Statistical Foundations of NLP: Testing Methods

Gülce Erdoğan 21802781 EEE

Abstract - This study examines bigram collocation identification in Fyodor Dostoyevsky's literary corpus. Using three statistical hypothesis tests (t-test, chi-square test, and likelihood ratio test) with different window sizes (1 and 3) the hypothesis testing scores are found for collocation candidates. The presence of collocations is found by comparing test scores to threshold values. The study provides valuable insights into computational linguistic analysis challenges, particularly in the context of literary texts.

1 Introduction

In this assignment, we examine the collocations within the corpus of Fyodor Dostoyevsky via three statistical hypothesis testing methods being student's t-test, chi-square test, and the likelihood ratio test. This corpus contains six of Dostoyevsky's novels and it is preprocessed and analyzed to identify bigram collocations. We develop functions for each testing method, adjusting calculations for varying window sizes (1 and 3) to maintain result accuracy. In the end, we compare the final test scores with the threshold values to better examine whether the bigrams are collocations or not.

2 Results

2.1 Part 1: Corpus Preprocessing

2.1.1 Part A

In this assignment, I used the "Fyodor Dostoyevsky Processed.txt" corpus that was provided to us by our professors. This corpus was collected by combining six novels by the author Dostoyevsky: Crime and Punishment, Devils, Notes from the Underground, The Brothers Karamazov, The Gambler, and The Idiot. Some preprocessing tasks such as standardization were already done; however, I need to further preprocessing to perform statistical tests on collocations. Therefore, in Part A, this corpus was downloaded and opened for this purpose.

2.1.2 Part B

In Part B, the text was tokenized using the NLTK library, and the total number of tokens was found as 1425758.

2.1.3 Part C

In Part C, after tokenizing the text, the part-of-speech (POS) tags of the words were found by using the pos_tag() function of the NLTK library. These tags were a necessary step towards lemmatization since they were utilized while locating the accurate lemma for each token and were also utilized in filtering the selection of collocation candidates.

2.1.4 Part D

In Part D, the lemmatization was done by using the WordNetLemmatizer in the NLTK library. The professors gave us the "custom_lemmatizer.py" file to be used for this purpose. Therefore, I found the lemmatized corpus via this class. The number of 5 different words was also examined the word "that", "the", "abject", "london", and "." were found as 19429, 48392, 21, 2, and 51738, respectively.

2.1.5 Part E

In Part E, bigram counts (frequencies) were calculated under two different configurations having distinct collocation window sizes using the lemmatized tokens. The first configuration has a collocation window size of 1 that forms bigrams via each adjacent word pair, and the second has a collocation window size of 3 that forms bigrams by including the next 3 words; thus, providing a broader perspective on word associations. Additionally, the frequencies of two different words were examined in this section, and both of the frequencies of the collocations "magnificent capital" and "bright fire" were found as 1.

2.1.6 Part F

In Part F, the collocation candidates for both window size configurations were found within the desired criteria in distinct substeps. The process began by eliminating all bigrams that did not possess the POS tag structure NOUN-NOUN or ADJ-NOUN. Then, the bigrams containing stopwords were removed according to the stopword list in GitHub source. Afterward, bigrams including any punctuation marks were also excluded by Python's built-in is_alpha() function. Finally, the bigrams that appeared less than 10 times were also discarded. As a result, two distinct sets of collocation candidates were obtained, each corresponding to one of the window size configurations, thereby setting the stage for subsequent statistical analysis.

In addition to the above, the FAQ section provided by our professors clarified an important ambiguity regarding the counting of collocation candidate frequencies. Two primary methodologies were proposed: one considers all appearances of a collocation candidate, regardless of their POS tags, while the second method exclusively counts

occurrences aligned with the POS tags of the candidate and merges them at the end. In my implementation, I adopted the first approach by first calculating all the bigram frequencies and then iterating through each bigram in that dictionary. For every bigram, the function verified its POS tags and if the bigram contained a POS tag that conforms to either an ADJ-NOUN or NOUN-NOUN structure, it was appended to the collocation candidates list without checking whether every instance of this bigram had ADJ-NOUN or NOUN-NOUN structure. This method allowed us to take all the instances of the collocation candidates regardless of their POS tags. Hence, more comprehensive candidate collocation occurrences were captured.

Additionally, whether two different bigrams were among the collocation candidates was examined in this section and the result was found as "No" for both "mr. skimpole" with window size 1 and "spontaneous combustion" with window size 3.

