Linguistics 384
Bag of words exercise

Step 1: Read the following 5 pairs of English-Chinese sentences. The code (W_i) following each Chinese word is used to help you identify the same Chinese word in different Chinese sentences.

It is proven that it is healthy to celebrate birthdays.

事实(W1) 证明(W2) 庆祝(W3) 生日(W4) 是(W5) 健康(W6) 的(W7)。

Statistics show that those people who celebrate the most birthdays become the oldest.

统计数字(W8) 表明(W9) 那些(W10) 庆祝(W3) 生日(W4) 最多(W11) 的(W7) 人(W12) 变得(W13) 最老(W14)。

So celebrate your birthday every year!

所以(W15), 每年(W16) 都(W17) 庆祝(W3) 你的(W18) 生日(W4) 吧(W19)!

Cited from Dr. Birthday's PhD dissertation Celebrate Your Birthdays to Get Healthier!

引自(W20) 生日(W4) 博士(W21) 的(W7) 博士(W21) 论文(W22) 庆祝(W3) 你的(W18) 生日(W4) 来(W23) 变得(W13) 更(W24) 健康(W6)!

Dissertation will be published by Birthday Press.

论文(W22) 将(W25) 由(W26) 生日(W4) 出版社(W27) 出版(W28)。

Step 2: Using the bag of words method, calculate the probability of *celebrate* being translated into each of the possible candidate Chinese words based on the training data you read just now. Note that it is very important that you DO NOT consider impossible Chinese candidate words. There are extra rows in the table.

Code of Candidate Chinese Word	Frequency	Probability
	, ,	Í
Total		