

Corpus Word Frequency Analysis

Jan Jugueta (jan.jugueta@hdr.mq.edu.au)
Macquarie University, North Ryde, NSW



Figure: Five front pages from *Neues Deutschland*. Adapted from *Neues Deutschland*, by Jan Jugueta, 2019, retrieved from <http://nd-archiv.de/>. Copyright 1974 by Neues Deutschland.



Corpus Word Frequency Analysis

Jan Jugueta (jan.jugueta@hdr.mq.edu.au)
Macquarie University, North Ryde, NSW



		1973-11-13_UnerverantwortlicherEntscheid
Um:		1973-11-14_Australienqualifiziert
Ge: 14;		1973-11-14_Kurzberichtet
Au:		1973-11-15_DiefaischenGarantiertderchenischenJunta
Be: 14;	Ku	1973-11-16_Bulgarienqualifiziert
Au: 15;		1973-11-19_Hollandnach-0GegenBelgienqualifiziert
br: 16;	Au:	1973-11-21_FIFA-Haltungbefremdend
Fu: 17;	Da:	1973-11-22_NunimNepstadionSiegfürNationalelf-10
Gu: WM: 18;		
Ch: Ma:	19;	
Na: Ho:	21;	
Te: Po:		TA: 22/11/1973
Fu: 20;		Nun im Nepstadion Sieg für Nationalelf 1:0
Gu: Du:		Die Entscheidung fiel durch ein Kopfballtor von Lauck
Ma: 21;		von unserem Sonderberichterstatter Max Schlosser
Pa: 22;		
Zu: 23;		Zum 13. Male standen sich am gestrigen Abend die Fußballnationalmannschaften Ungarns und
Fu: 24;		der DDR gegenüber, zum fünftmal dabei im Nepstadion von Budapest. Nach drei Niederlagen
hall: 25;		und einem Unentschieden an gleicher Stelle gelang unserer Auswahl nun der erste Sieg. An
zu: 26;		seiner Berechtigung gibt es keinen Zweifel. Er wurde gegen einen Gastgeber erzielt, der ihn
die: 27;		mit dem festen Willen auf den Rasen kam, um eine Freiheit enttäuschender Spiele, die ihn
Te: 28;		umfasste. Fünf Minuten später folgte die entsprechende Kostprobe dieser Partie, die ihn
Fu: 29;		zu einer Qualifikation für die Weltmeisterschaft kostete.
Gu: 30;		Der Punkt, den die ungarische Mannschaft erzielte, kostete dieser Partie, die ihn
Ma: 31;		an mancher Stelle berücksichtigt, dass nach dem Ausfall von Kreische nun mit Croy, Löwe und Streich noch
Na: 32;		drei weitere Stammspieler unserer Nationalmannschaft ersetzt werden mussten und mit den
Ho: 33;		de: 34; den Magdeburgern Tyll und Hoffmann zwei unserer Talente debütierten.
Ch: 35;		Unseren „Neuen“ soll an dieser Stelle ein erstes Lob gewidmet werden. Es impunierte, wie
Te: 36;		er couragierte sich an ihrer Aufgabe gingen. Insbesondere der erst 18-jährige Martin Hoffmann
Fu: 37;		war bei seinem Debüt eine wahre Größe. Mit einer tollen Mitte ins Netz
hall: 38;		„König“ Laupusinski und seine Schnelligkeit löste er nun als einsam in der gegenüberliegenden
zu: 39;		Hälfte Alarm aus. Da ihm noch nichts gelang und er in manchen Zweikämpfen gegen die an
die: 40;		internationaler Erfahrung durchweg reicher ungarischen Abwehrspieler noch den kürzeren
Te: 41;		zog, wird ihm niemand ankreiden. Auch Axel Tyll löste die ihm übertragenen Aufgaben gut,
Fu: 42;		obwohl er allerdings in einigen Fällen zu risikovoll und ungenau abspielte.
Gu: 43;		Die nächsten Minuten waren gebannt. Der strategische „Julian“ Peter Dücke bestritt sein 60.
Ma: 44;		und wie er das tat, beeindruckte mich sehr. Wie Götzenwald hatte, die
Na: 45;		Originalübertragungen in ungarischen Rundfunk und Fernsehen zu verfolgen, der hörte immer
Ho: 46;		wieder die „angehobenen“ Stimmen der Reporter, wenn der Jenner bei unsren
Ch: 47;		Angriffsaktionen im Ballbesitz kam. Hervorragend seine Vorbereitung des
Te: 48;		spielsentscheidenden Treffers.
Fu: 49;		Schließlich soll - ohne die Leistung der anderen zu schmälen - auch noch Wolfgang

Figure: A corpus of *Neues Deutschland* articles in .txt format. Adapted from Neues Deutschland, by Jan Jugueta, 2019, retrieved from <http://nd-archiv.de/>. Copyright 1974 by Neues Deutschland.