2.2 Part 2: Finding the Collocations

2.2.1 Part A

In the second part, the functions for 3 different hypothesis testing methods were developed to compute the scores of collocation candidates for the 2 distinct window sizes (1 and 3). The utilized testing methods were the student's t-test, chi-square test, and likelihood ratio test. Each testing method used bigram frequencies, individual word frequencies, and total word count to compute the respective scores. At the end of the computation, I separately rank the scores according to different testing methods and list the top 20 collocation candidates for each situation. The results of this ranking can be observed in Figures 1–6.

Regarding the FAQ provided by our professors, I need to consider the impact of larger window size (3 in our case) on the total count of the bigrams and individual word frequencies. Due to the increased window size, the total number of bigrams and the probabilities of unigrams were tripled; therefore, I need to adjust it while calculating the hypothesis testing scores. To solve this issue, I utilized the second approach that was mentioned in the FAQ where I calculated the word counts directly from the list of all bigrams rather than the tokenized corpus. I also paid attention not to count the same word twice when deriving the count from bigrams. Hence, with this method, I maintained the accuracy of my calculations.

Rank	Word	Score	$c(w1w2)$	$c(w1)$	$c(w2)$
1	stepan trofimovitch	22.619069	512	525	513
2	pyotr stepanovitch	22.547831	509	834	509
3	varvara petrovna	20.534433	422	474	507
4	katerina ivanovna	20.239065	410	427	635
5	nikolay vsyevolodovitch	17.657104	312	518	312
6	fyodor pavlovitch	17.052922	291	306	461
7	old man	16.857563	289	1356	2546
8	nastasia philipovna	15.583748	243	417	251
9	young man	15.140569	232	776	2546
10	old woman	14.353161	208	1356	1047
11	yulia mihailovna	14.175298	201	215	202
12	pyotr petrovitch	13.100114	172	834	331
13	lizabetha prokofievna	13.074941	171	185	177
14	great deal	12.790646	164	1202	237
15	dmitri fyodorovitch	12.720228	162	427	327
16	evgenie pavlovitch	12.563966	158	227	461
17	thousand rouble	11.980513	144	614	543
18	long time	11.793031	143	1074	2623
19	go away	11.511941	136	1858	1342
20	mavriky nikolaevitch	11.487925	132	149	132

Fig. 1. Top collocation candidates, student's t-test (window size 1)

Rank	Word	Score	$c(w1w2)$	$c(w1)$	$c(w2)$
1	stepan trofimovitch	1387728.43	512	525	513
2	ippolit kirillovitch	1359440.81	41	43	41
3	lef nicolaievitch	1359440.81	41	43	41
4	avdotya romanovna	1341882.35	112	119	112
5	yulia mihailovna	1326304.32	201	215	202
6	nikodim fomitch	1316081.54	24	24	26
7	lizabetha prokofievna	1273169.85	171	185	177
8	mavriky nikolaevitch	1263071.68	132	149	132
9	trifon borissovitch	1235650.87	39	45	39
10	rodion romanovitch	1205266.43	82	97	82
11	mihail makarovitch	1197632.52	21	25	21
12	gavrila ardalionovitch	1173526.56	58	61	67
13	arina prohorovna	1120123.57	39	44	44
14	varvara petrovna	1056418.47	422	474	507
15	semyon yakovlevitch	1034362.57	37	51	37
16	kuzma kuzmitch	983274.48	20	29	20
17	daria alexeyevna	980371.30	19	21	25
18	darya pavlovna	935804.97	49	62	59
19	katerina ivanovna	883755.64	410	427	635
20	pyotr stepanovitch	869957.83	509	834	509

Fig. 2. Top collocation candidates, chi-squared test (window size 1)

Rank	Word	Score	$c(w1w2)$	$c(w1)$	$c(w2)$
1	ippolit kirillovitch	69.521040	41	43	41
2	lef nicolaievitch	69.521040	41	43	41
3	nikodim fomitch	69.473603	24	24	26
4	mihail makarovitch	68.994638	21	25	21
5	trifon borissovitch	68.572449	39	45	39
6	kuzma kuzmitch	68.402672	20	29	20
7	avdotya romanovna	68.340263	112	119	112
8	semyon zaharovitch	68.147640	10	51	10
9	semyon yakovlevitch	67.912991	37	51	37
10	mavriky mavrikyevitch	67.912991	11	149	11
11	rodion romanovitch	67.697947	82	97	82
12	o u	67.662663	14	225	14
13	von sohn	67.592173	19	74	19
14	mavriky nikolaevitch	67.527414	132	149	132
15	de cominges	67.478888	17	243	17
16	father iosif	67.313039	19	1139	19
17	marya kondratyevna	67.217767	26	125	26
18	dmitri prokofitch	67.162036	23	427	23
19	sofya matveyevna	66.788045	51	135	51
20	sofya semyonovna	66.729425	71	135	71

