# Exploratory Data Analysis

The first step in building a predictive model for text is understanding the distribution and relationship between the words, tokens, and phrases in the text. The goal of this task is to understand the basic relationships you observe in the data and prepare to build your first linguistic models.

Tasks to accomplish

1. **Exploratory analysis** - perform a thorough exploratory analysis of the data, understanding the distribution of words and relationship between the words in the corpora.
2. **Understand frequencies of words and word pairs** - build figures and tables to understand variation in the frequencies of words and word pairs in the data.

Questions to consider

1. Some words are more frequent than others - what are the distributions of word frequencies?
2. What are the frequencies of 2-grams and 3-grams in the dataset?
3. How many unique words do you need in a frequency sorted dictionary to cover 50% of all word instances in the language? 90%?
4. How do you evaluate how many of the words come from foreign languages?
5. Can you think of a way to increase the coverage -- identifying words that may not be in the corpora or using a smaller number of words in the dictionary to cover the same number of phrases?

# Modeling

The goal here is to build your first simple model for the relationship between words. This is the first step in building a predictive text mining application. You will explore simple models and discover more complicated modeling techniques.

*Tasks to accomplish*

1. Build basic n-gram model - using the exploratory analysis you performed, build a basic [n-gram model](http://en.wikipedia.org/wiki/N-gram) for predicting the next word based on the previous 1, 2, or 3 words.
2. Build a model to handle unseen n-grams - in some cases people will want to type a combination of words that does not appear in the corpora. Build a model to handle cases where a particular n-gram isn't observed.

*Questions to consider*

1. How can you efficiently store an n-gram model (think Markov Chains)?
2. How can you use the knowledge about word frequencies to make your model smaller and more efficient?
3. How many parameters do you need (i.e. how big is n in your n-gram model)?
4. Can you think of simple ways to "smooth" the probabilities (think about giving all n-grams a non-zero probability even if they aren't observed in the data) ?
5. How do you evaluate whether your model is any good?
6. How can you use [backoff models](http://en.wikipedia.org/wiki/Katz%27s_back-off_model" \t "_blank) to estimate the probability of unobserved n-grams?

*Hints, tips, and tricks*

As you develop your prediction model, two key aspects that you will have to keep in mind are the size and runtime of the algorithm. These are defined as:

1. **Size**: the amount of memory (physical RAM) required to run the model in R
2. **Runtime**: The amount of time the algorithm takes to make a prediction given the acceptable input

Your goal for this prediction model is to minimize both the size and runtime of the model in order to provide a reasonable experience to the user.

Keep in mind that currently available predictive text models can run on mobile phones, which typically have limited memory and processing power compared to desktop computers. Therefore, you should consider very carefully (1) how much memory is being used by the objects in your workspace; and (2) how much time it is taking to run your model. Ultimately, your model will need to run in a Shiny app that runs on the [shinyapps.io](https://www.shinyapps.io/) server.

*Tips, tricks, and hints*

Here are a few tools that may be of use to you as you work on their algorithm:

* object.size(): this function reports the number of bytes that an R object occupies in memory
* Rprof(): this function runs the profiler in R that can be used to determine where bottlenecks in your function may exist. The profr package (available on CRAN) provides some additional tools for visualizing and summarizing profiling data.
* gc(): this function runs the garbage collector to retrieve unused RAM for R. In the process it tells you how much memory is currently being used by R.

There will likely be a tradeoff that you have to make in between size and runtime. For example, an algorithm that requires a lot of memory, may run faster, while a slower algorithm may require less memory. You will have to find the right balance between the two in order to provide a good experience to the user.

Peer Graded Assignment: Milestone Report

## Instructions

The goal of this project is just to display that you've gotten used to working with the data and that you are on track to create your prediction algorithm. Please submit a report on R Pubs (<http://rpubs.com/>) that explains your exploratory analysis and your goals for the eventual app and algorithm. This document should be concise and explain only the major features of the data you have identified and briefly summarize your plans for creating the prediction algorithm and Shiny app in a way that would be understandable to a non-data scientist manager. You should make use of tables and plots to illustrate important summaries of the data set. The motivation for this project is to: 1. Demonstrate that you've downloaded the data and have successfully loaded it in.2. Create a basic report of summary statistics about the data sets.3. Report any interesting findings that you amassed so far.4. Get feedback on your plans for creating a prediction algorithm and Shiny app.

**Review criteria**

1. Does the link lead to an HTML page describing the exploratory analysis of the training data set?
2. Has the data scientist done basic summaries of the three files? Word counts, line counts and basic data tables?
3. Has the data scientist made basic plots, such as histograms to illustrate features of the data?
4. Was the report written in a brief, concise style, in a way that a non-data scientist manager could appreciate?

# [My submission](https://www.coursera.org/learn/data-science-project/peer/BRX21/milestone-report/submit)

Submit your assignment

Title

Give your project a descriptive title

Please submit the URL of an R Pubs document describing your exploratory analysis (http://rpubs.com/) be sure that the url that you submit is http and not https).