Body Word Freq	Title World Freq	Title Subtitle Word Freq
	word	freq
wahl	qualifiziert	3
allweltmeisterschaft	entscheid	1
l	unverantwortlicher	1
de	australien	1
erischen	berichtet	1
e	kurz	1
nationalen	chilenischen	1
er	falschen	1
nenischen	garantien	1
a	junta	1
zentralslager	bulgarien	1
r	belgien	1
en	holland	1
ern	befremdend	1
its	fifahaltung	1
geber	nationalelf	1
	neustadion	1
	sieg	1

Figure: Results from Body, Title and Title + Subtitle word frequency analysis. Screenshot by Jan Jugueta.



Corpus Word Frequency Analysis Workflow



ND Logo Copyright
Neues Deutschland

Icon made by
Freepik from
www.flaticon.com

CC Image courtesy of
www.kissclipart.com

CC Image courtesy of
www.wikimedia.org

Icon made by
Freepik from
www.flaticon.com

Corpus Word Frequency Analysis automates the process from extracting the correct data from the .txt files to creating the .csv files. The user will still need to save their corpus in a .txt file format to be compatible with the software.



**Nun im Nepstadion
Sieg für Nationalelf 1 : 0**

Die Entscheidung fiel durch ein Kopftor von Leudk



Peter Duda (dunkles Trikot, Bildmitte) touchte wiederholt gefährlich vor dem ungarischen Tor auf. Hier sind er von Balint (Nummer 3) gestoppt.

Figure: Article 'Nun im Nepstadion Sieg für Nationalelf 1:0' from *Neues Deutschland*, retrieved from <http://nd-archiv.de/>. Copyright 1973 by Neues Deutschland.

1973-11-22_NunimLeopstadionSiegfürNationalfeiert!-0 ~

Nun im Neugstadion Sieg für Nationalfeiert 1:0
Die Entscheidung fiel durch einen Kopftreffer von Lauch
und die Schiedsrichterurteile von Schüssler und Schässner

Zu 13. Mal standen sich an gestrigen Abend die Fußballnationalmannschaften Ungarns und der DDR gegenüber, zum fünfzehnten, dabei im Neugstadion von Budapest. Nach drei Niederlagen und einem Unentschieden an gleicher Stelle gelang unserer Auswahl nur der erste Sieg. An seiner Berechtigung gibt es keinen Zweifel. Er wurde gegen einen Gastgeber erzielt, der mit seinem Aufgebot nicht die Qualifikation für die Fußball-Weltmeisterschaften kostete, mit dieser Partie die Periode des Haubauabschnitts erfolgreich zu beginnen. Und das Lied gewinnt noch an Gewicht, wenn man bedenkt, daß die ungarische Mannschaft aus einer Gruppe bestand, die nach strengem Schema nach drei Auswärtsspielen eine nationale Mannschaft ersetzte, was weiter unten und bei den Hauptschreibern Tilly und Hoffmann zwei unserer Talente debütierten.

Unserer „Jesuit“ war an dieser Stelle ein erstaunliches Lobj geworden. Es ließgestern, wie er sich selbst ausdrückte, eine „große Freude“ über den Sieg und über den ungarischen Martin Hoffmann, widmete auch unsere ungarischen Journalistkollegen große Anerkennung. Mit seinem ungewöhnlichen Laufstil und seiner Schnelligkeit löste er mehr als einmal in der ungarischen Presse Aufsehen. Er war ein großer Star, der auf dem Platz seine Fertigkeiten an der internationalen Erfahrung durchweichen konnte. Ein ungarischer Abwehrspieler noch den kurzen zog sie gegen ihn Nämlich ankreiden. Auch Axel Tilly löste die im Übertragenen Aufgaben gut, wobei er seine technischen Fertigkeiten und seine schnelle Reaktionen beweisen konnte.

Die nächste Anerkennung gebührte unserm gestrigen „Jubilar“, Peter Dücke bestreit sein 100. Länderspiel. Und wie er das tat, beeindruckt einmal mehr. Nur Gelegenheit hatte, die einzigen Übertragungen in ungarischer Handlung und Farbe zu verfolgen, der „Duke“ war immer der Atem im Geschehen. Seine Leistung war der Voraussetzung der Tengely-Mannschaft in Ballbesitz kam. Herrenvorgang seine Vorbereitung des spielsuchenden Treffers. Schließlich soll, wie die Leistung der anderen zu schmälen - auch noch Wolfgang Blöschwitz und Rainer Kühn - der Torhüter von Carl Zeiss Jena eine seiner bisher besten Leistungen in der Auswahl.

