Statistical Natural Language Processing
Assignment 2

Sven Stauden 2549696
Janis Landwehr 2547715
Carsten Klaus 2554140

Exercise 1

1.1

Let A, B be arbitrary sets

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

A and B can be partitioned into

$$A = A \setminus B \cup A \cap B$$
$$B = B \setminus A \cup A \cap B$$

By the definition of probability measures we can write

$$P(A) = P(A \setminus B) + P(A \cap B)$$
$$P(B) = P(B \setminus A) + P(A \cap B)$$

We can assemble the findings above

$$\begin{aligned} & P(A) + P(B) \\ & = P(A \setminus B) + 2 \cdot P(A \cap B) + P(B \setminus A) \\ & = P((A \setminus B) \cup (A \cap B) \cup (B \setminus A)) + P(A \cap B) \\ & = P(A \cup B) + P(A \cap B) \end{aligned}$$

And therefore $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

1.2

First we have to compute the probabilities of the individual events $P(A = a) = \sum_{b \in B} P(A = a, B = b)$

$$
\begin{aligned}
P(X = 0) &= 0.32 + 0.08 = 0.4 \\
P(X = 1) &= 0.48 + 0.12 = 0.6 \\
P(Y = 0) &= 0.48 + 0.32 = 0.8 \\
P(Y = 1) &= 0.08 + 0.12 = 0.2
\end{aligned}
$$

X and Y are stochastically independent if $P(X = x|Y = y) = P(X = x)$ holds.

$$
\begin{aligned}
\frac{P(X = 0, Y = 0)}{P(Y = 0)} &= P(X = 0|Y = 0) = 0.4 = P(X = 0) \\
\frac{P(X = 0, Y = 1)}{P(Y = 1)} &= P(X = 0|Y = 1) = 0.4 = P(X = 0) \\
\frac{P(X = 1, Y = 0)}{P(Y = 0)} &= P(X = 1|Y = 0) = 0.6 = P(X = 1) \\
\frac{P(X = 1, Y = 1)}{P(Y = 1)} &= P(X = 1|Y = 1) = 0.6 = P(X = 1)
\end{aligned}
$$

This consludes that the random variables X and Y are independently distributed.

Exercise 2

2.1

For a language which underlies Zipfan distribution, the occurrence probability of a word w with rank $rank(w)$ can be calculated numerically by

$$P(w) \approx \frac{c}{rank(w)} \tag{1}$$

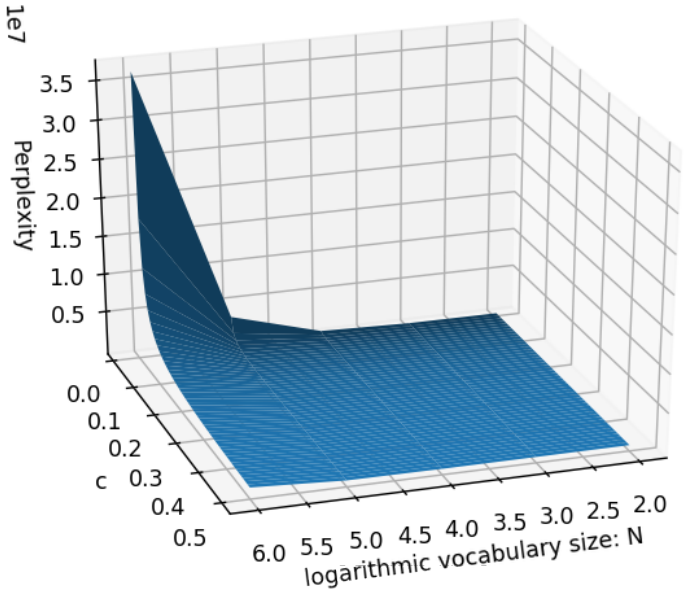
To compute the perplexity for English texts with $c = 0.5$ and a vocabulary size N, we consider the probability words from all N ranks. So we have:

$$PP = exp\left(-\frac{1}{N}\sum_{i=1}^{N} log(\frac{c}{n})\right) \tag{2}$$

The perplexity of $N = 10000$ and $c = 0.1$ is 36808.27.

2.2.

We computed the perplexity after (2) and plotted the output in figure .



One can observe that with increasing vocabulary size and decreasing values for c the perplexity increases.

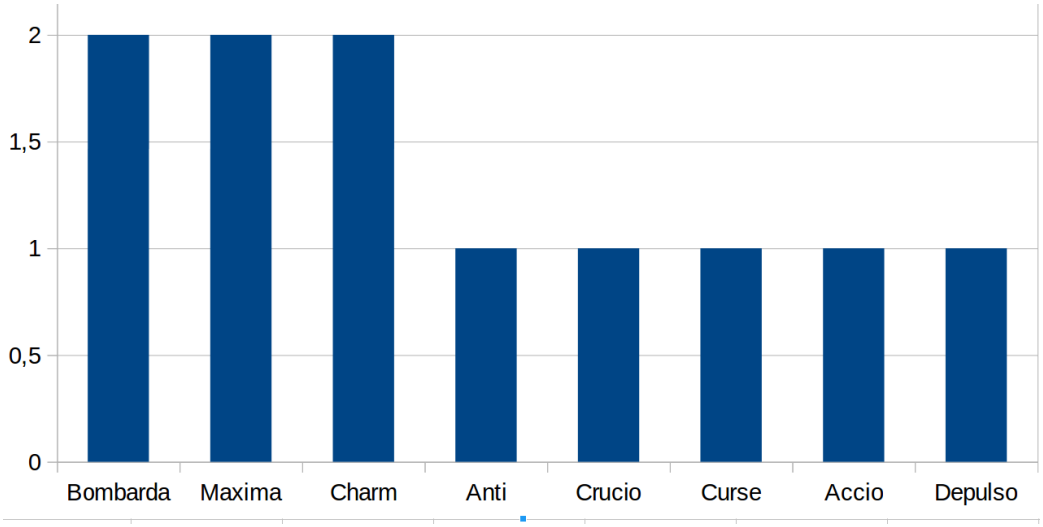
Exercise 3

3.1

First we enumerate all possible magic spells.

- Anti Bombarda Maxima
- Bombarda Maxima
- Crucio Curse
- Accio Charm
- Depulso Charm

As the probability of all magic spells is uniform, we can reconstruct the occurrence probabilities of the words by considering the histogram of words out of which the spells consist. (see figure).



A language obeys Zipf's distribution if for a word rank r holds that

$$P(r) \sim \frac{1}{r}$$

Ranking the words and computing their probability reveals that the distribution obeys Zipf's law but it is hardly to recognize as the considered language consist of to few words.

Spell	Rank r	$P(r)$	$\frac{1}{r}$
Bombarda	1	$\frac{2}{11}$	1
Maxima	2	$\frac{2}{11}$	$\frac{1}{2}$
Charm	3	$\frac{2}{11}$	$\frac{1}{3}$
Anti	4	$\frac{1}{11}$	$\frac{1}{4}$
Crucio	5	$\frac{1}{11}$	$\frac{1}{5}$
Curse	6	$\frac{1}{11}$	$\frac{1}{6}$
Accio	7	$\frac{1}{11}$	$\frac{1}{7}$
Depulso	8	$\frac{1}{11}$	$\frac{1}{8}$

Table 1: The table show that the probability of the rank relates proportionally to the rank

3.2

We compute the perplexity for a language with vocabulary size $N = 8$ and the respective word probabilities out of the occurrences.

$$PP = exp\left(-\frac{1}{8}(3 \cdot log(2/11) + 5 \cdot log(1/11))\right) = 8.482 \tag{3}$$