EEE 486/586

Statistical Foundations of Natural Language Processing

Assignment 1

The Hunt for Collocations

(Due 16/03/2025, 23:59 PM)

General Instructions

Groups: You are expected to work alone.

Assignment: In this homework, you will find collocations in bigram form with 3 hypothesis testing methods. These are **student's t-test**, **chi-square test** and **likelihood ratio test**. You will use the corpus in "Jane Austen Processed.txt" that is provided. This corpus consists of the concatenation of six novels of Jane Austen: *Lady Susan, Mansfield Park, Northanger Abbey, Persuasion, Pride and Prejudice*, and *Sense and Sensibility*.

Submission Guidelines:

(i) You are expected to submit two documents in this assignment: one report and one answer sheet.

(ii) Follow the instructions in this assignment and fill in the answer sheet that is provided with the assignment. Fill the tables on the answer sheet and save the changes on the PDF.

(iii) Write a brief report (2-3 pages) that explains the details of your procedure part-by-part, and answer any further questions in the assignment. **Attach your code at the end of your report as an appendix**.

(iv) Name the report "report_SurnameNameID.pdf" and submit in the report field on Moodle. Name the answer sheet "answers_SurnameNameID.pdf" and submit it in the relevant field on Moodle.

(v) Failing to meet these requirements may result in loss of grades.

Important remarks:

- Collaboration and code sharing among students are prohibited.
- You are not allowed to use any NLP specific libraries unless instructed to do so.

- Properly label all your figures and tables throughout your report.
- Your reports will be evaluated based on the proper completion of tasks, clarity of presentation of results, sufficiency of discussions regarding the results, quality of writing, plots and organization of the report and your possible insights and comments.
- There might be slight deviations in your results depending on your implementation details. However, you are expected to find the correct collocations with reasonably accurate test scores.
- Please see the following for information about academic honesty and plagiarism: http://ascu.bilkent.edu.tr/Academic_Honesty.pdf

Part 1: Corpus Preprocessing

(a) We have already performed some preprocessing to standardize the task and provide the corpus in the file "Jane Austen Processed.txt". Download this corpus and proceed with the following preprocessing steps by yourself.

(b) First, tokenize the text with the nltk library of Python. Note that the required Python version for the nltk library is ≥ 3.5. Also, you should use an up-to-date version of the nltk library (≥ 3.5). Using older versions may lead to incorrect results. (Preferably nltk=3.9.1 with python=3.10.)

(c) After the previous step, use the nltk library to find POS (part-of-speech) tags of the tokens. Use the `pos_tag()` function with parameter `tagset='universal'`. These tags will be useful to find the correct lemma for each token and to filter the collocation candidates.

(d) Then, lemmatize the tokens using the WordNetLemmatizer in nltk. You may check "custom_lemmatizer.py" for an example use. Note: If you have not done a full installation of nltk, you may need to download WordNetLemmatizer first.

(e) You will use the lemmatized tokens to find bigram counts. Two configurations will be used where the size of the collocation window is changed:

- A collocation window of size 1 selects the second word of collocation candidates as the word which is 1 word to the right of the first word, ie.:
  "... word **word1 word2** word ..."
- A collocation window of size 3 selects the second word of collocation candidates from all 3 words to the right of the first word, ie.:
  "... word **word1 word2** word ..."
  "... word **word1** word **word2** word ..."
  "... word **word1** word word **word2** word ..."

You will find collocations using both configurations, separately, in Part 2.

(f) Find the collocation candidates for both configurations by following these steps:

- Eliminate all bigrams except those with POS tags **NOUN-NOUN** or **ADJ-NOUN**.

- Eliminate bigrams that include stopwords (https://gist.github.com/sebleier/554280).
- Eliminate bigrams including any punctuation marks. (Hint: You can use the `isalpha()` function).
- Eliminate bigrams that occur less than 10 times.

As a result, you should have two sets of collocation candidates, one for each window size.

Part 2: Finding the Collocations

(a) Write a function for each of the three hypothesis testing methods (student's t-test, chi-square test, likelihood ratio test) that calculates the scores of collocation candidates and sorts them by their scores. Fill in the tables in the answer sheet with the top 20 candidates. Indicate the scores, bigram counts and individual word counts. An example is given below in Table 1 for t-test scores on Tolstoy's *Anna Karenina*.

Rank	**Bigram**	**t-score**	**c(w1w2)**	**c(w1)**	**c(w2)**
1	stepan arkadyevitch	20.89272	442	547	510
2	alexey alexandrovitch	20.60396	431	628	524
3	sergey ivanovitch	15.59907	245	299	283
4	darya alexandrovna	11.66978	137	209	193
5	lidia ivanovna	9.475952	90	110	95
6	old man	8.718279	80	392	522
7	countess lidia	7.977724	64	164	110
8	agafea mihalovna	7.273684	53	74	64
9	great deal	7.05524	50	206	55
10	first time	6.347678	44	340	564
11	madame stahl	6.155495	38	121	46
12	sick man	5.945857	36	63	522
13	anna arkadyevna	5.93569	36	737	53
14	old prince	5.808735	35	392	164
15	young man	5.735658	35	207	522
16	next day	5.693718	33	96	308
17	several times	5.556931	31	71	86
18	marya nikolaevna	4.996019	25	65	31
19	princess varvara	4.880443	24	317	29
20	young men	4.819674	24	207	190

Table 1: Student's t-test table with collocation window size 1 for Anna Karenina

Hint: Keep in mind that using a collocation window of size 3 triples the total number of bigrams. Calculate probabilities and expected values accordingly.

Hint: If binomial distribution probabilities are too low to calculate, use a very small positive number instead.

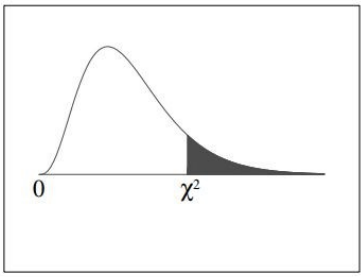
Part 3: Explaining the Statistical Tests

(a) Explain how you evaluated the t-score, chi-square score and log-likelihood score for the bigrams "good wish" and "high spirit" when you take the size of the collocation window as 1. Write down the formulas you used for evaluating these scores.

(b) Then, for a significance of $\alpha = 0.005$, decide, for all 3 tests, whether these bigrams are collocations or not. Use the following tables to make your decisions. (A t-distribution with many degrees of freedom is also equivalent to the standard normal distribution). **Put all of your explanations and decisions in your report**.

DF	A = 0.1	0.05	0.025	0.01	0.005	0.001	0.0005
∞	ta = 1.282	1.645	1.96	2.326	2.576	3.091	3.291
1	3.078	6.314	12.706	31.821	63.656	318.289	636.578
2	1.886	2.92	4.303	6.965	9.925	22.328	31.6
3	1.638	2.353	3.182	4.541	5.841	10.214	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.61
5	1.476	2.015	2.571	3.365	4.032	5.894	6.869
6	1.44	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.86	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.25	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587

Figure 1: One-tailed t table [1]



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750

Figure 2: Chi-square table [2]

References

[1] “T Table - T Distribution (Score, Chart)”, 2022. [Online]. Available: https://t-tables.net/. [Accessed: 15-Mar-2022].
[2] “Chi-Square Distribution Table” [Online]. Available: http://kisi.deu.edu.tr/joshua.cowley/Chi-square-table.pdf. [Accessed: 26-Sep-2020].