Dieses Pauschallob soll nicht der Eindruck verursachen, als müsse man an der gestrigen Partie nur die Leistung der eigenen Mannschaft bewundern. Das ist nicht so. Es mußte nicht mehr unserer Konzeption zu entsprechen. Auch müßten wieder zu viele Fehlerlabials, die von der ungarischen Auswahl meist sofort zu Gegenangriffen auf unsrer Tor genutzt wurden, registriert werden. Aber es gab auch hier kein Kapital an Spielpläne und Taktiken. Deutlich, wie sie nach einem ersten Direktwettkampf gegen die ungarische Mannschaft (1:1, Kosciusko, Fazekas, Nagy) oder aber die eigene Abwehr am Bransch und Blöschwitz mit großem Einsatz das gefährlichen Situationen noch bereinigen konnte.

Die ungarische Mannschaft hat sich in der zweiten Hälfte durch spielerische Akzente zu setzen - das sind für Cheftreiniger Buschburger und seine Schützlinge sicher zwei der wichtigsten Aufgaben, deren Lösung im Hinblick auf die Weltmeisterschaftsendrunde in Angriff zu nehmen.

In einer wenig benidenswerten Lage sind die Verantwortlichen des ungarischen Fußballs. 1980 Zuschauer im Neugstadion, das man aus früheren Jahren bei Länderspielen nur als überfüllt bezeichnete, waren kaum 10000. Ein gutes Spieler und Talentes fehlt es nicht, auch wie vor nicht. Aber als Mannschaft ist die ungarische Auswahl herzhaft von der europäischen Spalte ins Mittelmaar zurückgefegt worden. Ungars: Balázs, Kórus, Károly, Harazsanzi, Balázs, Hegedűsi - 1. Juhász, Kosciusko, Fazekas (64. László), Nagy, Béla, Kocsis (84. Stochel). Blöschwitz - B. Bransch, Weisse, Kühn, Blöschwitz; Tilly, Lauch, Dücke, Hoffmann, Schiedsrichter: Beck, Kutscherski.

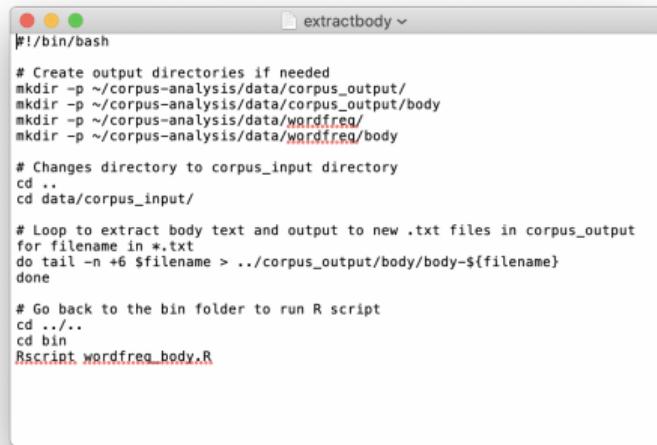
Figure: Article converted to .txt format. Date, title, author and body text saved on specific lines. Copyright 1974 by Neues Deutschland. Screenshot by Jan Jugueta.



Bash Unix Shell Script

Bash Unix Shell scripts have been written to extract *specific* information from the corpus. This allows for the textual analysis not to be skewed by irrelevant data. It also keeps the original .txt file unaltered. The Bash scripts:

- ▶ Extract body text only
- ▶ Extract title text only
- ▶ Extract title and subtitle text only



```
#!/bin/bash

# Create output directories if needed
mkdir -p ~/corpus-analysis/data/corpus_output/
mkdir -p ~/corpus-analysis/data/corpus_output/body
mkdir -p ~/corpus-analysis/data/wordfreq/
mkdir -p ~/corpus-analysis/data/wordfreq/body

# Changes directory to corpus_input directory
cd ..
cd data/corpus_input/

# Loop to extract body text and output to new .txt files in corpus_output
for filename in *.txt
do tail -n +6 $filename > ../corpus_output/body/body-${filename}
done

# Go back to the bin folder to run R script
cd ../../
cd bin
Rscript wordfreq_body.R
```

Figure: Screenshot of Bash Unix Shell script in .txt file format. Screenshot by Jan Jugueta.



