SNLP Assignment 1

Clemens Damke - 7011488

1. Smoothing

1.1. Simple bi-gram matrix

	</s>	and	capital	city	england	europe	in	is	largest	live	london	million	of	people	river	thames	the	western
<s>	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	1	0
and	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
capital	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
city	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
england	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
europe	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
in	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	1
is	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2	0
largest	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
live	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
london	2	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
million	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
of	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
people	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
river	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
thames	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
the	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
western	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

1.2. Add-1 smoothing

	</s>	and	capital	city	england	europe	in	is	largest	live	london	million	of	people	river	thames	the	western
<s>	1/22	1/22	1/22	1/22	1/22	1/22	1/22	1/22	1/22	1/22	3/22	1/11	1/22	1/22	1/22	1/22	1/11	1/22
and	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	2/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19
capital	1/19	2/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19
city	1/20	1/20	1/20	1/20	1/20	1/20	1/10	1/20	1/20	1/20	1/20	1/20	1/10	1/20	1/20	1/20	1/20	1/20
england	2/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19
europe	2/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19
in	1/21	1/21	1/21	1/21	1/21	1/21	1/21	1/21	1/21	1/21	1/7	1/21	1/21	1/21	1/21	1/21	1/21	2/21
is	1/21	1/21	1/21	1/21	1/21	1/21	2/21	1/21	1/21	1/21	1/21	1/21	1/21	1/21	1/21	1/21	1/7	1/21
largest	1/20	1/20	1/20	3/20	1/20	1/20	1/20	1/20	1/20	1/20	1/20	1/20	1/20	1/20	1/20	1/20	1/20	1/20
live	1/19	1/19	1/19	1/19	1/19	1/19	2/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19
london	3/22	1/22	1/22	1/22	1/22	1/22	1/22	3/22	1/22	1/22	1/22	1/22	1/22	1/22	1/22	1/22	1/22	1/22
million	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	2/19	1/19	1/19	1/19	1/19
of	1/19	1/19	1/19	1/19	2/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19
people	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	2/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19
river	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	2/19	1/19	1/19
thames	1/19	1/19	1/19	1/19	1/19	1/19	1/19	2/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19
the	1/21	1/21	2/21	1/21	1/21	1/21	1/21	1/21	2/21	1/21	1/21	1/21	1/21	1/21	2/21	1/21	1/21	1/21
western	1/19	1/19	1/19	1/19	1/19	2/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19	1/19

1.3. Kneser-Ney smoothing

Kneser-Ney might yield better results for the given corpus at least in some cases. For example `a = (largest, in)` and `b = (largest, largest)` get the same probability using Add-1-smoothing. Using Kneser-Ney `a` would correctly be considered more likely than `b` because the word `in` appears in more contexts (i. e. it has more distinct preceding words) than the word `largest`. This can be considered as an improvement. To see whether the overall result improves, one would have to compute the complete bi-gram matrix though and check for regressions.