**Introduction to Corpus Linguistics**

WiSe 2018-2019

by Dr. Natalia Levshina ©

**Session 7. Visualization of corpus frequencies**

The goals of our today’s session is to learn how to visualize the frequencies of words and any other units in a corpus.

**Data**

We’ll use a corpus of recipes from the BBC website <https://www.bbc.com/food/recipes>. The file is *recipes.txt*. Please save it locally from GitHub.

**R practice**

**1. Open the corpus in R.**

recipes <- scan(file = file.choose(), what = "character", sep = "\n", comment.char = "#", encoding = "UTF-8")

head(recipes)

#output omitted

**2. Now we need to parse the corpus with UDPipe.**

library(udpipe)

ud\_eng <- udpipe\_load\_model("english-ud-2.0-170801.udpipe") #change the name of the model if you work with another language!

Note: if you work with another language, and don’t have the parser in your directory, you need to download it first. For example:

ud\_eng <- udpipe\_download\_model(language = "english")

ud\_eng <- udpipe\_load\_model(ud\_eng)

recipes\_ud <- udpipe(x = recipes, object = ud\_eng)

head(recipes\_ud)

#output omitted

dim(recipes\_ud)

#3593 17

This means we have 3593 tokens (including punctuation) and 17 columns with different information. Which of these columns have we already used?

colnames(recipes\_ud)

[1] "doc\_id" "paragraph\_id" "sentence\_id" "sentence"

[5] "start" "end" "term\_id" "token\_id"

[9] "token" "lemma" "upos" "xpos"

[13] "feats" "head\_token\_id" "dep\_rel" "deps"

[17] "misc"

**3. How to create a bar chart (or bar plot).**

Let us extract all common nouns in the corpus and compute their frequencies:

recipes\_nouns <- recipes\_ud$lemma[recipes\_ud$upos == "NOUN"]

head(recipes\_nouns)

#[1] "melon" "piece" "bowl" "finger" "shape" "juice"

recipes\_nouns\_freq <- table(recipes\_nouns)

sort(recipes\_nouns\_freq, decreasing = TRUE)[1:20]

recipes\_nouns

minute pan bowl oil soup oven mixture chicken

36 25 20 20 19 17 16 15

heat pepper salt water egg garlic beetroot onion

14 14 13 13 11 11 10 10

leave sugar baking boil

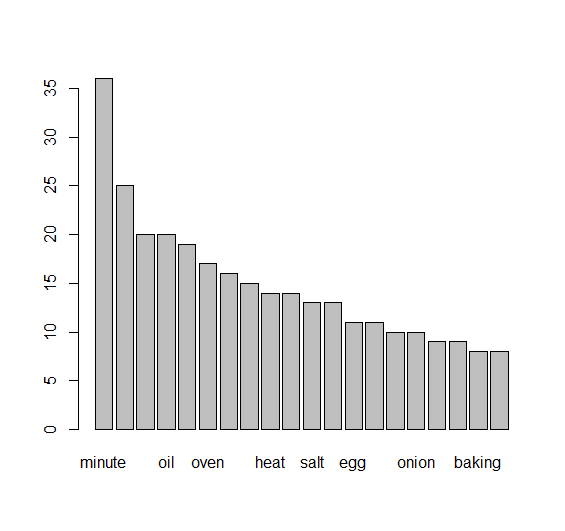
9 9 8 8

Let us save the top 20 nouns as a separate object called top20:

top20 <- sort(recipes\_nouns\_freq, decreasing = TRUE)[1:20]

Now we can make a bar plot:

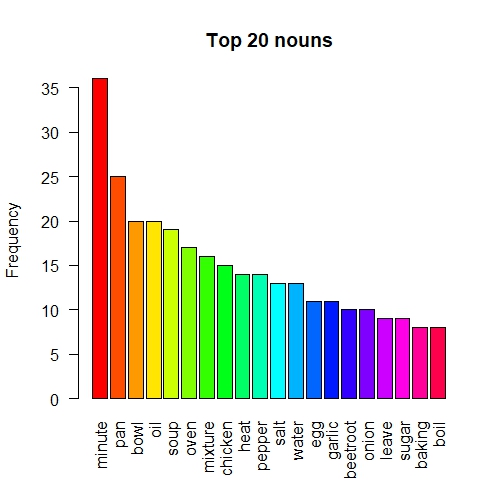
barplot(top20)



This doesn’t look very useful. Let us make an enhanced version:

barplot(top20, las = 2, col = rainbow(20), main = "Top 20 nouns", ylab = "Frequency")

The argument las specifies the orientation of the text labels (2 means orthogonal to the axis). The argument col specifies the colour. You can also use a specific colour, e.g. col = "blue". Next, main specifies the plot title, and ylab the label for the y-axis (i.e. the vertical axis). If you need to specify the label of the horizontal, x-axis, you can use the argument xlab, e.g. xlab = "Words".



