Sanna Tiirikainen, 014022498
Introduction to digital humanities
Hands-on assignment

1. Research question and corpus

I made a little research on the most common words of Hungarian love folk songs, and their collocates. I wanted to find out what kind of themes the songs include.

My corpus consists of 77 songs from a folk song collection *A magyar népköltész gyöngyei – A legszebb népdalok gyűjteménye* collected by Elek Benedek (1896). The collection is found as an e-book at http://mek.oszk.hu/11000/11002/11002.htm – it is a part of the Gutenberg project. I chose the songs from the *Boldog szerelem* (”happy love songs”) section.

2. Preprocessing the corpus

Since the corpus is normal book text, I needed to preprocess it some ways to make it easier to utilize. First, I copied the songs to an Excel file, one line in one cell, one song in one column. I removed the punctuation using the ”find and replace” tool. With that I also removed all the extra white spaces.

Then I used Google Refine to put all the text in lower case and trim the leading and trailing whitespaces. I explored the corpus with the cluster tool, as well, because I wanted to know if there were many duplicate lines in the songs. I wanted to remove the duplicate lines because I thought it would give a more realistic picture of the word frequency, and they also included the name of the song. This can be discussed, though, because the more a line or word is repeated in a song, the more important it probably is. With Google Refine, though, I could not use the clustering tool with the data in this form because it removes the rows in all columns.

I decided that it is safer to remove the duplicate lines with Excel, if several songs include similar lines. So, I put the lines (in every song) in alphabetical order, so it was easy to pick out the similar lines. Then I put all the songs into a txt file.

The next step was to preprocess my corpus with R.
I read the corpus in:

```
> docs <- Corpus(DirSource("C:/Users/Sanna/Desktop/szerelem"))
```

And installed some packages and libraries (also to use later):

```
> install.packages(”tm”, ”SnowballC”, ”wordcloud”, ”RcolorBrewer”)
> library(”tm”)
> library(”SnowballC”)
> library(”wordcloud”)
> library(”RcolorBrewer”)
```

Since R's tm package has a Hungarian stopword list, I removed the stopwords with the next code:

```
> docs <- tm_map(docs, removeWords, stopwords("hungarian"))
```

With `fix(docs)` I could save the new text file.

I then visualized the data with Voyant Tools at http://voyant-tools.org/tool/Links, but I noticed, the stopword list doesn't include all stopwords, and some inappropriate words were added to my collocate graph, such as *ha* (if) and *is* (also). I removed the common stopwords that were disturbing the analysis with the next R code:

```
docs2 <- tm_map(docs2, removeWords, c("ha", "is", "hát", "mind"))
```

Now my corpus is preprocessed.

3. Visualization

I wanted to get an image of my corpus by making some word clouds and collocate clusters. First I took a look at the most common words with the next code in R:

```
> dtm <- TermDocumentMatrix(docs2)
> m <- as.matrix(dtm)
> v <- sort(rowSums(m),decreasing=TRUE)
> d <- data.frame(word = names(v),freq=v)
> head(d, 10)
```

I resulted with the following table of the 10 most frequent words (the English equivalents added afterwards):

```
             word freq
rózsám     rózsám   38      → my rose
kis           kis   24      → little
barna       barna   22      → brown
piros       piros   22      → red
édes         édes   21      → sweet
lány         lány   21      → girl
leszek     leszek   17      → I will be(come)
három       három   16      → three
galambom galambom   15      → my dove
rózsa       rózsa   14      → rose
```

Then I ran the next code in R to make a wordcloud of max 50 most common words of the corpus:

```
> set.seed(1234)
> wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.words=
+   random.order=FALSE, rot.per=0.35, colors=brewer.pal(8, "Dark2"))
```

(More information about preprocessing and visualizing with R at http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know and https://rstudio-pubs-static.s3.amazonaws.com/31867_8236987cf0a8444e962ccd2aec46d9c3.html.)

Similar kinds of word clouds can be made also with other tools, such as Voyant Tools: http://voyant-tools.org/tool/Cirrus/?corpus=1452354554921.9113&docIndex=0&docId=d1452316137660.e31e490c-ea99-59d9-59ea-4d0f89b9180f.
I also used the Voyant Tools Links to reveal some collocate clusters, i. e. the most common words to appear together: http://voyant-tools.org/tool/Links/?corpus=1452354381234.3179.

