# NGRAMS – PART 2

## NATURAL LANGUAGE PROCESSING - CS 322.00

Blake Howald

September 27, 2017

## AGENDA

- Presentation – Maryam Hedayati, Sasha Mayn
- Questions
- Ngrams – Part 2
  - Smoothing
  - Evaluation
    - Training/Testing
    - Perplexity

# SMOOTHING

- Multiple methods to address data sparsity (also known as discounting).
- Redistributing probability mass to unseen data
  - Laplace (add one) smoothing
    - Also add *k* smoothing
  - Good-Turing discounting
    - Kneser-Nay and other "absolute" discounting methods

# LAPLACE (ADD ONE) SMOOTHING

***The cat in the hat***

| | The | Cat | In | Hat |
|---|---|---|---|---|
| The | 0 | 1 | 0 | 1 |
| Cat | 0 | 0 | 1 | 0 |
| In | 1 | 0 | 0 | 0 |
| Hat | 0 | 0 | 0 | 0 |

| | The | Cat | In | Hat |
|---|---|---|---|---|
| The | 0 | .25 | 0 | .25 |
| Cat | 0 | 0 | .25 | 0 |
| In | .25 | 0 | 0 | 0 |
| Hat | 0 | 0 | 0 | 0 |

| | The | Cat | In | Hat |
|---|---|---|---|---|
| The | 1 | 2 | 1 | 2 |
| Cat | 1 | 1 | 2 | 1 |
| In | 2 | 1 | 1 | 1 |
| Hat | 1 | 1 | 1 | 1 |

| | The | Cat | In | Hat |
|---|---|---|---|---|
| The | 0.05 | .1 | 0.05 | .1 |
| Cat | 0.05 | 0.05 | .1 | 0.05 |
| In | .1 | 0.05 | 0.05 | 0.05 |
| Hat | 0.05 | 0.05 | 0.05 | 0.05 |

# GOOD TURING SMOOTHING

***The cat in the hat***

| | The | Cat | In | Hat |
|---|---|---|---|---|
| The | 0 | 1 | 0 | 1 |
| Cat | 2 | 1 | 3 | 2 |
| In | 4 | 1 | 0 | 0 |
| Hat | 1 | 1 | 2 | 1 |

| ***Bucket:*** | occurrences |
|---|---|
| ***0 counts:*** | 4 |
| ***1 count:*** | 7 |
| ***2 counts:*** | 3 |
| ***3 counts:*** | 1 |
| ***4 counts:*** | 1 |

| | The | Cat | In | Hat |
|---|---|---|---|---|
| The | 0 | .05 | 0 | .05 |
| Cat | .1 | .05 | .15 | .1 |
| In | .2 | .05 | 0 | 0 |
| Hat | .05 | . 05 | .1 | .05 |

# GOOD TURING SMOOTHING

***The cat in the hat***

| | The | Cat | In | Hat |
|---|---|---|---|---|
| The | 0 | 1 | 0 | 1 |
| Cat | 2 | 1 | 3 | 2 |
| In | 4 | 1 | 0 | 0 |
| Hat | 1 | 1 | 2 | 1 |

| ***Bucket:*** | occurrences |
|---|---|
| ***0 counts:*** | 4 |
| ***1 count:*** | 7 |
| ***2 counts:*** | 3 |
| ***3 counts:*** | 1 |
| ***4 counts:*** | 1 |

| | The | Cat | In | Hat |
|---|---|---|---|---|
| The | (.35) | .05 | (.35) | .05 |
| Cat | .1 | .05 | .15 | .1 |
| In | .2 | .05 | (.35) | (.35) |
| Hat | .05 | . 05 | .1 | .05 |

P* (zero counts) = 1 Counts/Total Counts
P* (zero counts) = 7/20 = .35

# GOOD TURING SMOOTHING

***The cat in the hat***

| | The | Cat | In | Hat |
|---|---|---|---|---|
| The | 0 | 1 | 0 | 1 |
| Cat | 2 | 1 | 3 | 2 |
| In | 4 | 1 | 0 | 0 |
| Hat | 1 | 1 | 2 | 1 |

