Quiz1 Allen Herrera CS325

Task

- Download the unigram and bigram data. ☑
- Complete the following code to estimate P(w[0], ..., w[6]). ☑
- Submit a short report explaining how you derive this estimation. ☑
- References: IOUtils. ☑

Solution

```
String[] w = {"he","came","to","my","school","to","study"};
```

Chain Rule

P(w[0], ..., w[6])= P(w[6]|w[0],w[1],w[2],w[3],w[4],w[5]) * P(w[5]|w[0], ..., w[4]) * etc…

Markov Assumption Method ☑

P(w[0], ..., w[6])= P(w[0]) * P(w[1]|w[0])* P(w[2]|w[1])* P(w[3]|w[2])* P(w[4]|w[3]) * P(w[5]|w[4])* P(w[6]|w[5])

P(w[0]) is simply a unigram probability of the word “he”. Derived by counting the number of times the word he appeared in our sample (the .txt files) and divided by the total number of words observed. P(w[0]) = 0.0014331108804395238 or 0.14%

P(w[1]|w[0]) is a bigram as P(w[1]) “came” is dependent on the P(w[0]) “he” appearing right before it. This is consistent with the chain rule so far but will diverge into the markov assumption in the next step as we disregard w[0] when calc w[2]. P(w[1]|w[0]) = 0.004758864772863373 or 0.48% the total probability so far is 6.819980884530863E-6

P(w[2]|w[1]) = 0.2169154509097054 or 21.69% and total prob is 1.4793592287635838E-6

P(w[3]|w[2]) = 0.0036468028849340603 or 0.36% and total prob is 5.394931503308864E-9

P(w[4]|w[3]) = 7.9365518615475E-4 or 0.08% and total prob is 4.281715366550761E-12

P(w[5]|w[4]) = 0.018234547310248688 or 1.8% and total prob is 7.807514142038865E-14

P(w[6]|w[5]) = 8.853309990925741E-4 or 0.09% and total prob is 6.91223429580067E-17

Therefor the P(w[0], ..., w[6])= 6.91223429580067E-17 or .0000000000000006912%

TLDR: P(w[0], ..., w[6])= was estimated via markov’s assumption and bigrams, my commented test.java file is available in my quiz1 directory should you want to look at that.