SNLP 2016
Exercise 2

1) (1 point) Use set theory and the axioms defining a probability function to show that:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

2) (2 points) Consider the following joint probability distribution:

x	0	0	1	1
y	0	1	0	1
$p(X = x, Y = y)$	0.32	0.08	0.48	0.12

Compute the marginal distributions: $P(X)$, $P(Y)$ and the conditional distributions $P(X|Y)$ and $P(Y|X)$. Are these random variables independent?

3) (2 points) Suppose the vocabulary has $30k$ words. Let's say the average sentence length in your corpus is 5 words. How many parameters do you need to store and estimate in order to describe the probability distribution: $P(w_1, w_2 ... w_5)$?

4) (5 points) In this exercise you will compare the probability distributions $P(W_i|W_{i-1} = "of")$ and $P(W_i|W_{i-1} = "the")$. The distribution of the words given the previous word is "of" or "the" respectively.

- Download the Brown corpus from the web `http://www.nltk.org/nltk_data/` or through the python NLTK toolkit.
- Tokenize and lowercase each token.
- Estimate the conditional probability distributions $P(W_i|W_{i-1} = "of")$ and $P(W_i|W_{i-1} = "the")$ with maximum likelihood estimation.
- The expected value of the function $f(x) = -logP(x)$ is called the entropy of the probability distribution $P(X)$.It shows how unexpected an event is on average.
  $E[-log_2 P(X)] = \sum_{i=1}^{N} P(x_i) * -log_2 P(x_i)$
- Plot the frequency distribution (unnormalized frequency counts) or the probability distribution for the 50 most frequent tokens for both distributions. Based on the plots which distribution do you expect to have a higher entropy (i.e. which history gives worse predictions about what is going to come)?
- Compute the entropy for both distributions and verify your guess.