SLP Assignment B August 27, 2020

**Introduction to Prediction**

B1. Examine this data in surnames-dev.csv file, on Blackboard, then:

1. Form some intuitive hypotheses about the names of some of the nationalities. (These can be very naive, for example, starts with ***k***, then Russian or two or more ***i***s, then Italian.) Write these down, formally if you like.
2. Write a program to read a one-column file of surnames and output a two-column file of surnames plus predicted nationalities. Don’t be sophisticated yet; it’s fine to directly implement your naive hypotheses. A working template is available as angel-b1.py.
3. Run your program on musician\_surnames.csv (which will become available to you on August 31) and save the results.
4. Submit a report that includes 1) your hypotheses, 2) your program, 3) the results, 4) your thoughts on which two results you are most likely correct, which two are most likely wrong, and why.
5. Be prepared to explain your report, including your code, to the class.

Due September 1. Work mostly individually; submit individual reports. Estimated 2-3 hours.

B2. Extend it.

1. Write a driver to run your predictor for B1 (or an improved version) on a two-column surnames file, and to compute the accuracy of your predictions, and also the precision and recall numbers for the classification of Japanese surnames.
2. Run this new code on surnames-dev.csv, and note the performance.
3. Plan how to modify your code to improve precision, then do it, and record the results.
   1. Intuition: Increase the differences between the classification based on the language, use regular expressions to identify
   2. How? Using regular expressions, with better rules, that more unique to the languages
   3. A picture containing bottle, drawing

      Description automatically generatedResult?
4. Now plan how to modify your code to improve recall, and do the same.
   1. Intuition: Modify rules that seem to repetitive, or open so that the positive-accurate-classifications happen more often.
   2. How? Modify regular expressions that seem to general
   3. Result? A screenshot of a cell phone

      Description automatically generated
5. Write a report summarizing all your plans and all the results, and including your code. Highlight the best results, for inclusion on the leaderboard.
6. Be prepared to explain your report.

Due September 3. Work mostly individually; submit individual reports. Estimated 2 hours.

B3. Build a baseline character unigram model (bag of characters), and learn the weights for a Russian-language predictor. (Russian, since there's more data.) If in python, you’ll find the functions linear\_regression and np.matmul useful.

1. Hand in your code.
2. Report the weights it learned on the training data.
3. Report the accuracy and precision on both the training data and the test data in surnames-test.csv (available next week).
4. Find one non-Russian name that was misclassified as Russian, explain why the error happened, and how it could be fixed.
5. Conversely, discuss one Russian name that was misclassified as non-Russian.

Due September 8. Work in pairs. Estimated 2 hours per person.