EDAN20 Assignment 2
Language models

Matilda Andersson, ma4748an-s

September 2020

1 Introduction

The purpose of this laboratory report is to describe and comment on the language model composed in this assignment in order to calculate the probability distribution over sequences of words. Novels written by Selma Lagerlöf made up the corpus as was retrieved from the text file `Selma.txt`.

2 Process description

The following subsections contains a more in depth description of the process for creating the language model.

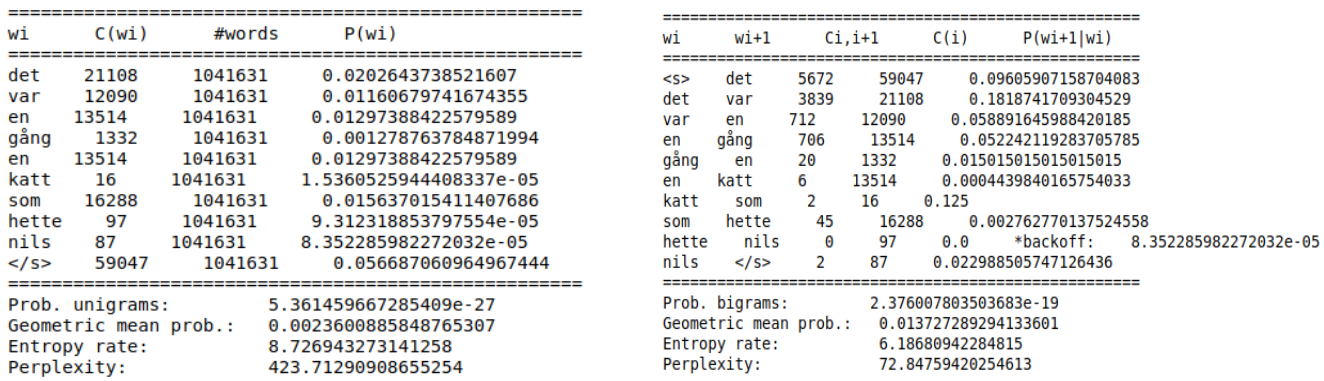
2.1 Segmenting the corpus

After retrieving the text corpus the first steps of the process focused on processing and modifying the text into a structure we easily could work with in the subsequent steps. The text alteration started with a tokenization process using the regex encoding `r"\p{L}+"` to find all words, which was followed by a segmentation step that replaced all non letters, or punctuation signs, with a white-space, and, using the regex encoding `r"[\.\;\:\?\!]\p{Z}+(\p{Lu})"` we were able to identify the endings of the sentences in the corpus. Each sentence mas marked with the markup code `r" </s>\n<s> \1"`.

2.2 Unigrams, bigrams and likelihoods

Continuing the process we moved on to counting and the number of uni- and bigrams present in the corpus. With this information we could then compute the likelihood of a sentence's probability and the results of the unigram and bigram calculations are presented in figure 1.

The number of bigrams identified were counted to 320124, while the theoretical maximum number of possible bigrams would be $n\hat{2}$, with n being the number of words in the text corpus. The difference is explained by the fact that words, and sequences of words, are repeated in a text, which will result in an overlap. Furthermore, not all words in the corpus will appear as combinations in the text, as nonsensical sequences of words such as "Me bookcase" will not be grammatically correct. A "backoff" strategy is used for analyzing and handling bigrams that previously are unseen, and instead takes the second word in the bigram and controls its existence among the unigrams, hence, using the words probability. Furthermore, the algorithm by default puts the probability to 1, in order to avoid a



```
=====================================================
wi      C(wi)      #words      P(wi)
=====================================================
det    21108     1041631     0.0202643738521607
var    12090     1041631     0.01160679741674355
en    13514     1041631     0.01297388422579589
gång    1332     1041631     0.001278763784871994
en    13514     1041631     0.01297388422579589
katt    16     1041631     1.5360525944408337e-05
som    16288     1041631     0.015637015411407686
hette    97     1041631     9.312318853797554e-05
nils    87     1041631     8.352285982272032e-05
</s>    59047     1041631     0.056687060964967444
=====================================================
Prob. unigrams:          5.361459667285409e-27
Geometric mean prob.:    0.0023600885848765307
Entropy rate:            8.726943273141258
Perplexity:              423.71290908655254
```

```
================================================
wi      wi+1      Ci,i+1      C(i)      P(wi+1|wi)
================================================
<s>    det     5672     59047     0.09605907158704083
det    var     3839     21108     0.1818741709304529
var    en     712     12090     0.058891645988420185
en    gång     706     13514     0.052242119283705785
gång    en     20     1332     0.015015015015015015
en    katt     6     13514     0.0004439840165754033
katt    som     2     16     0.125
som    hette     45     16288     0.002762770137524558
hette    nils     0     97     0.0     *backoff:    8.352285982272032e-05
nils    </s>     2     87     0.022988505747126436
================================================
Prob. bigrams:          2.376007803503683e-19
Geometric mean prob.:   0.013727289294133601
Entropy rate:           6.18680942284815
Perplexity:             72.84759420254613
```

Figure 1: Using unigrams and bigrams to determine a sentence's probability.

potential division by zero scenario that could occur if a word has not previously been seen. Thus, unfortunately, skewing the result if a sentence consists of mostly unseen words, as the probability value otherwise given by the unigram completely becomes neglected. In order to solve this issue, Norving's proposed calculation for computing the possibility of an unknown word could be used.

