Instructions. Do this quiz in partnership with exactly one other student. Write both your names at the top of this page. Discuss the answer to the question with each other, and then write your joint answer below the question. Use the back of the page if necessary. It is fine if you overhear what other students say, because you still need to decide if they are right or wrong. You have seven minutes.

Question. Using notation from the online lecture notes, in Gibbs sampling for latent Dirichlet allocation (LDA) with symmetric priors,

\[ p(z_i = j | z', \bar{w}) \propto \frac{q'_{jw_i} + \beta}{\sum_t q'_{jt} + \beta} \frac{n'_{mj} + \alpha}{\sum_k n'_{mk} + \alpha}. \]

Consider using LDA to learn a topic model from a collection of tweets. A tweet is a very short document, so most tweets are about a single topic. On the other hand, many words such as “awesome” can belong to many topics. Given these considerations, is \( \alpha \ll 1 \) preferable, or larger \( \alpha \)? And \( \beta \ll 1 \) or larger \( \beta \)? Explain your answer briefly.

Answer. The scalar \( \alpha \) is the number of pseudowords that belong to each topic \( j \) in each document \( m \). When \( \alpha \) is bigger, it is easier for different positions in the same document to be assigned to different topics. Hence, for tweets, we prefer \( \alpha \ll 1 \). On the other hand, \( \beta \) is the pseudocount of prior occurrences of each word in each topic. We want it to be easy for two appearances of a word to be assigned to different topics, so we prefer larger \( \beta \).

Additional note. Making topic models work well for microtext, such as tweets, is difficult. The assumption that each tweet has a single topic implies that a clustering model may be more appropriate than a topic model. Although some words
like “awesome” may belong to many topics, other words like “Quvenzhane” intuitively do not, so it may be beneficial to use a non-symmetric $\beta$ prior.