EDAN20
Harald Nordgren
900404-1936

Lab Report 1

We collected the corpus using *Selma.txt*.

It was tokenized using a modified version of *token_perl.pl*, searching for a capital letter followed an arbitrary number of letters in lowercase and then a construct of (a space plus any number of letters) repeated arbitrarily and ended by punctuation. This produced a good result, with the exception of one sentence in the corpus that for some reason starts with a lowercase letter and was thus missed.

Putting parentheses around most of the expression above, except for the punctuation sign at the end means that it can be swiftly removed when the text is then replaced using the s command and *<s> $1 <\/s>*.

We ran the unigram and bigrams *count* programs. As an upper bound for the amount of bigrams, we have every combination of words, i.e. the total number of unigrams squared. Because of grammar, however, that number is drastically lowered, and in our training set a lot fewer than all the grammatically correct bigrams will actually occur.

To cope with bigrams that don't occur in the corpus we can use *backoff* method, where a small non-zero probability is assigned to a bigrams, based on its individual unigrams.

We wrote programs to compute sentence probabilities for unigrams and bigrams. For the test sentence "Det var en gång en katt som hette Nils" the results were as follows: (where the missing value for *Ci,i+1* in the second table for "hette" och "nils" is because this bigram doesn't occur in the corpus and its probability was consequently "backed off".)

wi	C(wi)	#words	P(wi)
<s>	61661	1085893	0.0567836794232949
det	22087	1085893	0.0203399414122754
var	12850	1085893	0.0118335784464952
en	13921	1085893	0.0128198634672109
gång	1328	1085893	0.00122295658964557
en	13921	1085893	0.0128198634672109
katt	15	1085893	1.38135156962979e-05
som	16790	1085893	0.0154619285693894
hette	107	1085893	9.85364119669249e-05
nils	84	1085893	7.73556878992682e-05
</s>	61661	1085893	0.0567836794232949

Prob. unigrams: 2.53949626383163e-28

wi	wi+1	Ci,i+1	C(i)	P(wi+1|wi)
<s>	det	5782	61661	0.0937707789364428
det	var	4021	22087	0.182052791234663
var	en	753	12850	0.0585992217898833
en	gång	691	13921	0.0496372387041161
gång	en	21	1328	0.0158132530120482
en	katt	5	13921	0.000359169599885066
katt	som	2	15	0.133333333333333
som	hette	50	16790	0.00297796307325789
hette	nils		107	9.85364119669249e-05
nils	</s>	2	84	0.0238095238095238

Prob. bigrams: 2.62718476579265e-19