Natural Language Processing
Portfolio I

Leon F.A. Wetzel
Information Science
University of Groningen - Faculty of Arts
l.f.a.wetzel@student.rug.nl

February 2, 2021

Abstract

In this document, you can find the results and explanations for the assignments of the first part of the portfolio for the course Natural Language Processing, taught at the University of Groningen. The corresponding Python code can be found at https://github.com/leonwetzel/natural-language-processing.

1 Week 1 - Regular Expressions

We used https://regex101.com/ to test our regular expressions. In this document, we use ∧ to better display/represent a caret. Spaces cannot be displayed in LaTeXsadly, so these are not shown below.

1.1 Regular Expressions I

1. Set of all alphabetic strings.
   $\wedge[a-zA-z]+\$$
2. Set of all lower case alphabetic strings ending in b.
   $\wedge[a-z]*[b]\$$
3. Set of all strings from the alphabet a, b such that each a is immediately preceded by a b and immediately followed by a b.
   $\wedge(bab)+\$$

1.2 Regular Expressions II

1. Set of all strings with two consecutive repeated words (e.g., *"Humbert Humbert"* and *"the the"* but not *"the bug"* or *"the big and the bug"*).
   $\backslash b(\backslash w+)\backslash b\backslash s+\backslash 1$
2. All strings that start at the beginning of the line with an integer and that end with a word.
   $\wedge\backslash d+.+\backslash b[a-zA-Z]+\backslash b\$$
3. All strings that have both the word *grotto* and the word *raven* in them (but not, e.g., words like *grottos* that merely contain the word *grotto*).
   $\backslash bgrotto\backslash b.*|\backslash braven\backslash b.*$

1.3 ELIZA

Create a chatbot in Python using regular expressions. See the attached jupyter notebook for details.

1.4 Byte-Pair Encoding

Experiment with more or less aggressive forms of tokenization and segmentation into subwords using the training data and code as explained in the jupyter notebook.

2 Week 2 - N-gram Language Models

2.1 J&M exercise 3.1

Write out the equation for trigram probability estimation (modifying Eq. 3.11). Now write out all the non-zero trigram probabilities for the I am Sam corpus on page 41.

2.2 J&M exercise 3.2

Calculate the probability of the sentence i want chinese food. Give two probabilities, one using Fig. 3.2, and another using the add-1 smoothed table in Fig. 3.6.

2.3 J&M exercise 3.6

Suppose we train a trigram language model with add-one smoothing on a given corpus. The corpus contains V word types. Express a formula for estimating P(w3|w1,w2), where w3 is a word which follows the bigram (w1,w2), in terms of various N-gram counts and V. Use the notation c(w1,w2,w3) to denote the number of times that trigram (w1,w2,w3) occurs in the corpus, and so on for bigrams and unigrams.

2.4 J&M exercise 3.7

We are given the following corpus, modified from the one in the chapter: <s> I am Sam </s> <s> Sam I am </s> <s> I am Sam </s> <s> I do not like green eggs and Sam </s> If we use linear interpolation smoothing between a maximum-likelihood bi-gram model and a maximum-likelihood unigram model with $\lambda 1 = 1/2$ and $\lambda 2 = 1/2$, what is P(Sam|am)? Include <s> and </s> in your counts just like any other token.

2.5 N-grams in the notebook

Answer the two questions in the notebook ngrams_exercise.