CS565: Intelligent Systems and Interfaces

Words: Finding Collocations

Semester: Jan – May 2019

Ashish Anand

Associate Professor, Dept of CSE

IIT Guwahati

Announcements

- Scribe
  - Gouri Sankar Majumder, Nidhi Ahlawat: 24th Jan Lec

  - Source File also need to be submitted
    - DDMMYY-Name.tex [preferred format]
    - DDMMYY-Name.doc
    - Submission only through Canvas

  - Accept Canvas invitation
  - Send email to p.saptarshi for any issue pertaining to canvas, with subject prefix CS565:
  - All communication through canvas only
  - Saptarshi will send an email regarding submission through Canvas

Recap

- Finding Collocation
  - Mean & Variance Approach: Implicit way to handle all collocations including with varying distance

  - Issue with observation being a chance observation

  - Hypothesis Testing
    - Generic Setting: Hypotheses (Null and Alternative); Significance Level; Appropriate Statistics

    - t-test (one-tail vs two-tail)

    - Likelihood Function and Maximum Likelihood Estimator

Objective

- Continuing with statistical methods to find collocation
  - t-test
  - Pearson’s Chi-square test
  - Likelihood ratio test

Finding Collocation

Hypothesis Testing-based Methods

Hypothesis Testing: Mitigating the chance issue

- Objective: Whether the observation is significantly different than just being a random event

- Objective in our case: whether words occur together more frequently than they would have occurred together by chance

- Steps are
  - Formulate ***Null Hypothesis, H_0 :*** model random event appropriately
  - Decide Significance Level: Probability of rejecting H_0 when it is true
  - Compute the probability *p* that the *event (corresponding statistics)* occurs if H_0 is true.
  - Reject null hypothesis if *p* is less than the significance level

Statistical Test: t-test

- Null Hypothesis: *Sample is drawn from a normal distribution with mean μ*

- $t = \frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{n}}}$

Example: Study of men heights

Null Hypothesis, H_0 : Sample is drawn from general population of men with mean heights = 158 cm

Sample size, *N* = 200; Observed/sample mean = 169 cm; sample variance = 2600

$t \approx 3.05$

Critical value of t-statistics = ±2.83

Give your verdict

Question: How to use t-test in this problem?

- What are my samples?
- What is sample size?
- What is sample mean?
- What is expected mean?

Deciding sample answers all questions

- Consider corpus : collection of n-grams
- Samples: Indicator random variable corresponds to the target n-gram.
- Sample size: # of n-grams
- x_i ~ Bernoulli (p)

Using *t-test* for finding collocations

- Text corpus as a sequence of *N* bigrams
- $P(w_i)$ = # of occurrences of word w_i / total # of words [MLE]
- H_0 : $P(w_i, w_j) = P(w_i) * P(w_j)$ [occurrence of the two words are independent]

- Under null hypothesis, process of random occurrence of the bigram is a *Bernoulli Trial* with $p = P(w_i, w_j) = P(w_i) * P(w_j)$
- *Mean,* $\mu = p$; *variance* $= p(1-p) \approx p$
- Calculate $\bar{x}$ and std. dev.

Example

For the bigram *new companies*

P(new) = 15828 / 14307668

P(companies) = 4675 / 14307668

μ = P(new companies) = 3.615×10^{-7}

Actual occurrence of new companies = 8

t = 0.999932 < t_critical at 0.005 = 2.576

Give your verdict

t	$C(w^1)$	$C(w^2)$	$C(w^1\ w^2)$	w^1	w^2
4.4721	42	20	20	Ayatollah	Ruhollah
4.4721	41	27	20	Bette	Midler
4.4720	30	117	20	Agatha	Christie
4.4720	77	59	20	videocassette	recorder
4.4720	24	320	20	unsalted	butter
2.3714	14907	9017	20	first	made
2.2446	13484	10570	20	over	many
1.3685	14734	13478	20	into	them
1.2176	14093	14776	20	like	people
0.8036	15019	15629	20	time	last

Table 5.6 Finding collocations: The t test applied to 10 bigrams that occur with frequency 20.

