Statistical Natural Language Processing
Assignment 7

Sven Stauden 2549696
Janis Landwehr 2547715
Carsten Klaus 2554140

Exercise 1

a)

(i)

$$tf(\text{"}Eutopia\text{"}, d_1) = \frac{\#\text{"}Eutopia\text{"}}{\#\text{words in } d_1} = \frac{3}{100} = 0.03$$

(ii)

$$idf(\text{"}Eutopia\text{"}, D) = log_{10}\left(\frac{\#\text{documents}}{\#\text{documents with "Eutopia"}}\right) = log_{10}\left(\frac{10^8}{10^4}\right) = 4$$

(iii)

$$tf{-}idf(\text{"}Eutopia\text{"}, D) = tf(\text{"}Eutopia\text{"}, d_1) \cdot idf(\text{"}Eutopia\text{"}, D) = 0.03 \cdot 4 = 0.12$$

b)

When training a language model, if we use an overly narrow corpus, the probabilities **don't generalize**.

A narrow corpus is restricted on a specific setting without representing boarder characteristics of a language. Training on such a corpus restricts the model in a way so it is not able to generalize to differently styled texts.

c)

To avoid the problem of zero probabilities of trigrams one could adapt a back-off approach for the scoring. Students who strongly orientated on Shakespeare texts might also include known trigrams in their sentences and therefore get a maximum amount of points for a sentence. More probable but less specific Shakespeare terms, which contain known bigrams get a medium scoring. Sentences using Shakespeare specific single words receive some points while sentences which do not contain any Shakespeare specific word are rated with 0. First a sentence is checked for trigrams. If none occurs, one considers the bigrams and afterwards the unigrams.

Exercise 2

```
##### Coffee vs. Not Coffee #####

# Black vs. not black
black <- c(500, 500)
n_black <- c(1000 + 100 + 400, 50 + 1200 + 9450)

chisq.test(data.frame(black,n_black))
# X-squared = 1019.1, df = 1, p-value < 2.2e-16


# Beans vs not beans
beans <- c(1000, 50)
n_beans <- c(500 + 100 + 400, 500 + 1200 + 9450)

chisq.test(data.frame(beans,n_beans))
# X-squared = 5684.5, df = 1, p-value < 2.2e-16


# Leaves vs. not leaves
leaves <- c(100, 1200)
n_leaves <- c(500 + 1000 + 400, 500 + 50 + 9450)

chisq.test(data.frame(leaves,n_leaves))
# X-squared = 61.768, df = 1, p-value = 3.864e-15


# Rest vs. not rest
rest <- c(400, 9450)
n_rest <- c(500 + 1000 + 100, 500 + 50 + 1200)
```

```
chisq.test(data.frame(rest,n_rest))
# X-squared = 3710, df = 1, p-value < 2.2e-16


##### Tea vs. Not Tea #####

# Black vs. not black
black <- c(750, 1000)
n_black <- c(110 + 1300 + 400, 1500 + 200 + 7350)

chisq.test(data.frame(black,n_black))
# X-squared = 637.33, df = 1, p-value < 2.2e-16


#Beans vs not beans
beans <- c(110, 1500)
n_beans <- c(750 + 1300 + 400, 1000 + 200 + 7350)

chisq.test(data.frame(beans,n_beans))
# X-squared = 205.99, df = 1, p-value < 2.2e-16


# Leaves vs. not leaves
leaves <- c(1300, 200)
n_leaves <- c(750 + 110 + 400, 1000 + 1500  + 7350)

chisq.test(data.frame(leaves,n_leaves))
# X-squared = 4629.8, df = 1, p-value < 2.2e-16


# Rest vs. not rest
rest <- c(400, 7350)
n_rest <- c(750 + 110 + 1300, 1000 + 1500  + 200)

chisq.test(data.frame(rest,n_rest))
# X-squared = 2846.4, df = 1, p-value < 2.2e-16
```

Good features result in high chi square values, which also lead to small p-values.
For coffee, the highest chi square value of 5684.5 is reached by the feature *beans*. Therefore, *beans* seems to be the most relevant one for the prediction task. This is follewed by the feature *rest* with a value of 3710. The two least importent ones are *black* and *leaves* with values of 1019.1 and 61.768. While they have significantly smaller chi square values than the first two ones, all features have a p-value smaller 0.05 and thus could in general be seen as suitable for this task.
For tea, the features can be ordered as: leaves (4629.8), rest (2846.4), black (637.33) and beans (205.99). Therefore, the feature leaves is the best one for predicting the class *tea*.

