Word Cloud of the Discomfort Glare Review Paper 2024

Yulia Tyukhova

October 7, 2024

This document consists of:
1. Introduction
2. Procedure
3. Results
4. Sources

1. Introduction

This is a word analysis of the most frequently used words in the accepted version of a published comprehensive review paper titled **"Discomfort glare in outdoor environments after dark – A review of methods, measures, and models"**[1] excluding supplementary materials in the appendix, references, funding and other information.

Word clouds can be created using text mining methods that allow us to highlight the most frequently used keywords in a **corpus** - a collection of texts.

2. Procedure

2.1 Preparation

1. The accepted version of the manuscript is manually copied, pasted, and saved in a plain text file (.txt).
2. The following packages are installed (move from comments) and loaded (see comments for the purpose).

```
#Install
#options(repos = c(CRAN = "https://cloud.r-project.org/"))
#install.packages("tm")  #text mining
#install.packages("wordcloud") #word-cloud generator
#install.packages("RColorBrewer") #color palettes
#install.packages("textstem") #lemmatizing the words
#Load
library("tm")
library("wordcloud")
library("RColorBrewer")
library("textstem")
library("knitr")
```

3. Interactively choose and import the locally saved text file and use the Corpus() function from text mining (tm) package to load the text.

```
#Choose the locally saved file from your laptop
text <- readLines(file.choose())
#Load the data as a corpus
docs <- Corpus(VectorSource(text))
```

2.2 Cleaning and lemmatizing

Before counting the words and ultimately creating a word cloud[2], there is a number of cleaning steps that have to be completed such as removing numbers, extra white spaces, and so on. Removing **stop words** - common words that carry little or no meaningful information[3] - is also an important step. They can be divided into three groups: global, subject, and document-level stop words (for further information see reference[3]).

During the text preparation stage one can choose to do **stemming** or **lemmatizing**. Essentially, stemming removes a part of the word (e.g. word vector ('are', 'am', 'being', 'been', 'be') –> after stemming ("ar" "am" "be" "been" "be")), while lemmatization transforms the word into its base form using proper grammatical roots (word vector ('are', 'am', 'being', 'been', 'be') –> after lemmatizing ("be" "be" "be" "be" "be"))[4]. I used lemmatizing for normalizing the text.

A note on the word "LED": During the analysis, I noticed that I don't see any "LED" words that is quite central to this particular paper. It became clear that after all words transformed to the lower case and are lemmatized "LED" became "led" that in turn became "lead". I had to differentiate "LED" from "lead" by making up a new word during the analysis ("LEDr") and then returning the correct output after the analysis is done ("LED").

Ultimately, there are a few extra steps that need to be completed after the "standard" text cleaning depending on the peculiarities of your text. For the purpose of a word cloud, I had to eliminate names of the cited authors and a number of words such as "however" that did not add value to my word cloud.

```
#Remove numbers
docs <- tm_map(docs, removeNumbers)
#Remove punctuation
docs <- tm_map(docs, removePunctuation)
#Eliminate extra white spaces
docs <- tm_map(docs, stripWhitespace)
#Before I lemmatize the word 'LED', I need to "save" it, otherwise it becomes 'lead'.
docs <- tm_map(docs, content_transformer(function(x) gsub('LED', 'LEDr', x)))
# Convert the text to lower case (This will ensure stop words working)
docs <- tm_map(docs, content_transformer(tolower))
#Remove common English stop words
docs <- tm_map(docs, removeWords, stopwords("english"))
#Remove your own stop word. Specify your stop words as a character vector.
docs <- tm_map(docs, removeWords, c("however","cdm","others","might","kohko","abboushi",
                                    "bullough", "lin","since","also","bul","bennett","many",
                                    "well","can","waters","two","one","tyukhova","needs",
                                    "based","boer","villa","cie","kent","fotios",
                                    "tashiro","hopkinson","will",'al',"bommel","et"))
docs <- tm_map(docs, content_transformer(lemmatize_strings))

#Additional bugs I am fixing manually.Some punctuation was not removed and some words
#had to be displayed in a desired way (e.g. 'non-uniformity' instead of 'uniformity')
docs<-gsub(" - ", " ", docs)
docs<-gsub('’', '', docs)
docs<-gsub('"",', '', docs)
docs<-gsub('‘', '', docs)
docs<-gsub('"-', '', docs)
docs<-gsub('"discomfort', 'discomfort', docs)
docs<-gsub('nonuniform','non-uniform', docs)
docs<-gsub('nonuniformty','non-uniformty', docs)
docs<-gsub('ledr','led', docs)
docs<-gsub('datum','data', docs)
docs<-gsub('luminaires','luminaire', docs)
```

