GIT Department of Computer Engineering
CSE 603
Spring 2016

Homework 03
Due date: May 23rd 2016
N-Grams

N-grams has many advantages for modeling languages as we have seen in the classes. However, for Turkish, morphological analysis of the words makes the word counts very problematic. To address this problem in a simplistic way, we will use this approach: take the first k characters of the word as the word itself and calculate the N-Grams using this definition of the word. For example, for the word “geliyor” we will take only “gel” if k is 3.

We will change the number k in our experiments but usually it will be between 3-8.

1. Download the Turkish news set from http://www.kemik.yildiz.edu.tr/data/File/1150haber.rar
2. Calculate the 1-Gram, 2-Gram and 3-Gram tables for this set using k=3, k=4, and k=8 values. You will do this calculations on 95% of the news set.
3. Calculate perplexity of the 1-Gram, 2-Gram and 3-Grams with different k values on the remaining 5% of the news set. Make a table of your findings.
4. Write a program that assigns probabilities for a given Turkish sentence using the techniques we learned in the class. One sample output of the program would be

```
N-Grams: 2
value of k: 4
Enter sentence: Hava çok güzel
The probability of the sentence is: 0.00053
```

Write your report that contains sample runs of your system and table for perplexities. Which values of N-grams and k values models the language better for Turkish?