LUND UNIVERSITY, INSTITUTION FOR ELECTRICAL AND INFORMATION TECHNOLOGY

EDAN20 - Lab 2

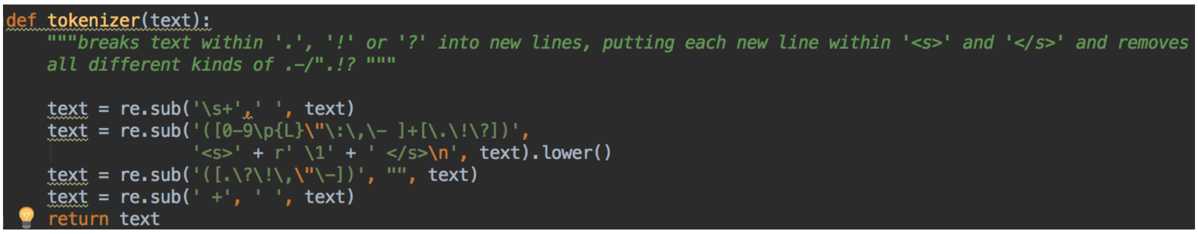
Gustav Fahlén elt14gfa@student.lu.se

October 3, 2018

1

This report is done with purpose to increase the understanding of python and how to use programming to analyze different texts, using language models.

This is done by starting off to divide the text in to one sentence on each line, marking the start of a sentence with a '<s>' and the end of a sentence with a '</s>'. All different kinds of punctuation was also supposed to be remove from the text, leaving only words containing letters or numbers and start and end flags. To form the chosen text in to that kind of structure the Regular expression operators were used, mainly re.sub, as seen in Figure 1.1.



```
def tokenizer(text):
    """breaks text within '.', '!' or '?' into new lines, putting each new line within '<s>' and '</s>' and removes
    all different kinds of .-/".!? """

    text = re.sub('\s+',' ', text)
    text = re.sub('([0-9\p{L}\"\:\,\- ]+[\.\!\?])',
                  '<s>' + r' \1' + ' </s>\n', text).lower()
    text = re.sub('([.\?\!\,\"\-])', "", text)
    text = re.sub(' +', ' ', text)
    return text
```

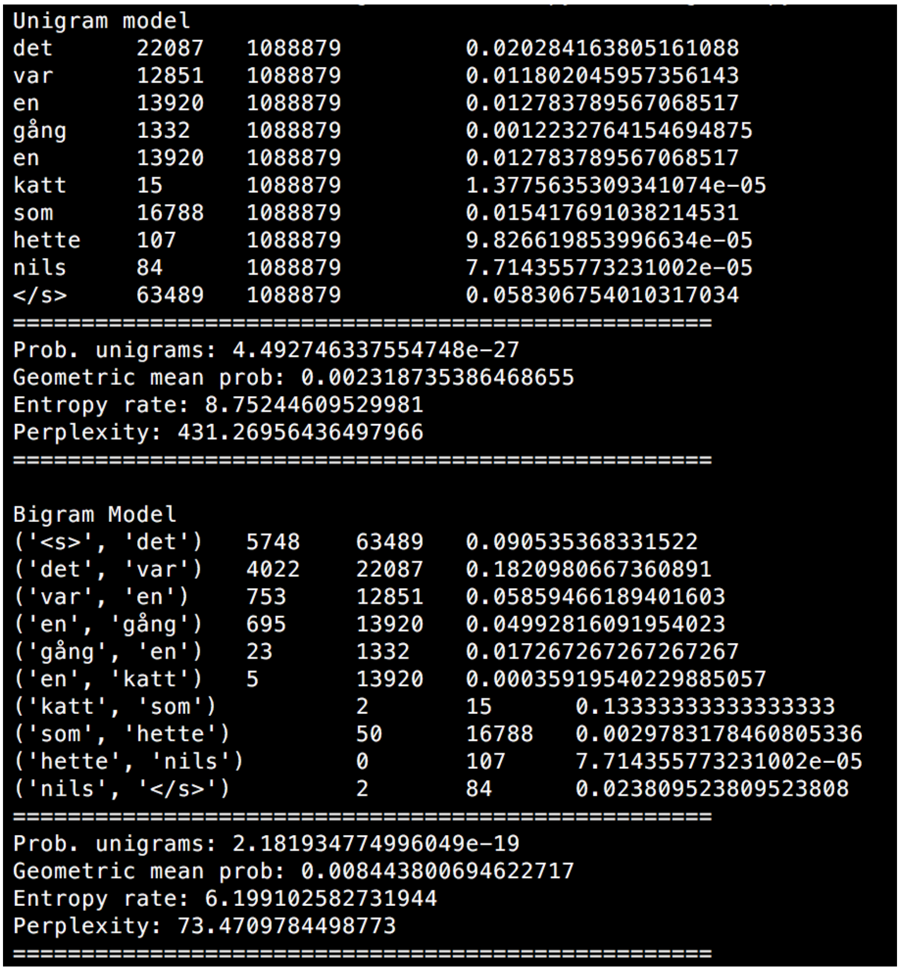
Figur 1.1: Describing the normalization of a corpus