Fig. 3. Top collocation candidates, likelihood ratio test (window size 1)

Rank	Word	Score	$c(w1w2)$	$c(w1)$	$c(w2)$
1	stepan trofimovitch	22.602372	512	1575	1539
2	pyotr stepanovitch	22.521479	509	2501	1526
3	varvara petrovna	20.518057	422	1421	1520
4	katerina ivanovna	20.220295	410	1281	1904
5	nikolay vsyevolodovitch	17.644302	312	1553	935
6	fyodor pavlovitch	17.041366	291	917	1381
7	old man	16.603300	290	4066	7633
8	nastasia philipovna	15.574371	243	1249	752
9	young man	15.025591	234	2327	7633
10	old woman	14.459309	215	4066	3140
11	yulia mihailovna	14.171001	201	645	606
12	o clock	13.221849	175	675	579
13	lizabetha prokofievna	13.071447	171	554	530
14	pyotr petrovitch	13.070649	172	2501	992
15	great deal	12.798607	165	3603	711
16	dmitri fyodorovitch	12.704857	162	1280	981
17	evgenie pavlovitch	12.552339	158	680	1381
18	ha ha	12.521412	157	677	677
19	thousand rouble	12.474037	157	1840	1629
20	hundred rouble	12.329844	153	1282	1629

Fig. 4. Top collocation candidates, student's t-test (window size 3)

Rank	Word	Score	$c(w1w2)$	$c(w1)$	$c(w2)$
1	stepan trofimovitch	461893.16	512	1575	1539
2	ippolit kirillovitch	453091.95	41	129	123
3	lef nicolaievitch	453091.95	41	129	123
4	avdotya romanovna	447144.47	112	357	336
5	yulia mihailovna	441833.13	201	645	606
6	nikodim fomitch	438661.54	24	72	78
7	lizabetha prokofievna	425730.11	171	554	530
8	mavriky nikolaevitch	420847.60	132	447	396
9	trifon borissovitch	411831.33	39	135	117
10	mihail makarovitch	411103.42	21	74	62
11	rodion romanovitch	404676.92	82	290	245
12	gavrila ardalionovitch	391097.91	58	183	201
13	arina prohorovna	373322.26	39	132	132
14	varvara petrovna	352056.50	422	1421	1520
15	semyon yakovlevitch	344737.95	37	153	111
16	o clock	334914.48	175	675	579
17	kuzma kuzmitch	327731.26	20	87	60
18	daria alexeyevna	326764.87	19	63	75
19	wisp tow	323415.10	14	54	48
20	darya pavlovna	311869.44	49	186	177

Fig. 5. Top collocation candidates, chi-squared test (window size 3)

Rank	Word	Score	c(w1w2)	c(w1)	c(w2)
1	semyon zaharovitch	63.155763	10	153	30
2	mavriky mavrikyevitch	62.924804	11	447	33
3	je suis	62.794400	10	138	39
4	thou wouldst	62.682332	11	506	39
5	dark brown	62.548382	10	623	45
6	wisp tow	62.528363	14	54	48
7	o u	62.443396	14	675	42
8	full speed	62.244908	14	675	48
9	eye glowed	62.241415	10	3588	57
10	eye shone	62.236514	11	3588	54
11	thousand roubles	62.172252	11	1840	57
12	literary matinée	62.063738	10	198	69
13	de cominges	62.061037	17	724	51
14	fit coughing	61.979217	10	603	72
15	kuzma kuzmitch	61.975095	20	87	60
16	mihail makarovitch	61.975095	21	74	62
17	printing press	61.959683	10	36	96
18	saith unto	61.919759	12	39	90
19	von sohn	61.905781	19	222	57
20	pan vrublevsky	61.852921	15	87	71

Fig. 6. Top collocation candidates, likelihood ratio test (window size 3)

2.3 Part 3: Explaining the Statistical Tests

2.3.1 Part A

In Part A of the final part of the assignment, I performed a detailed evaluation of the t-score, chi-square score, and log-likelihood score for two specified bigrams: ”head clerk” and ”great man”, within a collocation window size of 1. For each bigram, I first retrieved the count of the bigram occurrence, the count of the individual words, and the total word count from the corpus as inputs of the 3 methods.