R for Textual Analysis

The word frequency textual analysis is powered by R: a programming language used by researchers to interpret and manipulate data sets. The RScript was coded to:

- ▶ Remove stop words
 - ▶ Remove punctuation
 - ▶ Remove numbers
 - ▶ Convert all words to lower case
 - ▶ Count the frequency of words
 - ▶ Order them from most frequent to least frequent
 - ▶ Create a .csv file with the results

```
wordfreq_body.R | Source on Save | Run | Source | ...  
1 # Package and library required for this work  
2 if(require(tm)){  
3 install.packages("tm",repos = "http://cran.us.r-project.org")  
4 library(tm)  
5 }  
6 # The file path to where the .txt files are stored  
7 folder <- "./corpus-analysis/data/corpus_output/body"  
8 # Lists the files in the path  
9 filelist <- list.files(path=folder, pattern=".txt")  
10 # Creates a path to all .txt files in path folder  
11 filelist <- paste(folder, "/", filelist, sep="")  
12 # Reads text from .txt files  
13 readtext <- lapply(filelist, FUN = readLines)  
14 # Collapses elements into one element  
15 corpus <- lapply(readtext, FUN = paste, collapse=" ")  
16 # Convert corpus to something tm package can use  
17 VCorpus <- Corpus(VectorSource(corpus))  
18 # Using a new corpus name <VCorpus_clean> to keep the original VCorpus object untouched  
19 # Strip whitespace from corpus  
20 VCorpus_clean <- tm_map(VCorpus, stripWhitespace)  
21 # Convert corpus to lower case  
22 VCorpus_clean <- tm_map(VCorpus_clean, content_transformer(tolower))  
23 # Remove German stop words  
24 VCorpus_clean <- tm_map(VCorpus_clean, removeWords, stopwords("german"))  
25 # Remove punctuation  
26 VCorpus_clean <- tm_map(VCorpus_clean, removePunctuation)  
27 # Remove numbers  
28 VCorpus_clean <- tm_map(VCorpus_clean, removeNumbers)  
29 # Remove "-" and "--" from corpus  
30 VCorpus_clean <- tm_map(VCorpus_clean, removeWords, c("-", "-"))  
31 # Create Term-Document Matrix  
32 dtm <- TermDocumentMatrix(VCorpus_clean)  
33 # To get the list of frequency of words  
34 m <- as.matrix(dtm)  
35 v <- sort(rowSums(m), decreasing=TRUE)  
36 d <- data.frame(word = names(v), freq=v)  
37 # Output dataframe as a .csv file  
38 write.csv(d, file=paste0("./corpus-analysis/data/wordfreq/body/bodywordFrequency-", Sys.time(), ".csv"), row.names=FALSE)  
39
```

Figure: Screenshot of R script used in the RStudio environment. Screenshot by Jan Jugueta.



Results in .csv format

At the end of the automated process, Corpus Word Frequency Analysis will create .csv files with the frequency information. It will also timestamp the creation of the .csv onto the filename itself so that .csv files are never overwritten. The researcher can then import the .csv into other programs such as

- ▶ OpenRefine
- ▶ RStudio
- ▶ Excel

Body Word Freq		Title Word Freq		Title Subtitle Word Freq	
word	freq	word	freq	word	freq
fifa	8	qualifiziert	3	qualifiziert	3
auswahl	6	entscheid	1	brüskiert	1
fußballweltmeisterschaft	5	unverantwortlicher	1	entscheid	1
spiel	5	australien	1	fifa	1
wurde	5	berichtet	1	generalsekretariat	1
ungarischen	5	kurz	1	unverantwortlicher	1
adn	4	chilenischen	1	weltöffentlichkeit	1
chile	4	falschen	1	australien	1
ddr	4	garantien	1	berichtet	1
internationalen	4	junta	1	kurz	1
länder	4	bulgarien	1	chilenischen	1
tor	4	belgien	1	falschen	1
chilenischen	4	holland	1	garantien	1
junta	4	befremdend	1	junta	1
konzentrationslager	4	fifahaltung	1	bulgarien	1
udssr	4	nationalelf	1	belgien	1
treffen	4	nepstadion	1	holland	1
zypern	4	sieg	1	augen	1
bereits	4			befremdend	1
gastgeber	4			bekannten	1

Figure: Screenshot of R script used in the RStudio environment. Screenshot by Jan Jugueta.



Main Points

- ▶ Corpus Word Frequency Analysis is a simple, yet *quick* way to find out the most frequently used words in a corpus.
- ▶ It can run on all major operating systems, Windows, Mac OSX and Unix.
- ▶ Corpus Word Frequency Analysis utilises the programming languages of R and Bash and its code is available as Open Source.
- ▶ You can download Corpus Word Frequency Analysis from
<https://github.com/MQ-FOAR705/jugueta-corpus-analysis>