T-61.5020 Statistical Natural Language Processing

Exercises 6 — Collocations
Version 1.0

You can choose a subset of words for the exercise.

1. In Table 1 there is frequencies of several word pairs. Calculate the best candidates for collocations using the frequency method. Are the results better if you normalize the number of bigrams by the product of the number of its components?

2. Calculate the mean and variance for the place of occurrence for some word pairs in the table. Judge how well those can be used for determining collocations. How does the window size affect results? The given statistics were counted over a window of five words.

3. Use t-test and Pearson's chi-square test do determine whether the words pairs are collocations. Compare the results.

4. Find collocations using mutual information and compare to the previous results.

Table 1: *Word frequences. $C(a)$ stands for how many times a occurred in the test corpus. x marks some unknown word. There were total of 28 181 344 words in the corpus, and they were transformed to the base forms before counting the frequencies.*

s_1	s_2	$C(s_1)$	$C(s_2)$	$C(s_1, s_2)$	$C(s_1, x, s_2)$	$C(s_2, s_1)$	$C(s_2, x, s_1)$
hakea	työ	10435	26174	31	26	22	11
valkoinen	talo	3665	10767	710	2	1	6
herne	nenä	115	974	3	0	0	0
ja	olla	818046	1387476	7329	39979	3612	38162
venäjä	presidentti	27637	26855	717	216	10	24
vihainen	mielenosoittaja	589	1757	7	0	0	0
tuntematon	sotilas	1967	4806	154	4	0	0
aste	pakkanen	2879	1440	160	8	13	32
heittää	veivi	8126	21	5	0	0	1
kova	tuuli	20613	3916	279	16	9	12
liukas	keli	735	728	106	2	3	7
sekä	myös	50193	135637	138	124	34	244
oppia	lukea	2831	8952	21	4	7	1
olla	ula	1387476	44	3	2	1	2
ottaa	onki	38304	110	9	3	0	0

Some translations:

- hakea = apply for, työ = job
- valkoinen = white, talo = house
- herne = pea, nenä = nose, "herne nenässä" = "pissed off"
- ja = and, olla = be
- Venäjä = Russia, presidentti = president
- vihainen = angry, mielenosoittaja = demonstrator
- tuntematon = unknown, sotilas = soldier
- aste = degree, pakkanen = frost
- heittää = throw, veivi = crank, "heittää veivinsä" = "kick the bucket"
- kova = hard, tuuli = wind
- liukas = slippery, keli = weather, conditions
- sekä = as well as, myös = also
- oppia = learn, lukea = read
- olla = be, ULA = VHF, "olla ulalla" = "be confused"
- ottaa = take, onki = hook and line, "ottaa onkeensa" = "learn one's lesson"