Important: how to save the plot? In RStudio, go to the Plots panel, click on *Export > Save as Image*, then choose the name, format and size and click on “Save”. The file will be in your working directory.

**Exercise**

Make a bar plot with top 10 verbs.

**An important note**

Recipes often use verbs in the imperative. The imperative mood is specified in the feats column. For example, Mood=Imp|VerbForm=Fin. You can select all imperative forms as follow:

recipes\_imper <- recipes\_ud$lemma[grep("Mood=Imp", recipes\_ud$feats)]

This is how you can extract all other features, e.g. plural nouns (Number=Plur) or indefinite articles (Definite=Ind).

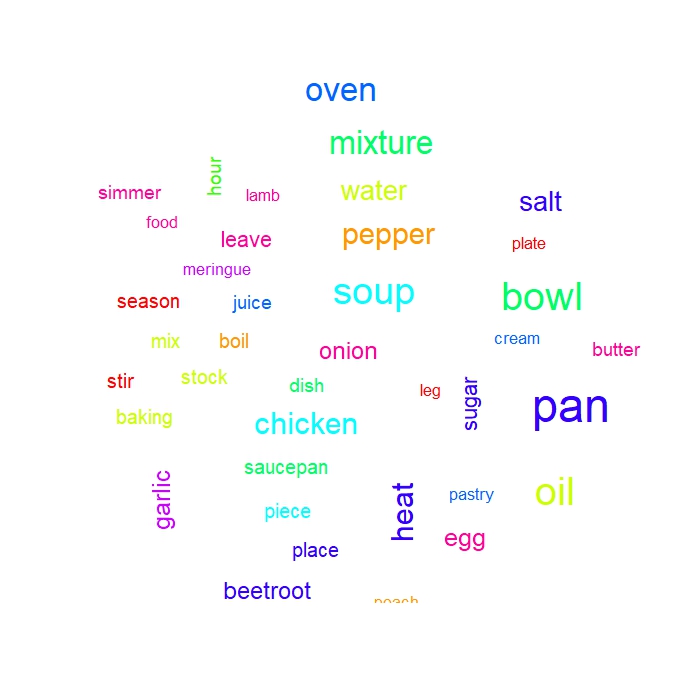
**4. How to make a wordcloud**

If you want to visualize more words, it is convenient to use wordclouds. This is how you can create a wordcloud for top 40 nouns.

library(wordcloud) #you need to install it first!

top40 <- sort(recipes\_nouns\_freq, decreasing = TRUE)[1:40]

wordcloud(words = names(top40), freq = top50, colors = rainbow(40), random.color = TRUE)



**Exercise**

Make a word cloud with the adverbs – those that are more frequent than 2.

Tip: you might need the following expression:

recipes\_adv\_select <- recipes\_adv\_freq[recipes\_adv\_freq > 2]

Can you make sense of the frequencies?

**5. How to make a pie chart.**

When there are few items, it can be useful to represent their frequencies with the help of a pie chart. Let us focus on parts of speech.

pos\_freq <- table(recipes\_ud$upos)

#ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM PART PRON PROPN PUNCT

#181 426 152 43 222 466 2 980 92 54 40 25 433

#SCONJ SYM VERB X

#38 18 419 2

Let us focus on the main notional categories: adjectives, adverbs, nouns and verbs, and create a vector with their frequencies manually:

pos\_freq\_main <- c(181, 152, 980, 419)

names(pos\_freq\_main) <- c("Adjective", "Adverb", "Noun", "Verb")

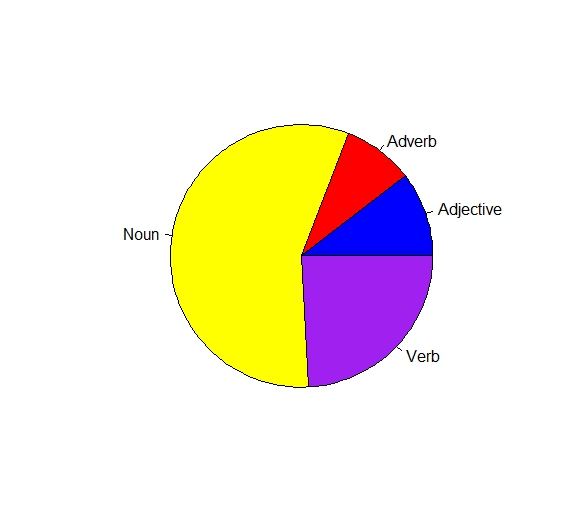
pos\_freq\_main

#Adjective Adverb Noun Verb

#181 152 980 419

Now we can create a pie chart. We will use manually defined colours:

pie(pos\_freq\_main, col = c("blue", "red", "yellow", "purple"))



**Exercise**

Make a pie chart with arguments (subjects, objects and obliques). Tip: check the column called *dep\_rel*. How can you interpret the results?

**Exercise**

Take your own corpus or one of the corpora that we have investigated already. Visualize the frequencies using the functions that you have learnt today and interpret the results.