Pearson’s Chi-square Test

- Does not require normal distribution assumption as in t-test
- Test for dependence or association
- Make a frequency or contingency table
- Compare observed frequency with expected frequency under independence assumption

Chi-square test: contd.

	w1 = new	w1 ≠ new
w2 = companies	8	4667
w2 ≠ companies	15820	14287173

$$X^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

O_{ij}: Observed frequency; E_{ij}: Expected frequency
X^2 is asymptotically χ^2 distributed.

Likelihood Ratio Test

- Two alternate hypotheses for occurrence frequency of a bigram w_1w_2
  - H_1: $p(w_2 \mid w_1) = p = p(w_2 \mid \neg w_1)$ -> Independence
  - H_2: $p(w_2 \mid w_1) = p1 \neq p2 = p(w_2 \mid \neg w_1)$ -> Association

- Calculate likelihood of observing w_2 'c_2' times when w_1 has occurred 'c_1' times
- Define Likelihood Ratio, $\lambda = L(H_1) / L(H_2)$
  - A number telling how much more likely is one hypothesis over the other.

Calculating Probabilities and Likelihood

- What we do
  - $p = c_2/N$; $p_1 = c_{12} / c_1$; $p_2 = (c_2 - c_{12}) / (N - c_1)$

  c_i: # of occurrence of w_i; c_{ij}: # of occurrence of w_{ij}

- Under the hood
  - Maximum Likelihood Estimate

Likelihood Ratio Test

	H_1	H_2
$P(w_2 \mid w_1)$	$p = c_2 / N$	$p_1 = c_{12} / c_1$
$P(w_2 \mid \neg w_1)$	$p = c_2 / N$	$p_2 = (c_2 - c_{12}) / (N - c_1)$
c_{12} out of c_1 bigrams are $w_1 w_2$	$b(c_{12}; c_1, p)$	$b(c_{12}; c_1, p_1)$
c_2-c_{12} out of N-c_1 bigrams are $\neg w_1 w_2$	$b(c_2 - c_{12}; N - c_1, p)$	$b(c_2 - c_{12}, N - c_1, p_2)$

$L(H_1) = b(c_{12}; c_1, p)\, b(c_2 - c_{12}; N - c_1, p)$
$L(H_2) = b(c_{12}; c_1, p_1)\, b(c_2 - c_{12}, N - c_1, p_2)$

$\text{Log } \lambda = \log (L(H_1) / L(H_2))$
$-2 \log L \sim \chi^2$

$-2 \log \lambda$	$C(w^1)$	$C(w^2)$	$C(w^1 w^2)$	w^1	w^2
1291.42	12593	932	150	most	powerful
99.31	379	932	10	politically	powerful
82.96	932	934	10	powerful	computers
80.39	932	3424	13	powerful	force
57.27	932	291	6	powerful	symbol
51.66	932	40	4	powerful	lobbies
51.52	171	932	5	economically	powerful
51.05	932	43	4	powerful	magnet
50.83	4458	932	10	less	powerful
50.75	6252	932	11	very	powerful
49.36	932	2064	8	powerful	position
48.78	932	591	6	powerful	machines
47.42	932	2339	8	powerful	computer
43.23	932	16	3	powerful	magnets
43.10	932	396	5	powerful	chip
40.45	932	3694	8	powerful	men
36.36	932	47	3	powerful	486
36.15	932	268	4	powerful	neighbor
35.24	932	5245	8	powerful	political
34.15	932	3	2	powerful	cudgels

Table 5.12 Bigrams of *powerful* with the highest scores according to Dunning's likelihood ratio test.

Source: Table 5.12 [FSNLP]

References

- Chapter 5 [FSNLP]