Investigation of Katz’ Back-off Formula: We will implement the Katz back-off formula of Section 15.7, again using text A and text B from our projects. Our alphabet will once again be non-overlapping pairs of consecutive letters, so that |X| = 729, |A| = 15,000 and |B| = 2,500. However, we will build a trigram model for this alphabet, so that we will have to deal, at least conceptually, with an alphabet X×X×X of size 387,420,489 for the Good-Turing estimate \( P_T(w_1,w_2,w_3) \) in equation (24) on page 271, and an alphabet X×X of size 531,441 for the corresponding bigram estimate required for computing \( Q_T(w_3|w_2) \) in equation (23). The estimate (2) in Homework 7 will play the role of \( f(w_3) \) in equation (23).

(a) Begin by creating bigram counts \( c_d(⟨w_2,w_3⟩) \) for all seen bigrams \( ⟨w_2,w_3⟩ ∈ X×X \). Unlike the way you converted text A & B from lower-case letters to symbols of \( X \) by chunking them into nonoverlapping pairs-of-letters, here you should extract overlapping bigrams \( ⟨w_2,w_3⟩ \). i.e., you should extract \( N−1 \) bigram tokens, where \( N = 15,000 \) is the length of text A.

i. Provisionally choose \( K = 7 \) for bigrams;

ii. Using the counts \( c_d(⟨w_2,w_3⟩) \) obtained above, compute the Good-Turing estimate \( P_T(w_2,w_3) \) as a function of the bigram count \( c_d(⟨w_2,w_3⟩) \), for \( 0 ≤ c_d(⟨w_2,w_3⟩) < 7 \);

iii. Check that for all \( i ∈ \{0, \ldots, K−1\} \), and any two bigrams \( ⟨w_2,w_3⟩ \) and \( ⟨w'_2,w'_3⟩ \),
\[
c_d(⟨w_2,w_3⟩) = i \quad \text{and} \quad c_d(⟨w'_2,w'_3⟩) = i + 1 \quad ⇒ \quad P_T(w_2,w_3) ≤ P_T(w'_2,w'_3),
\]
and, if necessary, revise your choice of \( K \) in Step 1(a)i to achieve this;

iv. Using the counts-of-counts, \( n_i = \) the number of bigrams with count \( c_d(⟨w_2,w_3⟩) = i \), compute the coefficient
\[
α = \frac{\sum_{i=2}^{K−1} in_i}{\sum_{i=2}^{K} in_i};
\]

v. For each seen “history” \( w_2 \), use the probability estimates of infrequent but seen bigrams from Step 1(a)ii, the \( α \) from Step 1(a)iv, and the relative frequency estimates of frequent bigrams to compute
\[
β(w_2) = \frac{1}{\sum_{(w_2,w_3)∈S_0} P(w_3)} \left[ 1 − α \sum_{(w_2,w_3)∈S^*} Q_T(w_3|w_2) − \sum_{(w_2,w_3)∈S_K} f_d(w_3|w_2) \right],
\]
where the sums range over $w_3 \in \mathcal{X}$ for a fixed $w_2$, $Q_T(w_3|w_2) = \frac{P_T((w_2,w_3))}{f_d(w_2)}$.

\[
S_0 = \{(w_2,w_3) : c_d((w_2,w_3)) = 0\} \text{ are the unseen bigrams,}
\]
\[
S^* = \{(w_2,w_3) : 1 \leq c_d((w_2,w_3)) < K\} \text{ are the infrequent seen bigrams,}
\]
\[
S_K = \{(w_2,w_3) : c_d((w_2,w_3)) \geq K\} \text{ are the frequent bigrams,}
\]

and $\hat{P}(w_3)$ in the denominator is the Good-Turing estimate (??) we developed earlier.

vi. Compute the back-off bigram estimate

\[
\hat{P}(w_3|w_2) = \begin{cases} 
  f_d(w_3|w_2) & \text{if } c_d((w_2,w_3)) \geq K, \\
  \alpha Q_T(w_3|w_2) & \text{if } 1 \leq c_d((w_2,w_3)) < K, \\
  \beta(w_2) \hat{P}(w_3) & \text{if } c_d((w_2,w_3)) = 0,
\end{cases}
\]

and check that it sums to unity for each seen history $w_2$.

Explain how the formula in equation (1) generalizes for unseen histories $w_2$.

(b) Repeat the exercise above with trigram counts $c_d((w_1,w_2,w_3))$, and compute the trigram estimate $\hat{P}(w_3|w_1,w_2)$ of equation (22) in Section 15.7 on page 271.

(c) Compute the preplexity of text B for the bigram and trigram models of Parts 1a and 1b.

2. **Comparison of various LM smoothing techniques:** Read the report


This report is for reading only; go through it at your own pace. Noting needs to be turned in on April 14!