For the t-score, I calculated the real mean ($\overline{X}$) by dividing the bigram count by the total count (N), and the expected mean (μ) as the product of individual word frequencies divided by N. The real mean was founded by MLE (p_{MLE}) and the expected mean was equal to the null hypothesis (H_0). The variance for small p was assumed to be the real mean. The t-score was then calculated using the formula [1]:

$$t = \frac{\overline{X} - \mu}{\sqrt{\frac{S^2}{N}}} = \frac{p_{MLE} - H_0}{\sqrt{\frac{p_{MLE}}{N}}} \tag{1}$$

For the chi-square score, I computed the observed and expected frequencies of the four possibilities (both words (O_{11}), first word only (O_{12}), second word only (O_{21}), neither word (O_{22})), and substituted these values into the chi-square formula (shortcut for 2-by-2) [1]:

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

$$X^2 = \frac{N \times (O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})} \tag{3}$$

For the log-likelihood score, I first calculated the probabilities and binomial distribution likelihoods under the null and alternative hypotheses and then computed the log-likelihood ratio using these likelihoods. The likelihood ratio was calculated using the formula where the likelihood of independence (L_{H1}) and the likelihood of dependence (L_{H2}) were calculated with binomial distribution formula and the probabilities below [1]:

$$likelihood = -2 \times log\left(\frac{L_{H1}}{L_{H2}}\right) \tag{4}$$

$$L(H1) = b(c_{12}; c_1, p) \times b(c_2 - c_{12}; N - c_1, p) \tag{5}$$

$$L(H2) = b(c_{12}; c_1, p_1) \times b(c_2 - c_{12}; N - c_1, p_2) \tag{6}$$

$$P(X = k) = C(n, k) \times p^k \times (1 - p)^{n-k} \tag{7}$$

$$p = \frac{c_2}{N}; \quad p_1 = \frac{c_{12}}{c_1}; \quad p_2 = \frac{c_2 - c_{12}}{N - c_1} \tag{8}$$

In the probabilities above, p is the probability of the second word in the corpus, p_1 is the empirical conditional probability of the second word given the first word, p_2 is the empirical conditional probability of the second word given the absence of the first word, c_{12} is the count of the bigram, c_1 and c_2 are the counts of the individual words and N is the total number of bigrams.

2.3.2 Part B

In Part B of the final stage of the assignment, for a significance level of λ = 0.005, we investigated whether the two bigrams ”head clerk” and ”great man” were collocations or not considering all three statistical tests. For this purpose, we looked into statistical tables corresponding to each test. For the t-test the degree of freedom was chosen as ∞ and for chi-square and likelihood ratio, it was chosen as 1 since the degree of freedom for 2x2 table is $(c-1)(r-1) = (2-1)(2-1) = 1$. According to the threshold values the two bigrams were found as collocation for all the hypothesis tests. The results can be observed in Figure 7,8

Test	Score	Threshold	Collocation? (Yes/No)
t-Test	4.674126	2.576	Yes
Chi-square Test	6294.816683	7.879	Yes
Likelihood Ratio Test	60.603199	7.879	Yes

Fig. 7. Tests results for (“head”, “clerk”)

Test	Score	Threshold	Collocation? (Yes/No)
t-Test	3.736722	2.576	Yes
Chi-square Test	117.402985	7.879	Yes
Likelihood Ratio Test	45.158766	7.879	Yes

Fig. 8. Tests results for (“great”, “man”)

3 Discussions & Conclusions

In summary, by using different statistical hypothesis testing methods on Fyodor Dostoyevsky’s works, we have studied how to identify word combinations in a literary collection. This research has given us a better understanding of the importance of word combinations in analyzing text and how window sizes and testing methods affect the process. The results of this study provide valuable insights into the difficulties and complexities of computational linguistic analysis and provide a solid groundwork for future investigations.

References

[1] A. Koç, *Lecture 6 collocations*, Slide show in Course [EEE 486/EEE 586 Statistical Foundations of Natural Language Processing], Bilkent University, 2023.