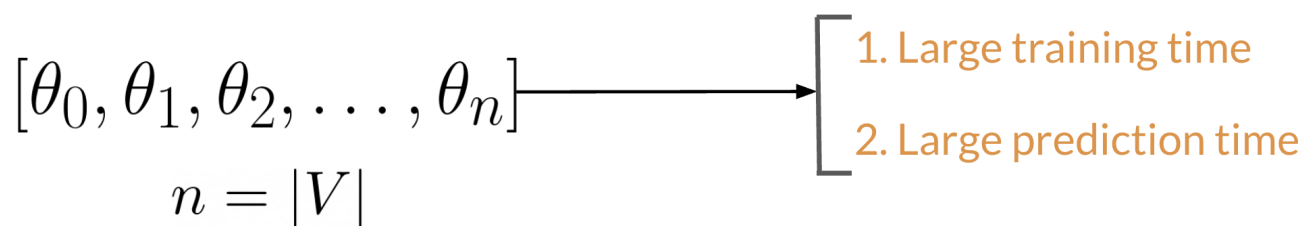# Vocabulary & Feature Extraction

Given a tweet, or some text, you can represent it as a vector of dimension V, where V corresponds to your vocabulary size. If you had the tweet "I am happy because I am learning NLP", then you would put a 1 in the corresponding index for any word in the tweet, and a 0 otherwise.

I am happy because I am learning NLP

All zeros!

$$[\,1\,,1\,,1\,,1\,,1\,,1\,,\ldots,0\,,\ldots,0\,,0\,,0\,]$$
$$1 \longleftrightarrow |V|$$

$$[\theta_0, \theta_1, \theta_2, \ldots, \theta_n]$$
$$n = |V|$$

1. Large training time

2. Large prediction time

As you can see, as V gets larger, the vector becomes more sparse. Furthermore, we end up having many more features and end up training $\theta\,V$ parameters. This could result in larger training time, and large prediction time.