Andrew Coonce
CS769 Advanced Natural Language Processing
Homework 1
January 27th, 2010

1) Solve x by hand.

$$(1 \quad 1)\begin{pmatrix}1 & 2\\ 3 & x\end{pmatrix}\begin{pmatrix}1\\ 1\end{pmatrix} = x^2$$

$$(4 \quad 2+x)\begin{pmatrix}1\\ 1\end{pmatrix} = x^2$$

$$6 + x = x^2$$

$$x^2 - x + 6 = 0$$

$$x = -2, -3$$

2) Compute the derivative (with respect to x) of the function

$$\frac{1}{1+e^{-x}}$$

$$\frac{d}{dx}(1+e^{-x})^{-1}$$
$$= \left(\frac{d}{dx}(1+e^{-x})\right)(1+e^{-x})^{-2}$$
$$= e^{-x}(1+e^{-x})^{-2}$$
$$= \frac{e^{-x}}{(1+e^{-x})^2}$$

3) Find the minimum of the function $f(x, y) = x + y$, where (x, y) must be on the unit circle.

$$x^2 + y^2 = 1$$

$$y = \pm\sqrt{1 - x^2}$$

$$\min\left(x \pm \sqrt{1 - x^2}\right)$$

$$\left(x \pm \sqrt{1 - x^2}\right)\frac{d}{dx} = 0$$

$$1 \pm \frac{x}{\sqrt{1 - x^2}} = 0$$

$$\pm x = \sqrt{1 - x^2}$$

$$x^2 = 1 - x^2$$

$$2x^2 = 1$$

$$x = \frac{1}{\sqrt{2}}$$

$$y = \sqrt{1 - \left(\frac{1}{\sqrt{2}}\right)^2} = \sqrt{1 - \frac{1}{2}} = \sqrt{\frac{1}{2}}$$

$$f(x, y) = \sqrt{\frac{1}{2}} + \sqrt{\frac{1}{2}} = \frac{2}{\sqrt{2}}$$

4) Let x be a random variable drawn from a Gaussian distribution with mean 0 and variance $\frac{1}{2\lambda}$. Write down the expression for $\log p(x)$.

$$p(x) = \frac{e^{-\frac{(x-0)^2}{2\left(\frac{1}{2\lambda}\right)^2}}}{\frac{\sqrt{2\pi}}{2\lambda}}$$

$$p(x) = \frac{2\lambda e^{-2x^2\lambda^2}}{\sqrt{2\pi}}$$

$$\log p(x) = -2x^2\lambda^2 \log\left(\frac{2\lambda}{\sqrt{2\pi}}\right)$$

5) Download the particular version of *Alice's Adventures in Wonderland* from http://pages.cs.wisc.edu/~jerryzhu/cs769/dataset/alice.txt. This is the document we'll be working on.

   a. Sentence Segmentation. Download MXTERMINATOR, a sentence boundary detector, from http://pages.cs.wisc.edu/~jerryzhu/cs769/code/jmx.tar.gz. Follow the instructions in MXTERMINATOR.html. If you use TCSH, simply do SETENV CLASSPATH MXPOST.JAR then you should be able to run it. Use the EOS.PROJECT that comes with the package. Apply it to *Alice*.
   b. Tokenization. Once you have segmented out sentences, it's time to separate individual words. Download the Penn Treebank tokenizer from http://pages.cs.wisc.edu/~jerryzhu/cs769/code/tokenizer.tar.gz. This is a UNIX SED program. Run it with SED -F. It needs an input file with one sentence per line. Apply the tokenizer to the processed *Alice* corpus.
   c. Stemming. Download and compile the Porter stemmer from http://pages.cs.wisc.edu/~jerryzhu/cs769/code/porter.c. Run the stemmer on *Alice* from the previous step. You will notice that it maps all words to lower case, and some words look funny.

Question 5.1. Do not strip punctuations or otherwise change the tokens out of the stemmer. How many word tokens and word types are there?

