Министерство образования и науки Российской Федерации

ФГАОУ ВО «Северо-Восточный федеральный университет имени М.К. Аммосова»

Институт математики и информатики

Кафедра информационных технологий

ОТЧЕТ

ПО НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

Магистранта 2 курса группы МАГ-ИВТ-16

Малгараевой Айгуль Жумабековны

Тема магистерской диссертации: «Применение машинного обучения для морфологического анализа слов якутского языка»

Место прохождения практики:

Кафедра Информационных Технологий ИМИ СВФУ

Рекомендуемая оценка \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Научный руководитель:

Павлов Александр Викторович, доцент кафедры ИТ \_\_\_\_\_\_\_\_\_\_\_\_

(подпись)

Якутск, 2018

Оглавление

**Элементы оглавления не найдены.**

# Введение

Якутский язык используется во многих средствах массовой информации, различных исторических хранилищах данных, такие как энциклопедии, учебники, художественная литература.

Развитие информационных технологий открывает перед нами широкие возможности в изучении данных ресурсов, однако, в повседневном общении на якутском языке, в различные социальных сетях, мессенджерах мы продолжанем испытывать трудности из-за малого количества специальных инструментов, таких, как клавиатуры на родном языке. Из-за этого возникает потребность замены якутских букв, к примеру, в слове оҕолор - ‘дети’, часто вместо буквы ҕ используется цифра “5” или в слове үөрүү ‘радость’ происходит замена букв ү,ө на у,о, что приводит не только к неправильному написанию слова, но также меняет его смысл.

# Статистико-комбинаторный метод моделирования языка

Статистико-комбинаторный подход моделирования языка был описан Н.Д. Андреевым. ("Статистико-комбинаторный методы в теоретическом и прикладном языковедении", М., 1967) Эксперименты по применению статистико-комбинаторной модели были проведены для различных корпусов языков, в том числе английского, французского и русского. Статистико-комбинаторное моделирование языка можно описать двумя этапами:

1) опираясь на статистику и комбинаторику, при этом не обращаясь к критерию правильности грамматики, сформировать систему языковых форм

2) затем, учитывая смысл текста, установить значения полученных форм и выявить правильность грамматики.

Работы Н.Д. Андреева были найдены, отреферированы и реализованы в виде программы на языке программирования Python (Unsupervised learning of morphology in the USSR, Franck Burlot, François Yvon, 2016)

Мы пытаемся реализовать данный алгоритм для якутского языка.  
В качестве корпуса языка были использованы статьи из газет на якутском языке, такие как «Кыым».

## Считывание статистики по корпусу якутского языка

Подготовка корпуса

Для начала составляем словарь, который содержит частоты слов в следующем виде {слово: число вхождений в корпус, ... }, например {"көппөҕү" : 4, "хазар" : 3, ...} и записываем их в словарь в json формат, для быстроты обработки. В итоге было загружено 386233 словоформ, 11.9 млн словоупотреблений.

Затем из полученного словаря подсчитываем независимые частоты каждого

символа, в виде {буква: число вхождений в корпус, ... }, а также общее

количество букв в тексте. В итоге получили:

{'а': 11101459, 'б': 2901561, 'в': 427404, 'г': 1311975, 'д': 1900403, 'е': 778639, 'ж': 37901, 'з': 91125, 'и': 5474794, 'й': 1316593, 'к': 2871767, 'л': 4443816, 'м': 1332271,'н': 6053266, 'о': 4046042, 'п': 914556, 'р': 5889206, 'с': 3247664, 'т': 6289053, 'у': 4051425, 'ф': 88899, 'х': 1971611, 'ц': 98852, 'ч': 637310, 'ш': 34279, 'щ': 13392, 'ъ': 6759, 'ы': 4767439, 'ь': 679475, 'э': 5704321,'ю': 29007, 'я': 166433, 'ё': 1635, 'ҕ': 800062, 'ҥ': 490070, 'ү': 1826789, 'һ': 1229530, 'ө': 1237080}), 84263863),

где 84263863 – общее количество букв.

Подсчитываем независимую вероятность букв в тексте - вероятность

букв, независимая от их расположения по позициям в слове.

