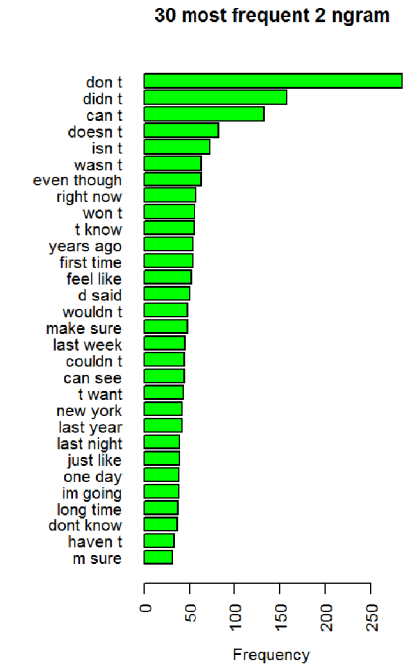


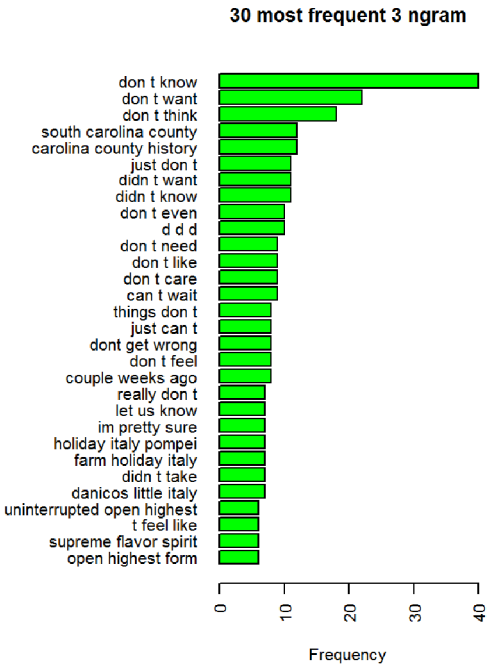




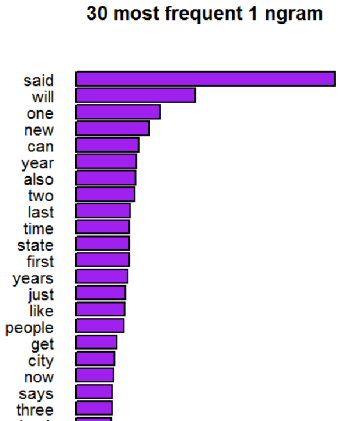
Blog

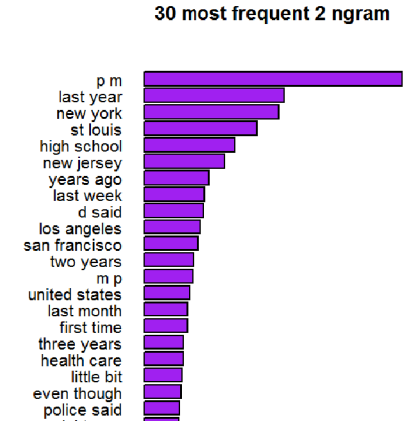


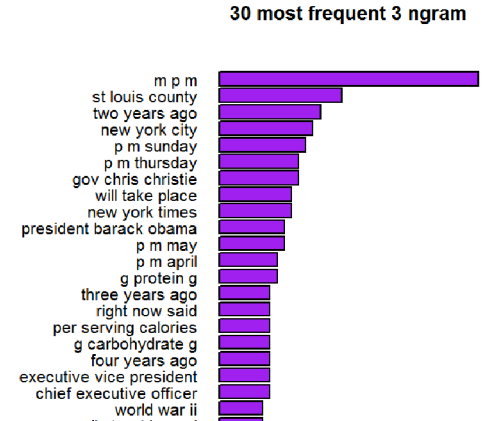


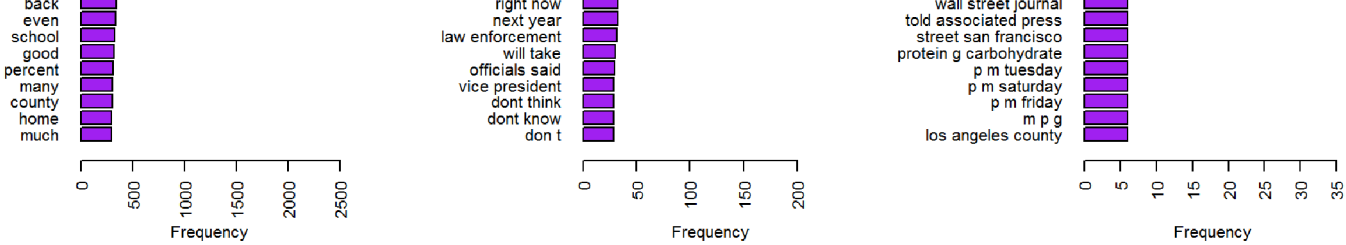


News







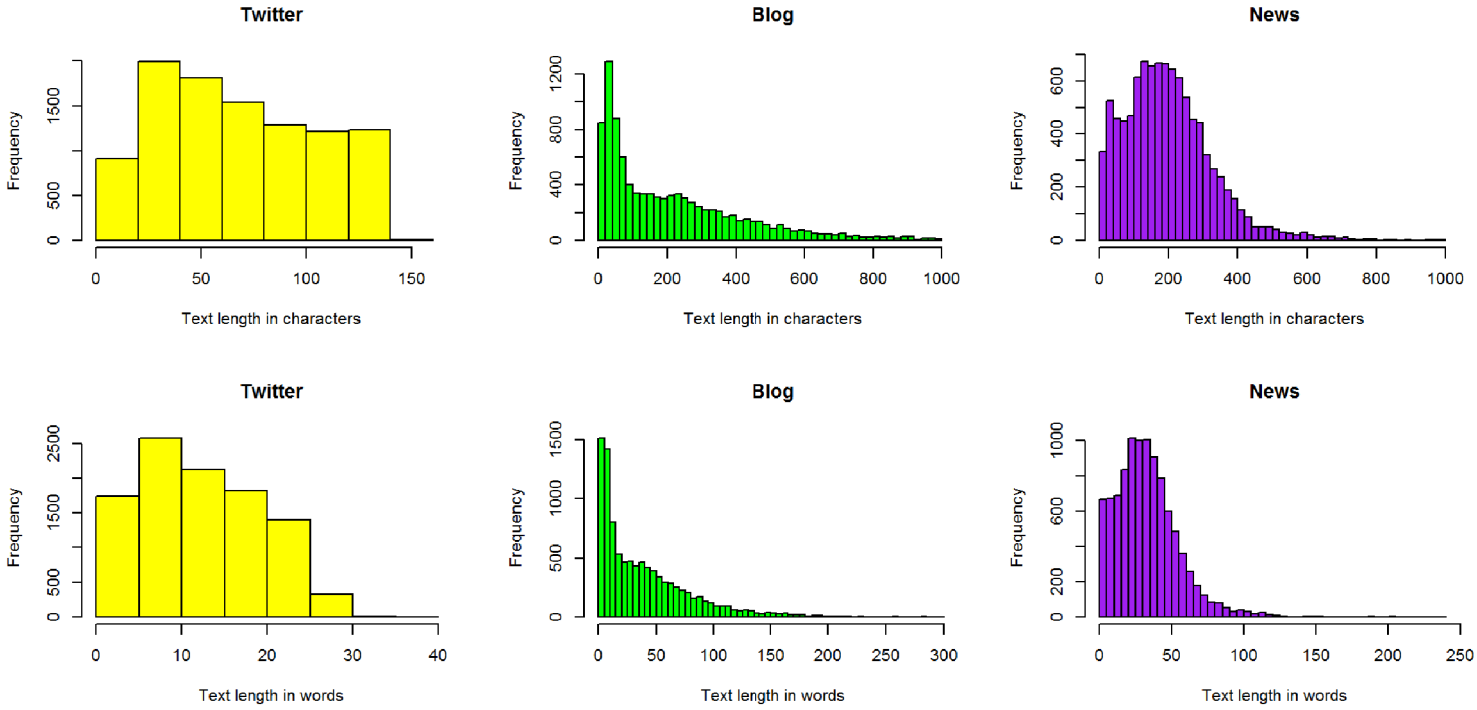


Unigram Clouds

In addition, the word clouds below show clearly the different dictionaries used in the different contexts. To interpret the plot, the following clues have to be applied: words with the same frequency have the same color, the more frequent a word is, the closer to the center of the cloud and the larger the font.

Text length distributions

The histograms below show the distribution of text length measured in number of characters and in number of words for the three files. The distribution are completely different as expected: shorter text with few words for Twitter, longer text with more words for the news and the blogs.



Stylistics analysis

A stylistics analysis has been conducted using a clustering technique (http://en.wikipedia.org/wiki/Cluster_analysis). The analysis has detected the difference between Twitter data and the other two sources grouping them in different clusters, as it is possible to see form the plot in the appendix (Stylistics analysis).

