Assignment 2: Answer Sheet

Name	Emirhan Koç
ID Number	22101553

Part 1.b

What is the number of tokens in the corpus? $N =$ 1899306

Part 1.d

What are the counts of the following words after lemmatization?

"that":	21406
"the":	76792
"negligent":	7
"london":	410
".":	68278

Part 1.e

How many times does the bigram ("fashionable","intelligence") occur in windows of size 1?

14

How many times does the bigram ("high","chancery") occur in windows of size 3?

6

Part 1.f

Is ("mr.","skimpole") among collocation candidates for window size 1? (Yes/No)

No

Is ("spontaneous","combustion") among collocation candidates for window size 3? (Yes/No)

No

Part 3.a

Is ("cursitor","street") a bigram according to the three tests with window size 1?

Test	Score	Threshold	Collocation? (Yes/No)
t-Test	3.6	2.576	Yes
Chi-square Test	25643	16.7	Yes
Likelihood Ratio Test	189	16.7	Yes

Table 7: Tests results for ("cursitor","street")

Is ("good","one") a bigram according to the three tests with window size 1?

Test	Score	Threshold	Collocation? (Yes/No)
t-Test	0.84	2.576	No
Chi-square Test	0.97	16.7	No
Likelihood Ratio Test	0.86	16.7	No

Table 8: Tests results for ("good","one")

Bigram	c_w1w2	w1	w2	t-score
sir leicester	384.0	3682.0	466.0	19.551794
old gentleman	372.0	3440.0	2118.0	19.090279
young lady	349.0	1872.0	1975.0	18.579049
old lady	293.0	3440.0	1975.0	16.909571
young man	269.0	1872.0	3177.0	16.211447
old man	259.0	3440.0	3177.0	15.737005
miss havisham	229.0	2093.0	317.0	15.110573
long time	181.0	1687.0	3494.0	13.223577
lady dedlock	175.0	1975.0	355.0	13.201460
last night	168.0	1309.0	1606.0	12.876655
dear sir	172.0	2518.0	3682.0	12.743251
bob sawyer	158.0	291.0	262.0	12.567134
young gentleman	152.0	1872.0	2118.0	12.159992
good deal	149.0	3196.0	364.0	12.156854
next day	133.0	622.0	2005.0	11.476029
miss pro	132.0	2093.0	160.0	11.474178
great deal	126.0	2304.0	364.0	11.186006
old woman	127.0	3440.0	1145.0	11.085777
young woman	120.0	1872.0	1145.0	10.851773
miss summerson	118.0	2093.0	200.0	10.842828

Size 1

t-scores

	Bigram	c_w1w2	w1	w2	t-score
467	sir leicester	384.0	11046.0	1398.0	19.458271
636	old gentleman	375.0	10320.0	6354.0	18.771249
166	young lady	351.0	5616.0	5925.0	18.423855
539	old lady	295.0	10320.0	5925.0	16.551193
895	young man	272.0	5616.0	9531.0	15.923209
14	old man	262.0	10320.0	9531.0	15.120285
626	miss havisham	230.0	6279.0	951.0	15.096954
752	lady dedlock	176.0	5925.0	1065.0	13.183226
360	ha ha	173.0	1389.0	1389.0	13.127402
571	long time	189.0	5061.0	10482.0	13.070718
411	last night	172.0	3927.0	4818.0	12.861881
916	bob sawyer	158.0	873.0	786.0	12.560399
676	dear sir	177.0	7554.0	11046.0	12.203598
634	good deal	149.0	9588.0	1092.0	12.056177
120	young gentleman	154.0	5616.0	6354.0	11.905176
73	next day	135.0	1866.0	6015.0	11.449549
899	miss pro	132.0	6279.0	480.0	11.443219
616	miss summerson	128.0	6279.0	600.0	11.255394
180	great deal	126.0	6912.0	1092.0	11.107083
822	old woman	133.0	10320.0	3435.0	10.993224

Size 3

Bigram	c_w1w2	w1	w2	t-score	chi-square
robinson crusoe	10	10	10	3.162269	1.899306e+06
chesney wold	91	104	105	9.539018	1.440284e+06
leo hunter	40	45	56	6.324412	1.205893e+06
dingley dell	20	28	25	4.472077	1.085308e+06
nathaniel pipkin	31	57	42	5.567583	7.623976e+05
bob sawyer	158	291	262	12.567134	6.217569e+05
serjeant buzfuz	40	119	48	6.324146	5.319853e+05
saint antoine	36	92	52	5.999637	5.144947e+05
serjeant snubbin	31	119	36	5.567405	4.260309e+05
gabriel grub	18	47	41	4.242422	3.193224e+05
dame durden	16	45	34	3.999815	3.177736e+05
john smauker	25	122	31	4.999635	3.138479e+05
stephen blackpool	32	178	38	5.656272	2.875041e+05
madame defarge	92	194	301	9.588690	2.751853e+05
south wale	11	37	23	3.316499	2.700406e+05
samuel slumkey	21	104	36	4.582171	2.236908e+05
job trotter	44	189	89	6.631991	2.185432e+05
masr davy	45	115	172	6.706731	1.943830e+05
jack maldon	25	112	58	4.999349	1.827044e+05
bayham badger	11	15	86	3.316430	1.781394e+05

