Assignment I

Natural Language Models and Interfaces 2012

4 April 2012

This exercise is meant to be done during the lab session and continued in your own time. You can work in teams of 2. Finish before the next lab session, and show the teaching assistant that you have completed the assignment to receive 3 points towards your final grade for the course.

The lab session today consists of two parts:

1. Brief introduction to unix/linux and some basic shell commands. Download a training corpus. Use unix utilities to view the corpus, to search for particular patterns, and to generate a word frequency distribution. Use `gnuplot` or another program to visualize the distribution.

2. Write a program (in java) to calculate the frequencies of ngrams, with $n \in \{1, 2, 3, \ldots\}$ as a parameter.

Useful resources:

- gnuplot cheatsheet `http://nishantkumar.com/cheats/gnuplot.html`
- bash cheatsheet `http://cheat-sheets.org/saved-copy/bash_ref.pdf`
- grep manual `info grep`

1 Obtaining and handling the training set

The assignments in the first few weeks will work with the OVIS corpus (`http://www.let.rug.nl/vannoord/Ovis/`). Start with creating a folder for your project, download the training set and look through it. In a terminal window you type:

```
mkdir nlmi12
cd nlmi
wget http://www.let.rug.nl/vannoord/Ovis/full_info.txt
less full_info.txt
```

(Exit the viewer by hitting `q`). The corpus contains questions and user-generated answers, and some semantic, syntactic and phonetic details about those answers. For now, we are only interested in the (transcribed) answers, marked with `A:`. You can filter out only these answers and remove the `A:` with the following commands:

```
less full_info.txt | grep '^A:' |
     sed 's/^A: //' > ovis-trainset.txt
less ovis-trainset.txt
```

Question 1 *What do the different parts of the grep and sed commands mean? Check the grep manual (under regular expression -¿ anchoring) to get the meaning of* `^`.

With regular expressions and the unix tools grep, sed, sort, and uniq it is very easy to generate a word frequency distribution:

```
less ovis-trainset.txt |  sed 's/\s\+/\n/g' | grep -v '^$' |
     sort | uniq -c | sort -g -r -k 1 >
     word-frequency-distribution.txt
%$
gnuplot
gnuplot> plot 'word-frequency-distribution.txt' using
           (log($0)):(log($1))gnuplot> set terminal jpeg
gnuplot> set output 'zipf.jpg'
gnuplot> plot 'word-frequency-distribution.txt' using
           (log($0)):(log($1))
```

(Exit gnuplot by typing 'quit').

Question 2 *What do the different parts of the grep, sed, uniq and sort commands mean? Use the unix 'info' utility to get information on these command (e.g.,* `info uniq`

Question 3 *What kind of distribution is the word frequency distribution? What does a straight line in log-log space mean?*

Question 4 *How does the word frequency distribution of the Ovis corpus compare to that from a different type of text? E.g., that of a Jane Austen novel (*`wget http://www-nlp.stanford.edu/fsnlp/statest/austen.txt`*).*

2 Extracting ngrams and counting frequencies

Write a Java program which takes as input a natural number n and a large text file (a corpus). The program should construct a table of all word sequences of length n in the corpus together with the number of times each sequence appears in the corpus (the corpus frequency of the sequence). The word sequences should be exactly as they appear in the corpus, so 'The' $\neq$ 'the'.

Take special care of the beginning and end of every sentences. Add $n-1$ 'dummy' start symbols `<s>` to the beginning of a sentences, and an end symbol `</s>` to the end of a sentence, such that when extracting trigrams from a sentence `a b c d` you arrive at the following set:

```
<s> <s> a
<s> a b
a b c
c d </s>
```

Make it possible for the program to take an additional argument m and report the m most frequent sequences (together with their frequencies and in decreasing frequency order).

Test your program with the ovis-trainset.txt corpus as the input corpus.

Question 5 *What are the top 10 most frequent bigrams, trigrams and tetragrams in the OVIS corpus?*