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. When someone types:

I went to the

the keyboard presents three options for what the next word might be. For example, the three words might be gym, store, restaurant. In this capstone you will work on understanding and building predictive text models like those used by SwiftKey.

This course will start with the basics, analyzing a large corpus of text documents to discover the structure in the data and how words are put together. It will cover cleaning and analyzing text data, then building and sampling from a predictive text model. Finally, you will use the knowledge you gained in data products to build a predictive text product you can show off to your family, friends, and potential employers.

You will use all of the skills you have learned during the Data Science Specialization in this course, but you'll notice that we are tackling a brand new application: analysis of text data and natural language processing. This choice is on purpose. As a practicing data scientist you will be frequently confronted with new data types and problems. A big part of the fun and challenge of being a data scientist is figuring out how to work with these new data types to build data products people love. The capstone will be evaluated based on the following assessments:

1. An introductory quiz to test whether you have downloaded and can manipulate the data.
2. An intermediate R markdown report that describes in plain language, plots, and code your exploratory analysis of the course data set.
3. Two natural language processing quizzes, where you apply your predictive model to real data to check how it is working.
4. A Shiny app that takes as input a phrase (multiple words), one clicks submit, and it predicts the next word.
5. A 5 slide deck created with R presentations pitching your algorithm and app to your boss or investor.

During the capstone you can get support from your fellow students, from us, and from the engineers at SwiftKey. But we really want you to show your independence, creativity, and initiative. We have been incredibly impressed by your performance in the classes up until now and know you can do great things.

We have compiled some basic natural language processing resources below. You are welcome to use these resources or any others you can find while performing this analysis. One thing to keep in mind is that **we do not expect you to become a world's expert in natural language processing**. The point of this capstone is for you to show you can explore a new data type, quickly get up to speed on a new application, and implement a useful model in a reasonable period of time. We think NLP is very cool and depending on your future goals may be worth studying more in-depth, but you can complete this project by using your general knowledge of data science and basic knowledge of NLP.

Here are a few resources that might be good places to start as you tackle this ambitious project.

- Text mining infrastucture in R (https://eventing.coursera.org/api/redirectStrict/8bHSb4_mq2OaXvz-wvndugh-bFKs45_CHN8YLwQI54C0o_MJB-6uX0MnwdTGMsqntnMJInpLyNHfk-ib5sqH7w.KNja5gJDOo0Kvtd88Lx3bA.Gn2_GNdsd3syZPbsx0dJiIyx0rMm9r95D3oDoF8FDiSAc2EQC8opuvg9-2zsHJK8Ji91yTkPgE2UV8bvaA04fEKGxWFudqCvUQWwK0-rOXEo7NB8h7dn5KJrpCU8WBlntKj9e1YHfDcZnqhd592o8dMLp3ixh7UkTyJ5HdEVT8CDq9M2ToIi1MAhR7Q-DqD3f3WvT4OmW6K-lnWxcpwVrWfgHsWophMJF-nYNNoQxsapdTw5kBepXA9EvyQRKH-NwO2G0ff1bD7dXuPuuZDOXUTga7sIBOqcVOvvGnJfm2tZExfuyQFtSJ3l-A9iGw0X)
- CRAN Task View: Natural Language Processing (https://eventing.coursera.org/api/redirectStrict/LHAMhYsxyHplywCePGSWuIaZ2cv98E8jm8uBYROFq9uyHiFU_evPhqFHEJg6ttdMBaYRNH-76OtihLi7KEkrpA.wH3EgYuPFndL0TpAqeu0Rw.BUS2Yd0MTWY1NNjrlAsNP59MbbJiJGuQpS6EOK7f47umewszhbS4qdiRKblZEp1fKgEuo-Q5RwpnO8tl0iCiYea0mHPk87Mev-UfnJvHM5dSWeng3DR63Jp56OuVeJZ24O66oqjCHtCLnXGTOqN9N1JL9r4rPrGRRgnoS9BMecHzZ2FM8qwC9NUdeSEeuyrGZmFVg-aPHn9BDei06-mEM-CPpfZxhOv4t_uxfdO7u8kPzY1PJFTLPmF-3hVzWdSQRJhdPRpD2F8fn6niflNf8r9ADuv0ayG8xGD3ooavo845z61KePK8UO0I5_bas11ciaOuLNH6FR_9xUHjVJYvurNbyo4c6ViAJiyQarloT_GycicfbrXQLJL5rCpXePScP3PszhvmreIbAkqQ1CBYvg)
- Coursera course on NLP (not in R) (https://eventing.coursera.org/api/redirectStrict/CRV0aRPrPLqFnEay-GhYx4V8BF1XveaNmTsHmdfE2QlKH4ddh3Y03nXPIQGL6TiZCxAcGzED484oj32upPqabg.tH3Q5GQvQ7rhqlv3Rp67tg.ynJzU0fKmGwK2RvfibJYmTC5OCz-QFr1TRnf2FYxzofedUEJaUIeI5yPcJ8-Bs6PT-vfj124eIkb-ibL6l1TZfkYJ9PuJ8iiGlOC1IBlRFp8BlrTO8ZIbllW4wBv34jtLFuL_rfV8JHmq0HabnPNmvwyd4ULppG7plrbmVDyzpxNn5lP8XGo4lWaLbhM2esGLo7WnSRNBlnkS9RNBwFAAJ4k8-4ZDUcw7XlkqzC-bQ2C8DPr5OrOA3jeNa39UJV7ZIJSv8opOGDx4cO81q2L-orwLGbp98Uka5WXVj1x0QE)