Брать текст из файла .txt (.doc/.docx)

Брать вероятности появления символов из таблицы для большого объема знаков.

Подумать про языки

Abc -> [‘a’:0] and [‘b’:1,’c’:1]