The last sentences from the given text 'Selma.txt' can be seen below, when applying it to the tokenizer method:

<s> hon hade fått större kärlek av sina föräldrar än någon annan han visste och sådan kärlek måste vändas i välsignelse </s>
<s> när prästen sa detta kom alla människor att se bort mot klara gulla och de förundrade sig över vad de såg </s>
<s> prästens ord tycktes redan ha gått i uppfyllelse </s>
<s> där stod klara fina gulleborg ifrån skrolycka hon som var uppkallad efter själva solen vid sina föräldrars grav och lyste som en förklarad </s>
<s> hon var likaså vacker som den söndagen då hon gick till kyrkan i den röda klänningen om inte vackrare </s>

The next task is to use and learn about n-grams. A n-gram means one or many adjacent words depending on the level of N. The purpose of this part of the lab is to learn how to count how many different N-grams their are in the chosen text.

These N-grams are then used in a following task to compute different kinds of likelihood of different n-grams from the sentence "det en var en gång en katt som hette nils." in the text 'Selma.txt' file. When the number of each n-gram is found in 'Selma.txt' are these values used to calculate Prob of Unigram, Geometric mean, entropy rateand Perplexity. The outprint from this method can be seen below in Figure 1.2.



```
Unigram model
det      22087   1088879         0.020284163805161088
var      12851   1088879         0.011802045957356143
en       13920   1088879         0.012783789567068517
gång     1332    1088879         0.0012232764154694875
en       13920   1088879         0.012783789567068517
katt     15      1088879         1.3775635309341074e-05
som      16788   1088879         0.015417691038214531
hette    107     1088879         9.826619853996634e-05
nils     84      1088879         7.714355773231002e-05
</s>     63489   1088879         0.058306754010317034
===================================================
Prob. unigrams: 4.492746337554748e-27
Geometric mean prob: 0.002318735386468655
Entropy rate: 8.75244609529981
Perplexity: 431.26956436497966
===================================================

Bigram Model
('<s>', 'det')   5748    63489   0.090535368331522
('det', 'var')   4022    22087   0.1820980667360891
('var', 'en')    753     12851   0.05859466189401603
('en', 'gång')   695     13920   0.04992816091954023
('gång', 'en')   23      1332    0.017267267267267267
('en', 'katt')   5       13920   0.00035919540229885057
('katt', 'som')          2       15      0.13333333333333333
('som', 'hette')         50      16788   0.0029783178460805336
('hette', 'nils')        0       107     7.714355773231002e-05
('nils', '</s>')         2       84      0.023809523809523808
===================================================
Prob. unigrams: 2.181934774996049e-19
Geometric mean prob: 0.008443800694622717
Entropy rate: 6.199102582731944
Perplexity: 73.4709784498773
===================================================
```

Figur 1.2: Likelihood of the sentene 'Det var en gång en katt som hette Nils.' in 'Selma.txt'

Depending on the chosen sentence and the order of the words in the sentence different results will be obtained. I think these kinds of facts can be quite interesting when doing research about a specific authors authorship and can also be used when improving spell correctors, since it tells which word or words is more likely to occur after one and another.