Additionally, we were tasked with writing five sentences on our own and run them through the bigram perplexity method. The sentences and their corresponding results are presented here in the following tables:

```
['<s>', 'idag', 'har', 'varit', 'en', 'bra', 'dag', '</s>']
=====================================================
wi      wi+1      Ci,i+1      C(i)      P(wi+1|wi)
=====================================================
<s>    idag     0     59047     0.0     *backoff:    9.60032871525521e-07
idag    har     0     1     0.0     *backoff:    0.004479513378538081
har    varit     210     4666     0.045006429489927134
varit    en     118     1721     0.06856478791400349
en    bra     39     13514     0.0028858961077401213
bra    dag     0     466     0.0     *backoff:    0.0009043509649770408
dag    </s>     150     942     0.1592356687898089
=====================================================
Prob. bigrams:         5.515065742378143e-18
Geometric mean prob.:   0.00696135588478838
Entropy rate:           7.166415953049683
Perplexity:             143.65017628033536
```

```
['<s>', 'den', 'gamle', 'mannen', 'suckade', 'medan', 'han', 'satte', 'sig', 'ner', '</s>']
================================================
wi      wi+1      Ci,i+1      C(i)      P(wi+1|wi)
================================================
<s>    den     1375     59047     0.023286534455603164
den    gamle     138     11624     0.011871988988300069
gamle    mannen     4     233     0.017167381974248927
mannen    suckade     0     511     0.0     *backoff:    4.800164357627605e-05
suckade    medan     0     50     0.0     *backoff:    0.0005452986710264959
medan    han     136     568     0.23943661971830985
han    satte     51     21589     0.0023623141414609292
satte    sig     184     459     0.4008714596949891
sig    ner     195     9250     0.02108108108108108
ner    </s>     36     1420     0.02535211267605634
================================================
Prob. bigrams:         1.5054396779813008e-20
Geometric mean prob.:   0.015775001027663077
Entropy rate:           5.986216090582471
Perplexity:             63.39143802567099
```

```
['<s>', 'det', 'blir', 'fasligt', 'kallt', 'i', 'sverige', 'på', 'vintern', '</s>']
=====================================================
wi      wi+1      Ci,i+1      C(i)      P(wi+1|wi)
=====================================================
<s>    det     5672     59047     0.096059071587040 83
det    blir     136     21108     0.006443054765965511
blir    fasligt     0     576     0.0     *backoff:    1.3440460201357294e-05
fasligt    kallt     0     14     0.0     *backoff:    3.840131486102084e-05
kallt    i     2     40     0.05
i    sverige     23     16508     0.0013932638720620305
sverige    på     0     56     0.0     *backoff:    0.013680468419238674
på    vintern     6     14250     0.0004210526315789474
vintern    </s>     23     100     0.23
=====================================================
Prob. bigrams:         2.948214052418311e-23
Geometric mean prob.:   0.005584134898459208
Entropy rate:           7.484450490860336
Perplexity:             179.07876836498406
```

```
['<s>', 'nils', 'är', 'en', 'hund', 'med', 'stora', 'bruna', 'ögon', '</s>']
=====================================================
wi      wi+1      Ci,i+1      C(i)      P(wi+1|wi)
=====================================================
<s>    nils     7     59047     0.00011854962995579793
nils    är     0     87     0.0     *backoff:    0.006038606761895527
är    en     304     6290     0.04833068362480127
en    hund     17     13514     0.0012579547136303093
hund    med     0     53     0.0     *backoff:    0.008774700445743262
med    stora     43     9140     0.004704595185995623
stora    bruna     0     1324     0.0     *backoff:    2.880098614576563e-05
bruna    ögon     6     30     0.2
ögon    </s>     106     442     0.2398190045248869
=====================================================
Prob. bigrams:         2.4819921614277553e-21
Geometric mean prob.:   0.008699214520247819
Entropy rate:           6.844899143344643
Perplexity:             114.95290726219642
```

```
['<s>', 'detta', 'är', 'en', 'konstig', 'mening', '</s>']
=====================================================
wi      wi+1      Ci,i+1      C(i)      P(wi+1|wi)
=====================================================
<s>    detta     298     59047     0.00504682710383254
detta    är     157     2482     0.06325543916196616
är    en     304     6290     0.04833068362480127
en    konstig     1     13514     7.399733609590054e-05
konstig    mening     0     7     0.0     *backoff:    0.00014880509508645575
mening    </s>     31     155     0.2
=====================================================
Prob. bigrams:         3.3978415515924226e-14
Geometric mean prob.:   0.01190929794260862
Entropy rate:           6.391767821456868
Perplexity:             83.96800590757242
```

Figure 2: Calculated sentence perplexity using bigrams.