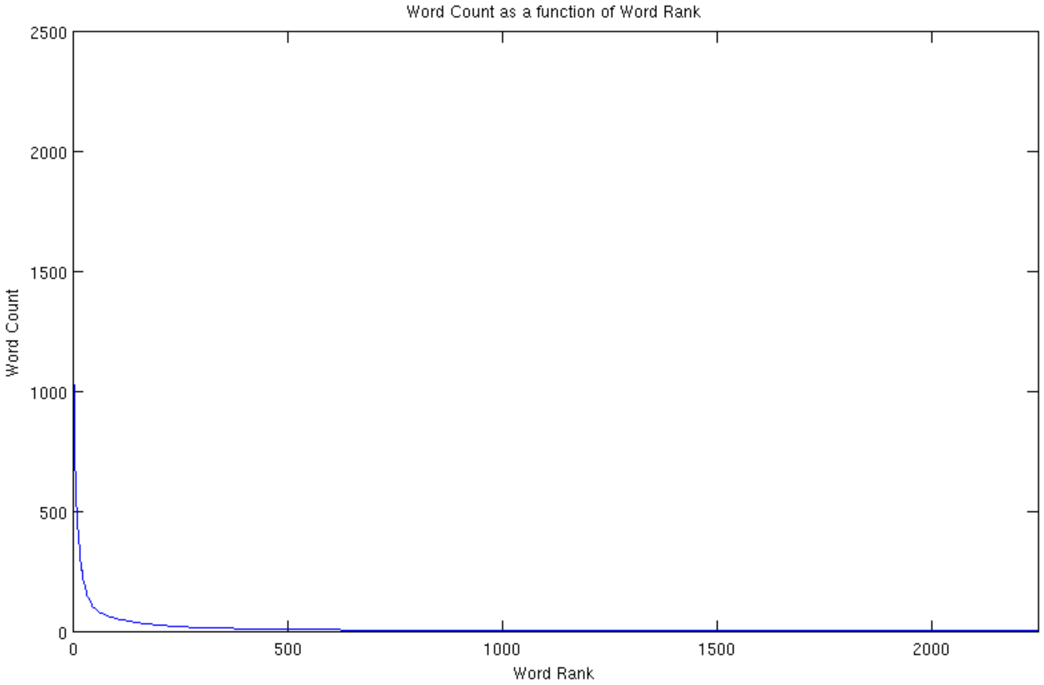
There are 27562 word tokens representing 2245 word types.

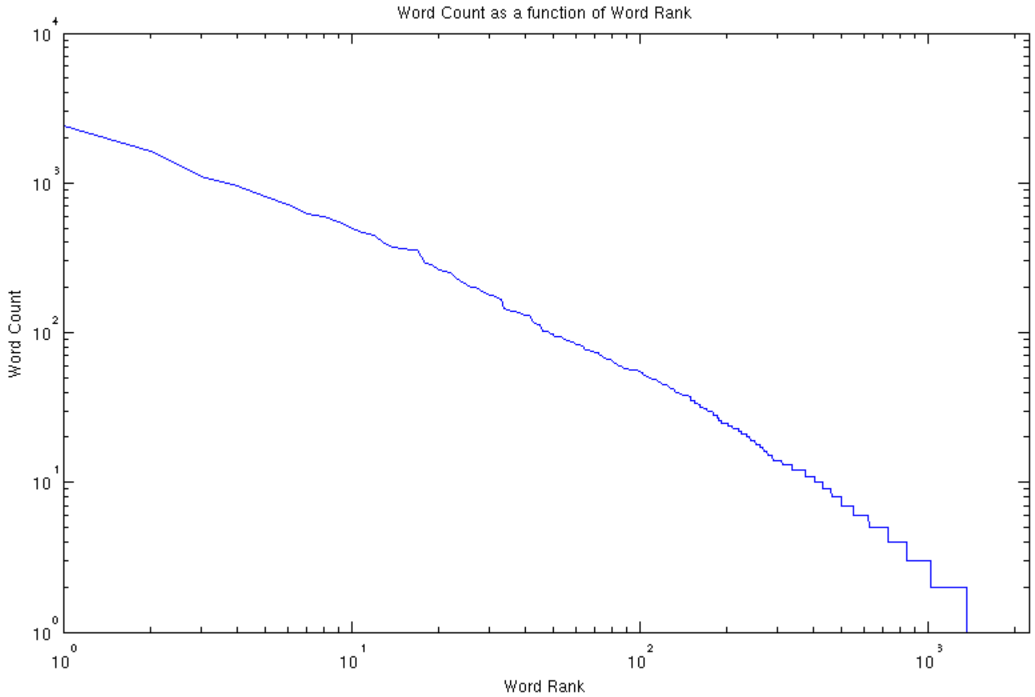
Question 5.2. List the top 10 most frequent words (they can be punctuations) and their counts.

The top 10 most frequent words, and their counts, are:

```
2418 ,
1618 the
1106 '
 961 .
 810 and
 720 to
 620 a
 596 it
 545 she
 499 of
```

Question 5.3. In Matlab, plot rank r (x-axis) vs. count f (y-axis) for all words. Each word would be a dot in such a plot. In a second plot, plot the same thing but use log scale on both axes.





Question 5.4. Assume the following relation: $f = ar^b$. Use Matlab's POLYFIT function to find a, b. Hint: tale log on both sides.

Using the following code segment:

```
polyfit(log(word_rank_scale), log(word_rank_numbers), 1)
= -1.3278  10.0814
```

I was able to determine that the log-log polynomial was of the form:

$$\log(f) = -1.3278\log(r) + 10.0814$$
$$f = 23894.4\, r^{-1.3278}$$