The next step is to build a table containing the frequency of the words called **document matrix**[2], in which column names are words and row names are documents. The function TermDocumentMatrix() comes from text mining package. I included an additional step of transforming certain words to the upper case.

```
#Build a term-document matrix
dtm <- TermDocumentMatrix(docs)
m <- as.matrix(dtm)

#Define terms to be displayed in uppercase
uppercase_terms <- c("led", "ugr","cbe","bcd")
#Make specific terms uppercase
rownames(m) <- sapply(rownames(m), function(term) {
  if (term %in% uppercase_terms) {
    return(toupper(term))
  } else {
    return(term)
  }
})
```

3. Results

This section shows the frequency and word cloud of the first 100 most used words in the accepted version of the research paper[1].

```
#This is a continuation of building a term-document matrix.
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
d1<-head(d,100)
#d1
```

```
#Simple table
kable(d1,row.names=FALSE, caption = "Word frequency first 100 words")
```

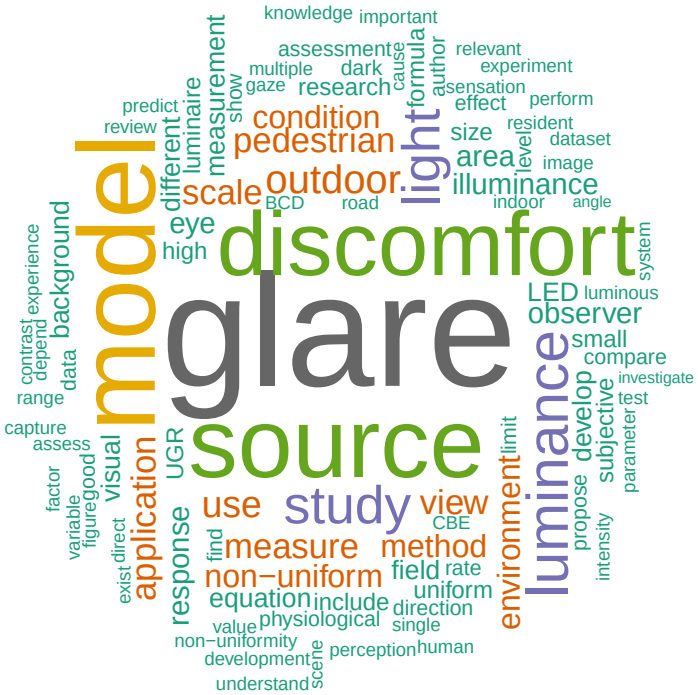
Table 1: Word frequency first 100 words

word	freq
glare	282
model	188
source	165
discomfort	149
luminance	91
study	81
light	80
outdoor	55
use	50
measure	46
view	45
scale	44
application	43
pedestrian	43
non-uniform	40
environment	39

word	freq
method	39
condition	37
observer	35
illuminance	34
response	33
area	32
eye	32
develop	31
different	31
measurement	31
visual	31
LED	30
background	28
field	28
equation	27
include	26
UGR	26
small	25
subjective	25
uniform	25
luminaire	24
size	24
formula	23
research	23
high	22
find	21
level	21
physiological	20
show	20
compare	19
data	19
effect	19
limit	19
propose	19
author	18
dark	18
good	18
assessment	17
direction	17
rate	17
experience	16
luminous	16
multiple	16
predict	16
road	16
assess	15
image	15
non-uniformity	15
parameter	15
resident	15
cause	14
CBE	14

word	freq
depend	14
direct	14
indoor	14
range	14
sensation	14
single	14
capture	13
contrast	13
development	13
factor	13
figure	13
perception	13
review	13
test	13
value	13
variable	13
BCD	12
dataset	12
exist	12
experiment	12
gaze	12
human	12
important	12
intensity	12
knowledge	12
perform	12
relevant	12
scene	12
system	12
understand	12
angle	11
behavioral	11

```
#Word cloud
set.seed(2345)
wordcloud(words = d$word, freq = d$freq, min.freq = 5,
          max.words=100, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"),
          scale=c(5,0.36))
```



4. Sources

1. Tyukhova, Y. 2024. "Discomfort glare in outdoor nighttime environments after dark – A review of methods, measures, and models". Building and Environment Volume 263, 111850, ISSN 0360-1323 https://doi.org/10.1016/j.buildenv.2024.111850

2. Text mining and word cloud fundamentals in R http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know

3. Stop words https://smltar.com/stopwords

4. Stemming vs lemmatizing
https://www.rdocumentation.org/packages/textstem/versions/0.1.4#examples