Data Science Specialization - Capstone Project

Kieso Jan

Coursera Course
This App can do English next words prediction

January 22th, 2022

Version 1.0

About this Project

- This project is part of the 10th course of the Coursera Data Science Specialization, **Data Science Capstone**.
- The deliverable focuses on analyzing a large corpus of text documents to discover the structure in the data.
- Also to figure out how words are put together to build a predictive text model.
- The Project Major Tasks:
  1. **Text data analysis**: analysis of the corpus to understand the relationship of words and word pairs
  2. **Predictive modeling**: build basic n-gram models and develop algorithms to facilitate text prediction
  3. **Shiny app development**: To build up a web-based Shiny app service, which is able to predict next words

Relevant Activities

- **Getting and cleaning the data**: profanity was first removed and words tokenized
- **Exploratory data analysis**: the frequencies of words and word paris were calculated
- **Modeling**: 2-7 gram models were built to facilitate word prediction
- **Prediction model**:
  1. Katz's back-off model was used to predict the next word
  2. The model iterates from 7-gram to 2-gram to find matches in the last n-1 words
  3. In the case of unseen n-gram, the most frequent word, 'the', is returned
  4. To improve efficiency, word pairs that appear less than 5 times in the corpus were removed

Shiny app Function Intro.

- Click **here** to open "Next Words Prediction" Shiny App
- The app takes in the following inputs:
  1. Select how many number of next words to predict.
  2. Typing English Words in textbox
- The predicted next word text will keep changing to show up by sequence of most used to less frequently used

========== **HAPPY PLAY !!** ==========

Reference

Relevant Resources Linkage

- This course is part of the **Coursera Data Science Specialization**
- The **Quanteda** package was used for data analysis and n-gram generation
- Read more about **Katz's back-off model**