The first step in analyzing any new data set is figuring out: (a) what data you have and (b) what are the standard tools and models used for that type of data. Make sure you have downloaded the data from Coursera before heading for the exercises. This exercise uses the files named LOCALE.blogs.txt where LOCALE is the each of the four locales en_US, de_DE, ru_RU and fi_FI. The data is from a corpus called HC Corpora (www.corpora.heliohost.org). See the readme file at http://www.corpora.heliohost.org/aboutcorpus.html for details on the corpora available. The files have been language filtered but may still contain some foreign text.

In this capstone we will be applying data science in the area of natural language processing. As a first step toward working on this project, you should familiarize yourself with Natural Language Processing, Text Mining, and the associated tools in R. Here are some resources that may be helpful to you.

- Natural language processing Wikipedia page
- Text mining infrastucture in R
- CRAN Task View: Natural Language Processing
- Coursera course on NLP (not in R)

Dataset

This is the training data to get you started that will be the basis for most of the capstone. You must download the data from the Coursera site and not from external websites to start.

- Capstone Dataset

Your original exploration of the data and modeling steps will be performed on this data set. Later in the capstone, if you find additional data sets that may be useful for building your model you may use them.

Tasks to accomplish

1. Obtaining the data - Can you download the data and load/manipulate it in R?
2. Familiarizing yourself with NLP and text mining - Learn about the basics of natural language processing and how it relates to the data science process you have learned

in the Data Science Specialization.

Questions to consider

1. What do the data look like?
2. Where do the data come from?
3. Can you think of any other data sources that might help you in this project?
4. What are the common steps in natural language processing?
5. What are some common issues in the analysis of text data?
6. What is the relationship between NLP and the concepts you have learned in the Specialization?

✓ Complete

👍 👎 ⚑