Name: _____
CS159: Natural Language Processing
PReview #7
Due in class Mar 30, 2015

Review

Suppose we want to build a language model off of the following training text (also available electronically):

> The sun did not shine. It was too wet to play. So we sat in the house All that cold, cold, wet day. I sat there with Sally, we sat there we two. And I said, "How I wish we had something to do!" Too wet to go out and too cold to play ball. So we sat in the house. We did nothing at all. So all we could do was to Sit! Sit! Sit! Sit! And we did not like it. Not one little bit. And then something went BUMP! How that bump made us jump! We looked! Then we saw him step in on the mat! We looked! And we saw him! The Cat in the Hat! And he said to us, "Why do you sit there like that?" "I know it is wet And the sun is not sunny. But we can have lots of good fun that is funny!" "I know some good games we could play," said the cat. "I know some new tricks," said the Cat in the Hat. "A lot of good tricks. I will show them to you. Your mother will not mind at all if I do."

Assume, too, that we pre-process the text by converting everything to lowercase and removing all punctuation (after we use it to separate sentences, of course!)

Give the counts of the bigrams "the cat," "with Sally," and "we said" under the following conditions:
1. Un-smoothed (MLE) count
2. Add-one smoothing
3. Add-lambda smoothing, with lambda = 0.1
4. Witten-Bell smoothing

■

Describe the data structures that your group is using for the language model and translation model for project 2.

■

Preview

Read about at least one of the smoothing techniques (other than Witten-Bell) in the Chen and Goodman paper linked from Project 2. Briefly describe the intuition behind that method, and how it modifies language model counts.

■

Read Sections 1 and 2 of this paper on Pharaoh. List three questions you have after reading it.

1.

2.

3.

■