Lab 2: You Can't Sit With Us

September 21st 2017

In this lab you'll use N-grams to perform both classification and generation against a sample of works from three famous authors.

The Situation

Authors have a tendency to dial the pettiness up to volume 11. In the local cafeteria, the authors only want to sit with other authors who have a similar writing style to their own. I know. PETTY. Your job in this assignment is to facilitate the pettiness of one such author. He has two authors who want to sit with him, and he needs to decide which one to accept as his lunchtime conversation.

Part 1: 40 points

Choose three authors who have books available online at Project Gutenberg (link is posted on piazza). Pick one of them to be the author who needs to select a lunchmate (the chooser). The other two authors will be the candidates. You can choose these authors randomly or specifically choose authors who you think of as having dissimilar/similar styles. Or authors that you like. Or that have funny names. Or whatever.

Using n-gram language models as your basis of reasoning, argue the case for who should be the chooser's lunchmate. Obviously the model's you construct will yield an overall prediction, but you should also dig into the details a bit and describe some of the specific n-grams that contributed to your prediction.

Part 2: 20 points

Karma is a bitch. The author who got left out of the lunch now gets to be the chooser. Of the two authors who excluded him, who should he choose to sit with.

Part 3: 40 points

Using the language models you created, generate two sample conversations that might be taking place at each of these lunches. These conversations probably won't make much sense. How might you alter your n-gram models to get your authors closer to the same page?

Step 7

Yes, I am always going to call this Step 7. Your main deliverable here is a short (~2 page) report on the authors you selected, the analytics you performed, and an argument for who these analytics recommend as lunchmates. Your report should also include the resulting conversations between the authors. In a single zip file, include your report (a pdf), as well as your code. Submit this using the command:

provide comp150nlp prose prose.zip