CSCI 375 HW1

Haoyu Sheng

September 2017

1 Written Exercises

1.1 Problem 1

Note that there are 24 words in total and the word ”green” has never appeared. Thus:
P($< s >$ I do not like green eggs and Sam $< /s >$)$= P(< s >) * P(I| < s >) * P(do| < s > I) * P(not| < s > Ido)... * P(< /s > |IdonotlikegreeneggsandSam)$ which cannot by computed due to division by 0.

1.2 Problem 2

Note that the word ”green” has never appeared. Thus:
P($< s >$ I do not like green eggs and Sam $< /s >$)$= P(< s >) * P(I| < s >) * P(do|I) * P(not|do) * ... * P(< /s > |Sam) = 1 * \frac{3}{4} * \frac{1}{4} * \frac{1}{1} * \frac{1}{1} * \frac{0}{1} * \frac{1}{0} * \frac{1}{1} * \frac{1}{1} * \frac{3}{4}$, which is undefined

1.3 Problem 3

Note that though all words appeared in the corpus, the bigram ”do like” and ”eggs $< /s >$” have never appeared.
Without Smoothing:
P($< s >$ I do like eggs$< /s >$) $= 1 * \frac{3}{4} * \frac{1}{4} * \frac{0}{1} * \frac{1}{1} * \frac{0}{1} = 0$

With add-1 Smoothing:
P($< s >$ I do like eggs$< /s >$) $= \frac{4}{14} * \frac{2}{14} * \frac{1}{11} * \frac{2}{11} * \frac{1}{11} = \frac{16}{260876}$

Question 2.1:

```
('Sam </s>', 4)
('</s> <s>', 4)
('<s> I', 4)
('I am', 4)
('am Sam', 3)
('Sam I', 2)
('do not', 2)
('am </s>', 2)
('not like', 2)
('like eggs', 2)
```

Question 2.2:

```
('Sam </s>', -1.252762968495368)
('</s> <s>', -1.252762968495368)
('<s> I', -1.252762968495368)
('I am', -1.252762968495368)
('am Sam', -1.466337068793427)
('do not', -1.7047480922384253)
('not like', -1.7047480922384253)
('like eggs', -1.7047480922384253)
('eggs and', -1.7047480922384253)
('and Sam', -1.7047480922384253)
```

Question 2.3:

```
('</s> <s>', -1.405533083755584)
('. </s>', -1.5188142128976665)
('. .', -3.212593273765518)
('of the', -3.3273099531494004)
('<s> the', -3.3751022708793035)
(', but', -3.6610290563215995)
('<s> a', -3.683525827637901)
(', and', -3.7105932143377203)
('in the', -3.883611319477772)
('is a', -4.058345505807142)
(', the', -4.253080120010176)
('the film', -4.330191107753882)
('of a', -4.37209759180373)
('to the', -4.409295662789433)
('to be', -4.458566711796216)
('and the', -4.506895616719751)
('in a', -4.533210254791982)
("<s> it's", -4.538839763013341)
```

```
('it is', -4.659113560580396)
('the movie', -4.670415693993646)
```

Question 2.4:

```
Generated by 1-gram
<s> up supposing that degree my , was to the wished in to had expecting case what admire his
right but </s>
Generated by 2-gram
<s> `` advertising Do partner Email filette Nonsense mumbling gulf containing risks extent
affirmations cakes recipient repressing Left establishment head-stones says </s>
Generated by 3-gram
<s> I was accurate You discriminated glamour instead publication voluntarily enunciated UT
harbourage walking strenuously redistribution sprinkled ghost Ghosts lulled carrying </s>
Generated by 4-gram
<s> `` Yes , midsummer choice credulity curled orange lectures smelt DAMAGES kidnapping
inhabitant When excellent kin proves richly prevent hall-front </s>
Generated by 5-gram
<s> `` I am an clattering apothecary intrust ceases prisoned grown place. appanage wave
pencil-head dish antipodes breadth Jewels clever deliberated </s>
```

The sentences generated all seem to have similar randomness and a lack of structure and almost none of the models was able to reach the end of a sentence before a threshold of 20 words. This is due to the suboptimal nature of add-one smoothing that takes away too much probability from sensical n-grams and assign them to arbitrary combination of words. Also, as n increases, the number of unique n-grams conditional to n-1 words is getting smaller and smaller, and add-one smoothing is extremely detrimental in cases like this since it then assigns 1/n+V probability to all possible combinations.