NLU-Assignment1 Report

Akash Mishra
SR No- 14421
akashmishra@iisc.ac.in

Abstract

This document contains the report of Assignment1 in which we have to fit a language model over given datasets and then check the perplexity of models.

1 Introduction

There are two datasets given-
1. Brown Corpus
2. Gutenberg Corpus

First we need to devide the given datasets into train, developement and test part. First we need to fit our model using training dataset, then tune hyperparameter(if any) using developement dataset and finally test and check the accuracy of model using test dataset. We have been given four different settings to implement and evaluate our model-
1.)Train over Brown Corpus and test over Brown Corpus
2.)Train over Gutenberg Corpus and test over Gutenberg Corpus
3.)Train over combined corpus(Brown+Gutenberg) and test over Brown Corpus
4.)Train over combined corpus(Brown+Gutenberg) and test over Gutenberg Corpus

2 Results

- I have used perplexity of a model as a metric for comparison.
- For calculating perplexity, I have used two basic algorithms-
  - Simple Backoff algorithm
  - Katzs Backoff algorithm
- Both algorithms have been applied over only bigram model
- for Simple Backoff Algorithm, I have devided given dataset into 8:2 training and testing dataset
- for Katz's Backoff Algorithm, I have devided data into 8:1:1 training, development and testing dataset

S1 : train = Brown , test = Brown	
Simple Backoff Algorithm	205.180663798
Katz's Backoff Algorithm	228.9162561742418
S2 : train = Gutenberg , test = Gutenberg	
Simple Backoff Algorithm	110.107190282
Katz's Backoff Algorithm	116.4113297346987
S3 : train = Brown+Gutenberg , test = Brown	
Simple Backoff Algorithm	252.251452239
Katz's Backoff Algorithm	272.9773085668053
S4 : train = Brown+Gutenberg , test = Gutenberg	
Simple Backoff Algorithm	115.650172752
Katz's Backoff Algorithm	117.89298781323258