

For each Speech:

Tokenize speech \rightarrow words

~~Tokenize words~~

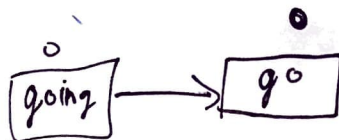
For each word in words:

1. tokenize word
2. if word in dict.key
 $\cdot \text{dict}[\text{word}].\text{count}++$

else

$\text{dict}[\text{word}] = 1$

return dict



$[\{D_1\}, \{D_2\}, \dots, \{D_5\}]$

KEY	VAL
WORD	COUNT
...	...
...	...

$$w = \frac{\text{COUNT}}{\Lambda_{ki}} \cdot \log_2 \left(\frac{n}{\Lambda_{ki}} \right)$$

$\Lambda_{ki} = \# \text{ docs containing term } i$

$\Lambda_{ki} = \text{highest count of term } i \text{ in all docs}$