**My Speech**

It’s been centuries that several researchers have been making lots of investigations about the morphological and lexical analysis in different languages such as English, Russian and etc. For creating our own algorithm, we had to make huge amount of research to understand the created applications regarding to this. Firstly, we made research about stemming algorithms and those which were created in English. Stemming is one of the most popular ways that has been used in linguistics by reducing the words into their initial form which is also called base form. As an example, we can choose “arguing”, “argues”, “argued” which the initial form is “argue”. There were lots of algorithms, but we chose three best ones to investigate. Lovins algorithm is one of the oldest and popular algorithms that has been created for English language and the main idea is that there’s a table for all the suffixes and prefixes and for usage of this algorithms the longest suffix should be deleted. Another algorithm is Porters algorithm and it’s similar to Lovins; however, in this algorithm suffixes are made by smaller suffixes. For instance

EED – EE

Levenshtein algorithm is another algorithm and we actually used this algorithm for our research. Informally, the Levenshtein distance between two words is the minimum number of single-character edits required to change one into another (kitten-sitten, sitten-sittin, sittin-sitting). It is zero when the strings are equal, so you might guess it is as same as strcmp.

Another topic that I would like to talk is data cleaning that we have done during this timeline. Data cleaning is the way of finding and deleting unnecessary and inaccurate data from the dataset. So, you might ask why did we need this? Well, we needed data cleaning because of statistical analysis that we are going to explain now. We have made statistical analysis by cleaning our dataset and diving the sentences into words and checking to which part of speech they belong. Due to having lack of resources, we acquired the data from 85 thousand Azerbaijani news and made data cleaning which brought us more than 3 million sentences without duplicates. After acquiring the data, we made statistical analysis about the implementation of words and we got this statistics. As you might see from the graph, we got more than 13 million words within this data and 6 million of them are nouns, 1 million of them are adjectives and verbs respectively and others.