Для вычисления независимых вероятностей букв делим количество

вхождения каждой буквы в текст на общее количество букв в тексте и

представим полученные данные в виде графика (рис. 1.):

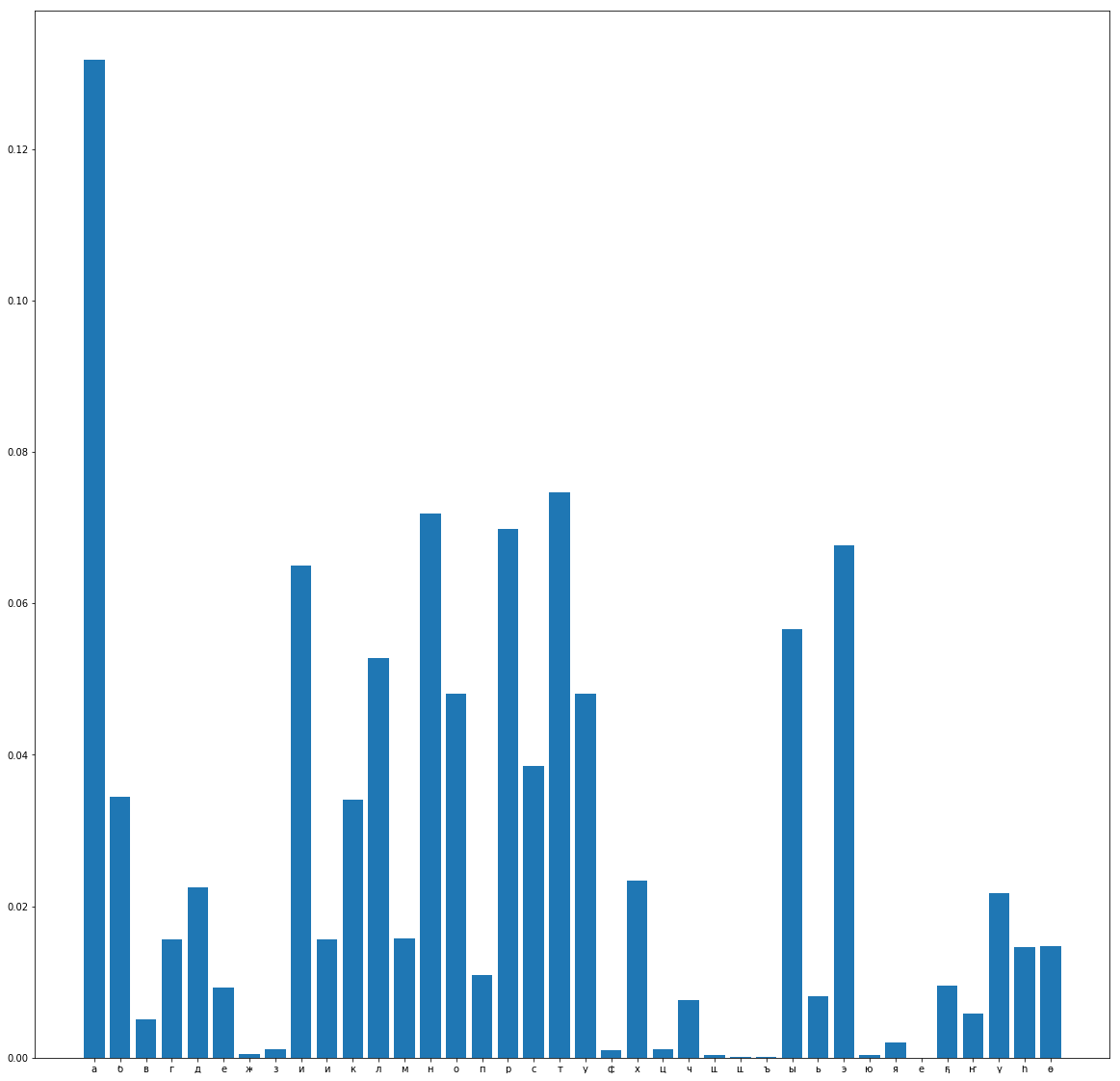


Рис. 1.

Следующим шагом является вычисление средней длины слова, получаем значение 7.1053179762290535, можем округлить это значение и считать длину равной 7.

Предыдущие вычисления проводились для получения условной вероятности буквы – вероятности буквы, зависящей от