Aditya Agashe (ava9)
Rohit Biswas (rb625)
John Hughes (jlh457) Due: 9/12/16

CS 4740 Project 1: Part 1 Analysis

Introduction

For Part 1 of Project 1, we had to make an unsmoothed unigram and bigram model to generate random sentences based on a given dataset.

Preprocessing

We processed the following data from the text files since we assumed that it was irrelevant text or that it didn't add value/meaning to the language model:

- Email headers ("From : ", "To : ", "Subject : ", "Re : ", "writes : ", wrote : ")
- Special characters ("][\^_`(){*+-#\$%&/<=>@|~<>\|\\")
- Email addresses ('[\w\.-]+@[\w\.-]+')

Next we turned the text into word tokens (we used a library called NLTK for this) and ran the unigram or bigram random sentence generation models. Note that in our model we distinguished words based on capitalization ("hi" and "Hi" were treated as different word tokens).

Unigram Random sentence generation

To generate the unigram random sentences, we started by creating a list of the cumulative probabilities of each of the word tokens. We then generated a random number and used it to select the next word in a sentence. When we hit an ending character (".!?;"), the sentence was returned. Below you will find some of the sentences that our model generated using the "motorcycles" corpus:

- "jerk dont pinkie , fuel Aug the i THE 's knowledge escape anyone GS the Cellular to the the read DoD is miracle Thomas Power , this Honda needed backpack Karras for 600F2 there previous holding Reef am of I ."
- "the illustration complying waves Bruiser it Wheelies ."
- "readers motorcycle also to EVER corner side going n't turn you wonder you June in in go that flywheels and watch but big 14 on bloody SCMA ."
- "This years every the lo of Chief a DoD helmet been , a Neurourology the News beam seen guess ."
- "everything still experience , want advantage to off be Rogers fill Thanks 's that ."

Aditya Agashe (ava9)
Rohit Biswas (rb625)
John Hughes (jlh457) Due: 9/12/16

Bigram Random sentence generation

To generate the bigram random sentences, we preprocessed the text to add a beginning marker (“|”) at the start of the list of word tokens. Then we parsed through the list of word tokens and after every ending character (".!?;") we added the marker (“|”). Therefore, each sentence started with a “|” character and used the bigram model (based on the conditional probabilities given the previous word) to randomly select the next word. When we hit an ending character (".!?;"), the sentence was returned. At the end we removed the “|” character from the beginning and end of the sentence (they were added by the model for sentence generation, and were not actually part of the sentence). Below you will find some of the sentences that our model generated using the “motorcycles” corpus:

- “Now why would be annual buyer 's testimony that the time I deleted Any lock including substantial lubrication when I console myself with the name tod johnson jumps to the newbies are located .”
- “the ' 69 Impala convertible The bike paint jobs and a mirror while I know all was essentially a car to be a lot of accidents that you , Contractor Large .”
- “A promise to hit 30 MC Power Arc II , is the dog suddenly appeared and go undetected for these three times , since the ground on a buddy 's far I do n't have the limits of opinion that story ... Nigel Tufnel , consider police officers read here 's better than the perp .”
- “Whacha mean you pointed me , Every nerve aware Red Barchetta Straining the scale with it proudly displayed again .”
- “Lets just try my heels to go to Chicago by the Milwaukee machine ;”

Team Contributions

Aditya, Rohit, and John worked together to plan, develop, design, code (preprocessing of text, read text, unigram model, unigram random sentence generation, bigram model, bigram random sentence generation), and test the unsmoothed unigram and bigram random sentence generation models and write the report.