Homework 3

Collocations, Hypothesis testing, N-gram Language Models

Natural Language Processing

Instructor: Arjun Mukherjee

Q1: Suppose that our null hypothesis on the mean height of men of *Swahili* tribe is 4ft. We are given a sample of 100 tribal men whose mean height is 6ft and standard deviation of 1.8ft. We want to know whether the 100 tribal men belong to Swahili tribe or not.
(a) What is the null hypothesis?
(b) What is the alternate hypothesis?
(c) Using a t-test, what is our confidence that the 100 tribal belong to Swahili? In other words compute the p-value to ascertain whether the 100 tribal belong to the Swahili population or is from a different population of taller tribes. You may use this t-table to compute your p-value and refer to one-tail values.
(d) Based on the standard thresholds, can we reject the null hypothesis? [4+4+4+4 = 16 points].

Q2: On a certain collection of news articles during World War II, containing a total of 100,000 terms, the following terms (words/phrases) appeared with the following frequencies (upon removal of stopwords/lemmatization/punctuations and other preprocessing):

Term	Frequency
adolf	150
hitler	200
industrial	700
revolution	900
adolf hitler	175
industrial revolution	250
hitler industrial	4
hitler revolution	14
revolution hitler	25

Assume the null hypothesis (H_0) that all words occur independently of other words, i.e., $P\ w_1 w_2\ = P(w_1)P(w_2)$ where w_1 and w_2 are single letter words. Also consider that bigrams $(w_1 w_2)$ are generated as a Bernoulli trail with a success probability of $P(w_1 w_2)$. We want to figure out where the following terms (a-e) form a collocation or not. For each case compute the p-value using a t-test. You may use this t-table to compute your p-value and refer to one-tail values. Also for each case (a-e), assuming a critical value of t = 2.576 corresponding to the p-value of 0.005 state whether the bigram is a statistically significant collocation or not?

Apply these assumptions in your computation. Assume that sample mean, x (expected number of occurrences of the collocation) is computed using corpus frequency. Expected number of occurrence of the collocation according to the null hypothesis (H_0), μ is $P\ w_1 w_2\ = P(w_1)P(w_2)$. The sample variance should be computed using the variance of the Bernoulli distribution with success probability $= x$. Note that the variance of the Bernoulli distribution is given by $p\ 1-p\ \approx p$ when p is very small, i.e., $p \ll\ 1$. [5+5+5+5 = 20 points].

(a) adolf hitler
(b) hitler industrial
(c) hitler revolution
(d) revolution hitler
(e) industrial revolution

Q3: This is based on Hypothesis testing of differences (refer to Section 5.3.2 FSNLP). Hypothesis testing of differences is a technique used for determining whether there is any statistical significance (indicated by low p-values) between samples from two different groups. The t-test for computing the difference of two groups of samples is given by $t = \ \overline{x_1}\ -\ \overline{x_2}\ \left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^{-\frac{1}{2}}$. Where $\overline{x_1}$ and $\overline{x_2}$ are the sample means of two groups of population (i.e, scores of A and scores of B in the following example) and σ_1^2 and

σ_2^2 are the sample variance. There is a subtle difference between the standard variance and sample variance. The sample variance for a group of samples $\{x_1, \ldots x_n\}$ is defined as $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} x_i - x^{\ 2}$ where $x = \frac{1}{n} \sum_{i=1}^{n} x_i$.

Now Consider two individuals, A and B who are tested on ten different intelligent quotients tests. The score (out of 100) on each test scored by A and B are tabulated below. Can we say that A is more intelligent than B? What is our confidence in the above statement? [5 points]

Hint: compute t and look up p-value using a 1-tailed using this t-table.

IQ Test	A's	B's Score
1	80	79
2	85	75
3	83	76
4	67	60
5	71	73
6	60	58
7	72	64
8	76	68
9	78	75
10	58	56

Q4: Recall that the Chi-squared (χ^2) measures the difference between two random variables using the expected observed counts to compute whether a bigram is a collocation or not. The Chi-squared (χ^2) value for a given contingency table for bigram $w_1 w_2$ with expected observed counts as:

	w_1	$\neg w_1$
w_2	W	X
$\neg w_2$	Y	Z

is computed using this expression, $\chi^2 = \frac{N\ WZ - XY\ ^2}{(W+Y)(X+Z)(W+X)(Y+Z)}$. Considering the following two contingency tables which summarizes the dependence of occurrence of w_1 and w_2 in a certain corpus. Please refer to Table 5.8 and Section 5.3.3 FSNLP to get a better understanding of the problem. [2+2+2 points]
(a) Compute Chi-squared (χ^2) measure for "garden soil".
(b) Compute the Chi-squared (χ^2) measures for "watch dog".
(c) Which out of these two bigrams tends to be more likely of a collocation. Justify your reasoning.
Hint: You don't require any p-value computations here as the Chi-squared (χ^2)value itself is sufficient to determine which out of "garden soil" or "watch dog" is more likely to be a collocation.