Size 1

chi-square

Bigram	c_w1w2	w1	w2	t-score	chi-square
robinson crusoe	10	30	30	3.162230	633088.666667
chesney wold	91	312	315	9.537660	479973.182137
leo hunter	40	135	168	6.323948	401910.900053
dingley dell	20	84	75	4.471897	361742.666868
quebec malta	10	45	48	3.162161	263776.805661
nathaniel pipkin	31	171	126	5.567100	254091.190295
bob sawyer	158	873	786	12.560399	207041.666368
pinch snuff	26	135	153	5.098320	186439.759089
serjeant buzfuz	40	357	144	6.323151	177275.091781
saint antoine	36	276	156	5.998760	171450.243659
serjeant snubbin	31	357	108	5.566564	141968.972061
knife fork	37	327	189	6.080999	126152.177554
dodson fogg	51	339	348	7.138561	125538.599112
kenge carboy	31	453	108	5.566237	111871.580355
lincoln inn	39	129	630	6.242735	106574.985898
gabriel grub	18	141	123	4.241930	106416.804869
dame durden	16	135	102	3.999401	105903.199481
john smauker	25	366	93	4.998816	104582.637365
ho ho	12	90	90	3.463695	101275.520739
dedlock baronet	66	1065	240	8.118564	96995.622200

Size 3

Bigram	c_w1w2	w1	w2	t-score	chi-square	likelihood-ratio
doctor manette	73	807	163	8.536062	7.683815e+04	877.154264
low voice	89	575	870	9.406283	2.991879e+04	873.445898
leo hunter	40	45	56	6.324412	1.205893e+06	843.047732
miss murdstone	83	2093	299	9.074465	2.076840e+04	755.789785
serjeant buzfuz	40	119	48	6.324146	5.319853e+05	746.247246
spinster aunt	50	62	936	7.066840	8.176449e+04	700.625591
job trotter	44	189	89	6.631991	2.185432e+05	698.725226
lord chancellor	47	317	100	6.853305	1.322873e+05	686.706737
masr davy	45	115	172	6.706731	1.943830e+05	683.922665
saint antoine	36	92	52	5.999637	5.144947e+05	667.541345
miss flite	57	2093	93	7.536373	3.162492e+04	653.913397
several time	76	287	3494	8.657409	1.081004e+04	626.315729
first time	113	1518	3494	10.367754	4.360830e+03	624.101093
nathaniel pipkin	31	57	42	5.567583	7.623976e+05	618.543570
uriah heep	44	220	148	6.630742	1.128656e+05	612.311606
miss betsey	53	2093	85	7.267345	2.991716e+04	610.756607
old lady	293	3440	1975	16.909571	2.348412e+04	598.672124
miss havisham	229	2093	317	15.110573	1.498521e+05	598.672124
old man	259	3440	3177	15.737005	1.118455e+04	598.672124
young lady	349	1872	1975	18.579049	6.200043e+04	598.672124

Size 1

LRT

Bigram	c_w1w2	w1	w2	t-score	chi-square	likelihood-ratio
dedlock baronet	66	1065	240	8.118564	96995.622200	835.055933
doctor manette	74	2421	489	8.578228	26221.385219	724.244091
dodson fogg	51	339	348	7.138561	125538.599112	702.666269
leicester baronet	59	1398	240	7.673519	59014.583002	699.984590
leo hunter	40	135	168	6.323948	401910.900053	670.460172
lord chancellor	53	951	300	7.273266	56006.772518	635.353068
dear sir	177	7554	11046	12.203598	1805.879004	598.672124
good deal	149	9588	1092	12.056177	11807.944929	598.672124
lady dedlock	176	5925	1065	13.183226	27653.741677	598.672124
long time	189	5061	10482	13.070718	3477.503673	598.672124
little man	136	12024	9531	9.937369	670.262067	598.672124
serjeant buzfuz	40	357	144	6.323151	177275.091781	596.430604
mender road	45	138	1236	6.703768	67572.564579	585.019283
miss murdstone	83	6279	897	9.002000	6812.872196	572.504394
masr davy	45	345	516	6.703573	64734.350166	570.733217
job trotter	44	567	267	6.629270	72789.065900	567.828386
lincoln inn	39	129	630	6.242735	106574.985898	552.483395
spinster aunt	50	186	2808	7.058136	27188.209087	545.139108
saint antoine	36	276	156	5.998760	171450.243659	542.686299
uriah heep	46	660	444	6.774774	41059.858945	536.628281

Size 3