Divyansha Sehgal

Rachel Kwak

Project 1: Part 1

The preprocessing decisions we made are as follows: 1) we decided that each line in the training data files would count as a sentence 2) since the tokenization was already done in the training files, we decided not to take any further steps in tokenizing besides splitting the words by spaces 3) to make a better distinction of where each sentence starts, we decided to put in start tokens. We decided that putting end tokens would be redundant because all the sentences in the training data ended with periods.

Our sentences generated by the unigram model was set to have a length greater than 5 and to not start with a punctuation. We allowed capitalized words in the sentences.

Here are examples of random sentences generated by our unigram model:

1) see class in The Allows and to The state easygoing the -- you to filmmakers is .
2) Just but dumb is the never The on 's At It homosexual , Dead Solaris objective to and its rueful movie suspense originality post-production stands apply but the ; asks Carvey point cable a It major film .
3) amazing even none Marquis due a bone-crushing structure together ... its referential drama .

Here are examples of random sentences generated by our bigram model:

1) Michael Gerbosi 's great laughs , and epic battle scenes .
2) The most heinous crime should have more objective measurements it 's birth .
3) Yakusho and increasing weariness as all odds in sensibility and romance and outer -- and some wonderfully weaves this `` Sade .