***Bucket:*** occurrences
***0 counts:*** 4
***1 count:*** 7
***2 counts:*** 3
***3 counts:*** 1
***4 counts:*** 1

| | The | Cat | In | Hat |
|---|---|---|---|---|
| The | .35 | .05 | .35 | .05 |
| Cat | .1 | .05 | .15 | .1 |
| In | .2 | .05 | .35 | .35 |
| Hat | .05 | . 05 | .1 | .05 |

| | The | Cat | In | Hat |
|---|---|---|---|---|
| The | .214 | .856 | .214 | .856 |
| Cat | .999 | .856 | … | .999 |
| In | … | .856 | .214 | .214 |
| Hat | .856 | .856 | .999 | .856 |

UPDATE COUNTS ((c+1)Nc+1/Nc)
**1 count:** 2 x 2 Counts / 1 Counts = 2 (3/7) = .856
(**0 count:** 1 x 1 Counts / 0 Counts = 1 (.856/4) = .214)
**2 count:** 3 x 3 Counts / 2 Counts = 3 (1/3) = .999
…

# GOOD TURING SMOOTHING

***The cat in the hat***

| | The | Cat | In | Hat |
|---|---|---|---|---|
| The | 0 | 1 | 0 | 1 |
| Cat | 2 | 1 | 3 | 2 |
| In | 4 | 1 | 0 | 0 |
| Hat | 1 | 1 | 2 | 1 |

***Bucket:*** occurrences
***0 counts:*** 4
***1 count:*** 7
***2 counts:*** 3
***3 counts:*** 1
***4 counts:*** 1

| | The | Cat | In | Hat |
|---|---|---|---|---|
| The | .35 | .05 | .35 | .05 |
| Cat | .1 | .05 | .15 | .1 |
| In | .2 | .05 | .35 | .35 |
| Hat | .05 | . 05 | .1 | .05 |

| | The | Cat | In | Hat |
|---|---|---|---|---|
| The | .214 | .856 | .214 | .856 |
| Cat | .999 | .856 | … | .999 |
| In | … | .856 | .214 | .214 |
| Hat | .856 | .856 | .999 | .856 |

UPDATE PROBABILITIES (New Counts/Total Counts)
**0 count:** .214/20 = .017
**1 count:** .856/20 = .042
**2 count:** .999/20 = .049
…

| | The | Cat | In | Hat |
|---|---|---|---|---|
| The | .017 | .042 | .017 | .042 |
| Cat | .049 | .042 | … | .049 |
| In | … | .042 | .017 | .017 |
| Hat | . 042 | .042 | .049 | .042 |

# INTERPOLATION & BACKOFF

- Interpolation (sum of λs = 1) - learned from the *Holdout Set*

$$\hat{P}(w_n|w_{n-2}w_{n-1}) = \lambda_1 P(w_n|w_{n-2}w_{n-1}) + \lambda_2 P(w_n|w_{n-1}) + \lambda_3 P(w_n)$$

- Katz Backoff
  - Back off to increasingly shorter histories until you have an estimation based on observation.
  - Full version requires a discounted probability (P*) and a normalization factor ($\alpha$) to ensure that we're not adding more probability mass than is actually there.

# EVALUATION - PERPLEXITY

- Perplexity
  - Intrinsic evaluation
  - Sum all probabilities of test sentences raised to the – 1 / N (where N is the total number of word tokens encountered in the test set)
  - Lower the perplexity (good!) the higher the conditional probability – the better the model is at capturing sequence patterns
  - Perplexity typically lowers for higher order N-gram models

| | Unigram | Bigram | Trigram |
|---|---|---|---|
| **Perplexity** | 962 | 170 | 109 |

  - MUST exclude test data from the training data (values artificially high)
  - If (truly) random – Perplexity will be (approach) the size of the vocabulary.