Exercise 3

Plot 1)

In plot 1 the occurrence of both features x and y is **redundant** as the absence of one feature does not lead to a loss of information relevant for the clustering.

Plot 2)

For plot 2 the feature y is **irrelevant**. Projecting all date onto the y-axis mergers all data so no clustering is possible anymore while projecting all data onto the x-axis allows a good distinction of the data into the clusters.

Data for feature relevance)

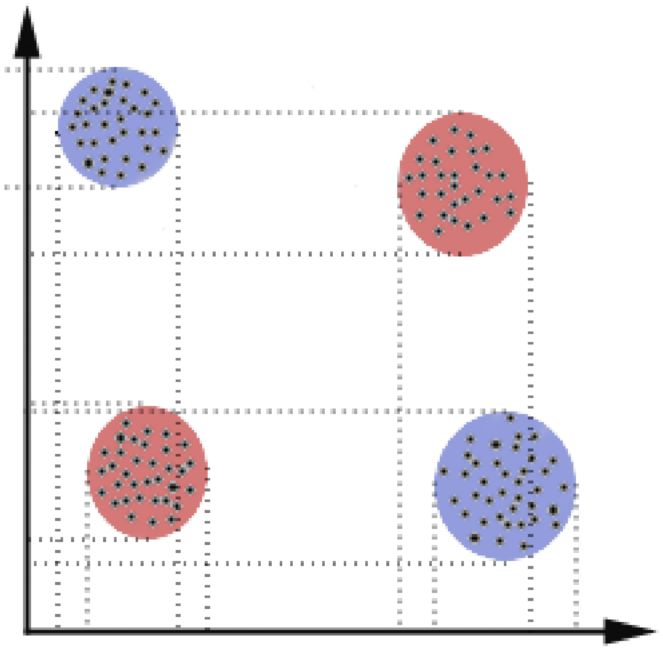

Figure 1: A 4-blob data cluster requires all 2 dimensions for a useful clustering. (We ignore feature transformation for this exercise)

For the given plot [1], both features x and y are relevant for the clustering if one ignores the possibility of data transformation. Discarding one axis would definitely lead to an overlapping of two classes.

Exercise 4

The dimension of the problem has decreased from $|V|$ features to 10 features.

Considering the best features of the PMI, we can observe that the words are not really category specific and therefore not really relevant. We assume that many features have the same PMI value and therefore too many features are not distinguishable by only considering the PMI.

[1] Image from `https://home.deib.polimi.it/matteucc/Clustering/tutorial_html`

PMI - 10 features

```
-- BIOLOGY --
poorwhit
henshaw
townsend
ghastli
erythrocephalu
eager
matan
candidissima
timhali
sportsmanship

-- CHEMISTRY --
mcl
overh
superphosph
luster
pto
sno
azot
foul
graduat
myrban

-- PHYSICS --
goh
upright
vw
conduc
ceaseth
venu
viewd
incrass
pin
milder
```

Mutual Information - 10 best features

The mutual information returns very category characteristic values which might be very good features for category classification. In contrast to the PMI, the MI averages over all possible events which leads a better consideration of all data.

```
-- BIOLOGY --
game
bird
wild
water
kill
state
compound
protect
air
year

-- Chemistry --
acid
oxygen
carbon
form
hydrogen
oxid
element
prepar
solut
sulphur

-- Physics --
colour
ray
refract
light
prism
reflect
glass
```

```
distanc
part
ring
```