# coding: utf-8

# The above is needed to remove the two special

# Non-ascii characters I discovered in our song

# lyric text files.

import re, string

with open('all\_lyrics.txt', "r") as file:

text = file.read()

text = text.lower()

# Get rid of all punctuation with the exception of apostrophies. County needs it's lovin', hootin' & hollerin'

RemovePunc = "[" + re.sub("'","", string.punctuation) + "]"

text = re.sub(RemovePunc, "", text)

# Just some instances of odd quotation marks I noticed in one of our songs; they're gone now

text = re.sub("”", "", text)

text = re.sub("“", "", text)

# Removal of unnecessary text that may have accidentally been recorded

# by our web-crawler. The presence of any legitimate uses of these words

# or phrases is outweighed by their use in song formatting, explained more

# in the preliminary results section of our final report.

text = re.sub("repeat chorus", "", text)

text = re.sub("chorus", "", text)

text = re.sub("instrumental break", "", text)

text = re.sub("instrumental", "", text)

text = re.sub("miscellaneous", "", text)

# Some of what our Webparser recovered was simply these two

# lines, for some reason. This removes them.

text = re.sub("original source", "", text)

text = re.sub("lyricwiki", "", text)

file.close()

with open(‘all\_lyrics.txt', "w") as file:

file.truncate(0)

file.write(text)

file.close()