GIT Department of Computer Engineering
CSE 554/484
Spring 2017

Homework 01
Due date: Apr 16th 2017

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. For this reason we will use n-grams of the characters instead of the words or word stems for this homework. In other words, the 1-gram value of character “a” will tell us how many times the letter “a” occurred in our training set.

You will demo your final results and you will prepare a report that shows your findings by following these steps:

1. Download the Turkish news set from http://www.kemik.yildiz.edu.tr/data/File/1150haber.rar
2. Calculate the 1-Gram, 2-Gram, 3-Gram, 4-Gram and 5-Gram tables for this set using 95% of the news set.
3. Calculate perplexity of the 1-Gram to 5-Gram models using the chain rule with the Markov assumption for each sentence. You will use the remaining 5% of the news set. Make a table of your findings in your report.
4. Implement also the simple interpolation technique that we learned in the class. The example below interpolates from 3-Gram to 1-Gram. You will do this for 5-Gram to 1-Gram. Select the best lambda values empirically. Put the perplexity results in your report table with the interpolation method and show the lambda values.

$$\begin{aligned}\hat{P}(w_n|w_{n-1}w_{n-2}) &= \lambda_1 P(w_n|w_{n-1}w_{n-2}) \\ &\quad +\lambda_2 P(w_n|w_{n-1}) \\ &\quad +\lambda_3 P(w_n)\end{aligned} \qquad \sum_i \lambda_i = 1$$

5. Write a program that assigns probabilities for a given Turkish sentence using the techniques we learned in the class. Note that during the demos, you should not calculate the N-gram values from the training set because you should prepare them in advance. Some sample outputs of the program would be

Use interpolation: N
N-gram: 4
Enter sentence: Hava çok güzel
The probability of the sentence is: 0.00053

Another run:

Use interpolation: Y
Enter sentence: Hava çok güzel
The probability of the sentence is: 0.000067

Notes:

- Submit your HW to the moodle page. Your submission should include
  - The source code
  - The report
  - The data files for your trained n-grams
- During your demo, you will download your submission from the moodle page, compile it and run the demo code. You will not use any development environments during the compilation or the demo. All the tasks should be done in command line shell either in UNIX or in Windows.
- Your homework will not be graded if you do not make a demo.