	$w_1 = garden$	$w_1 \neq garden$
$w_2 = soil$	15	50
$w_2 \neq soil$	200	400

	$w_1 = watch$	$w_1 \neq watch$
$w_2 = dog$	20	50
$w_2 \neq dog$	200	1000

Q5: Consider the following contingency table for the occurrence of the word "hitler" in two different corpuses, C_1 and C_2 [2+3+3 = 8 points].

	C_1=WWII News Archives	C_2 =9/11 Attacks News archive
$w =$ hitler	50	5
$w \neq$ hitler	9000	23000

(a) What does the Chi-squared (χ^2) measure in this case?
(b) Compute Chi-squared (χ^2) and p-value using this table (assume degree of freedom, df = 1).
(c) Can we say appearance or non-appearance of the word w = hitler is conditioned upon the choice of corpus (C_1 or C_2) assuming a confidence level of 95%? Justify your answer as to why or why not?

Q6: Consider the following contingency table for the occurrence of the word "war" in two different corpuses, C_1 and C_2 [2+3+3 = 8 points].

	C_1=WWII News Archives	C_2 =9/11 Attacks News archive
$w =$ war	2050	5005
$w \neq$ war	7000	18000

(a) What does the Chi-squared (χ^2) measure in this case?
(b) Compute Chi-squared (χ^2) and p-value using this table (assume degree of freedom, $df = 1$).
(c) Can we say appearance or non-appearance of the word "war" is conditioned upon the choice of corpus (C_1 or C_2) assuming a 99.9% confidence? Justify your answer as to why or why not?

Q7: What is the expression for the probability of the following sentence in terms of the probability of all unigrams [2 points].
$I\ am\ taking\ this\ course\ to\ learn.$

Q8. Derive the probability of the following sentence using chain rule and bigram approximation [4 points].

$$S = w_1\ w_2\ w_3\ \ldots\ w_n$$

Q9. Derive the probability of the following sentence using chain rule and trigram approximation [5 points].

$$S = w_1\ w_2\ w_3\ \ldots\ w_n$$

Q10. Compute these unigram and bigram probabilities based on the text snippet below. Each word has been assigned a part-of-speech tag, which is indicated after the slash. You need not solve the fraction in decimal but retain the fraction with the counts in numerator and denominator. Show the numerator and denominator (e.g., 5/10) rather than just the resulting probability (e.g., 0.5) will ensure you go the counts correct! [2 + 2 + 2 + 2 + 2 = 10 points]

I/PRO am/VERB a/ART nobody/NOUN.
Nobody/NOUN is/VERB perfect/ADJ.
Therefore/ADV I/PRO am/VERB perfect/ADJ.
(a) $P(perfect)$
(b) $P(VERB)$
(c) $P(am \mid I)$
(d) $P(ADJ \mid VERB)$
(e) $P(VERB \mid ART)$

Q11: (a) What does perplexity measure? (b) Two different language models A and B trained using different smoothing techniques on the same training data yielded a perplexity of 100 and 85 respectively on a particular held out test set. Which language model is better and why? (c) Which language model produces a higher probability of the corpus? [2 + 4 + 2 = 8 points]

Q12: This is based on the general definition of Perplexity. Consider a corpus containing $i = 1 \ldots m$ sentences. Each sentence s_i contains n_i words. Also let $M = \sum_{i=1}^{m} n_i$ denote the total number of words in the corpus.
Probability of the corpus, $C = \prod_{i=1}^{m} P(s_i)$
Perplexity of the corpus, $PPX = 2^{-l}$ where $l = 1/M \sum_{i=1}^{m} \log_2 P(s_i)$
Now, the actual computation goes like this. Using the trained ngram language model, one usually computes the probability of each sentence using the chain rule.
i.e., if the sentence $s_i = w_{i,1}\ w_{i,2}\ w_{i,3}\ \ldots . w_{i,n_i}$ then,
$P\ s_i = \prod_{j=1}^{n_i} P(w_{i,j} | w_{i,j-1})$ when we use a bigram model and $P\ s_i = \prod_{j=1}^{n_i} P(w_{i,j})$ when using a unigram model. Once the value of $P\ s_i$ is obtained, it is plugged in the PPX definition to obtain the perplexity value.

Now consider that we use another representation of the corpus, $= \{w_1 \ldots . w_M\}$, i.e, numbered by words alone and not considering the sentence structure.
Show that, when one is using a unigram language model, the following two expressions of perplexity in (A) and (B) are equivalent [10 points]:

(A) $PPX = 2^{-l}$ where $l = 1/M \sum_{i=1}^{m} \log_2 P(s_i)$

(B) $PPX = 1/ \sqrt[M]{\prod_{i=1}^{M} P\ w_i}$