Johanna Petersson
Adi10jpe@student.lu.se

Report assignment 1

Collection and normalizing a corpus

Splitting a running text without headings or table of contents, or any other unusually formatting is not that hard up until a point. The absolute most sentences in Swedish ends with a punctuation and starts with a capital letter. My guess would be that that takes it up to about 90% accuracy. It is after that it gets hard. Given that standard punctuation, includes a dot, question mark, exclamation mark and perhaps some more, what happens when one of those is used in a sentence, or a sentence ends with the punctuation and then something like a citation sign? There are more special cases than can be tested for. I knew all this going in having worked with text data before and it is even more apparent comparing my results with the ones given in the instructions for the assignment.

Counting unigrams and bigrams

The number of unigrams in a text is the number of unique words in the text. The number of possible bigrams for the same text is the number of unigrams to the power of the number of unigrams, a lot. But it is limited to the number of words in the text, there can't be more bigrams than there are actual bigrams in the text.

As to handling unseen bigrams in the corpus, one strategy is the back off strategy that we used in this assignment. In this case it is simply backing off to using the probability of the unigram instead of the bigram

Computing the likelihood of a sentence

Given that there is a proper tokenization method along with normalization it isn't that hard to calculate the probability of a sentence using unigrams or bigrams. It is just a collection of simple computations.

What I observed from this assignment was mostly how much the result can differ between two different tokenization. Comparing my results with the ones in the instructions for the assignment the numbers doesn't look that close when computing sentence probability and its other values while individual values for unigrams or bigrams are pretty close.

Result for the sentence "Det var en gång en katt som hette Nils"

w_i	C(w_i)	#words	P(w_i)	
<s>	-		-	-
det	21647	1048418	0,020647	
var	12563	1048418	0,011983	
en	13593	1048418	0,012965	
gång	1305	1048418	0,001245	
en	13593	1048418	0,012965	
katt	15	1048418	0,000014	
som	16492	1048418	0,015730	
hette	105	1048418	0,000100	
nils	84	1048418	0,000080	
</s>	48768	1048418	0,046516	

Sannolikhet för meningen: 4.34866198405341E-27
Entropy för meningen: 7.961044271603936
Perplexity för meningen: 249.17996618738627

w_i	w:(i+1)	C(w_i,w_i+1)	C(w_i)	P(w_i+1 | w_i)
<s>	det	4413	48768	0.0904896653543307
det	var	3912	21647	0.18071788238554995
var	en	730	12563	0.05810714001432779
en	gång	672	13593	0.04943721032884573
gång	en	23	1305	0.017624521072796936
en	katt	4	13593	2.942691091002722E-4
katt	som	2	15	0.13333333333333333
som	hette	50	16492	0.0030317729808391953
hette	nils	0	105	backoff: 8.012071521091778E-5
nils	</s>	1	84	0.011904761904761906

Sannolikhet för meningen: 1.0381199339631934E-18
Entropy för meningen: 5.430975689103229
Perplexity för meningen: 43.14064044525362