Conclusion

To summarize, context matters. The dictionary and style used are different in different contexts and in most cases also if the user is the same.

Predictive Model and Shiny App

Recent algorithms for natural language processing NLP (http://en.wikipedia.org/wiki/Natural_language_processing) have been based on machine learning (http://en.wikipedia.org/wiki/Machine_learning) techniques. Therefore, the predictive model of this project has been built using the best practices of these techniques (for more details, see machine learning course (https://WWW.coursera.org/course/ml)). The approach can be summarized in the following steps:

1. Usually, data are more important than algorithm and often, "it's not who has the best algorithm that wins. It's who has the most data" [prof Andrew Ng]. Therefore, first prepare a huge data set reaching a trade-off with memory limits. Unlike exploratory analysis, in a text predictive model every word

has its own importance so it is better not to apply any filter like "stop words" filter.

2. Start with a simple algorithm implemented quickly and test it. In this case, the algorithm is based on n-grams backoff techniques (backoff (http://en.wikipedia.org/wiki/Katz%27s_back-off_model)). In addition, the model includes three predictive models, one for each different context.
3. Improve the algorithm applying advanced techniques to refine the prediction and get a better accuracy. The additional techniques chosen are: managing unseen words (smoothing (http://en.wikipedia.org/wiki/N-gram#Smoothing_techniques)), spelling correction, dictionary matching and grammatical disambiguation (part of speech tagging (http://en.wikipedia.org/wiki/Part-of-speech_tagging)).

The Shiny App will give the user a powerful tool to write quicker than usual on a keyboard device. The application implements sentence completion and spell checking services and at the same time gives the user the possibility to switch between different contexts in order to make the algorithm learn faster, adapt better to different typing style of the user and eventually offer a better typing experience.

Appendix

Summary statistics

- Twitter file

```
## [1] "Total number of 1 ngram: 14503"
## [1] "Frequency of the 10 most frequent 1 ngrams"
##   dont    can    day thanks   will   good   love    get   like   just
##    375    381    386    393    417    422    452    490    521    611
## [1] "Total number of 2 ngram: 61609"
## [1] "Frequency of the 10 most frequent 2 ngrams"
##        let know      looks like        im going        just got
##              31              32              33              35
## looking forward  happy birthday       dont know      last night
##              35              36              39              52
##       cant wait       right now
##              68              71
## [1] "Total number of 3 ngram: 64986"
## [1] "Frequency of the 10 most frequent 3 ngrams"
##         happy new year          happy th july          will get sick
##                      4                      4                      4
##         dont even know looking forward seeing   happy valentines day
##                      5                      5                      6
##         im pretty sure            let us know      happy mothers day
##                      6                     10                     12
##          cant wait see
##                     14
```

- Blog file

```
## [1] "Total number of 1 ngram: 30700"
## [1] "Frequency of the 10 most frequent 1 ngrams"
## people   know    now    get   time   like   just    can   will    one
##    629    633    690    768    973   1077   1115   1215   1369   1417
## [1] "Total number of 2 ngram: 191212"
## [1] "Frequency of the 10 most frequent 2 ngrams"
##      t know       won t   right now even though      wasn t       isn t
##          55          56          58          64          64          73
##     doesn t       can t      didn t       don t
##          83         133         158         285
## [1] "Total number of 3 ngram: 212776"
## [1] "Frequency of the 10 most frequent 3 ngrams"
##                   d d d              don t even             didn t know
##                      10                      10                      11
##             didn t want              just don t carolina county history
##                      11                      11                      12
##   south carolina county             don t think             don t want
##                      12                      18                      22
##             don t know
##                      40
```

- News file

```
## [1] "Total number of 1 ngram: 30009"
## [1] "Frequency of the 10 most frequent 1 ngrams"
## time last  two also year  can  new  one will said
##  517  521  564  572  580  602  702  806 1151 2500
## [1] "Total number of 2 ngram: 173089"
```

```
## [1] "Frequency of the 10 most frequent 2 ngrams"
## los angeles      d said   last week   years ago  new jersey high school
##          53          56          57          61          76          85
##    st louis    new york   last year         p m
##         106         126         131         241
## [1] "Total number of 3 ngram: 192120"
## [1] "Frequency of the 10 most frequent 3 ngrams"
## president barack obama         new york times        will take place
##                      9                     10                     10
##     gov chris christie           p m thursday             p m sunday
##                     11                     11                     12
##          new york city          two years ago        st louis county
##                     13                     14                     17
##                  m p m
##                     36
```

Stylistics analysis