4. Analyzing

The ”problem” with the Hungarian language is that it is agglutinative, and the words are often inflected, so it is hard to get a picture of the most common lemmas (a word and all its different forms) of the corpus. Therefore the visualizations are not very highly informative.

Since I couldn't find any good tokenizator to get the words to their basic form, I decided to look deeper in the data to find out about word clusters of the lemmas that seem to be most common ones. More specifically, in case of adjectives, I wanted to find out which words they modify; when it came to nouns, I wanted to know what kind of modifiers they have.

For this purpose I used AntConc's Clusters/N-Grams tool. I looked for all forms of the words with regular expression, for example for 'girl' in all grammatical cases I used the regex [a-z]*l[e]?ány[a-z]* ('girl' can be *lány* or *leány*, and it can have suffixes in the end, and modifiers in the beginning, e.g. *kislány*, little girl).

For the 100 most common word forms, see the file ”top 100 words”, based on http://voyant-tools.org/tool/CorpusTypeFrequenciesGrid/?corpus=1452354554921.9113&query=. However, the list's top 10 words are a little different that the 10 ones R gave. I am going to use the Voyant's word list from now on.

I counted all the word forms of the seemingly most frequent lemmas with AntConc, and got the following results:

rózsa – rose	73
kis – little	26
barna – brown	27
piros – red	23
édes – sweet	31
lány – girl	40
lesz – will be	17
te – you	17*
három – three	17

* this word's inflected forms are hardly possible to find with a single regular expression, each should be found individually. Also, it is not clear which forms are actually forms of pronouns and which are forms of adverbs.

Then I looked for the most common associations, or collocates, for these words to see if they match with the collocate cluster I was given by Voyant Tools. I got the following data:

___ rózsa (___ rose)

rózsa	(rose)	6
bazsa	(peony)	6
te	(you)	3
édes	(sweet)	3

kis ___ (little ___)

lány	(girl)	12
angyal	(angel)	2
kalap	(hat)	2

barna ___ (brown___)

kis	(little)	4
szerető	(lover)	4

baba	(baby)	2
fattyú	(swan)	2
lány	(girl)	2

piros ___ (red ___)

alma	(apple)	5

édes ___ (sweet ___)

méz	(honey)	3
rózsa	(rose)	3
álom	(dream)	2
anya	(mother)	2
ló	(horse)	2

___ lány (___ girl)

kis	(little)	14
szép	(beautiful)	5
barna	(brown)	2
szőke	(blond)	2

___ lesz (will be ___)

barna	(brown)	2
fehér	(white)	2
piros	(red)	2
sárga	(yellow)	2

te___ (you ___)

rózsa	(rose)	3

három ___ (three ___)

csillag	(star)	2
kis	(little)	2
piros	(red)	2
szerető	(lover)	2

With the collocations, I get the picture that the songs are created around rather small groups of words – this may refer to repeating themes in the songs. The collocates I found with AntConc seem to follow the same pattern as the clusters and wordclouds.

Rózsa, or 'rose', is most certainly the most common word in these songs. It partly supports the results I got in my Bachelor's thesis ”Kukka-aiheiset sanat matyó-kansanryhmän kansanlauluissa” (2014), where I researched the different flower vocabulary in some Hungarian folk songs: about 25 % of the songs included flower related words, 40 % of which were 'rose'.

The words occuring in the collocate cluster, wordcloud and among the most common associations are: *rózsa* (rose), *te* (you), *édes* (sweet), *méz* (honey), *álom* (dream), *kis* (little), *lány* (girl), *barna* (brown), *szerető* (lover), *lesz* (will be(come)), *piros* (red).

The top frequency words and their collocates suggest that in this corpus, the Hungarian love folk songs

1) are rather romantic and positive (rose, dream)
2) include quite a lot of cherishing nicknames (rose, you)
3) tell stories about people (girl, lover)
4) describe the looks and nature of people (sweet, little, brown, red)

These results have connections to my findings on my Bachelor's thesis as well (where I did a lot of close-reading).