

Text Analysis and Word Vectors in R Workshop at Goethe University Frankfurt

Sara J Kerr M.A.*

25th January 2017

*PhD Candidate at An Foras Feasa, Maynooth University, Ireland.
Supported by a Pat and John Hume Scholarship.

Contents

1	Workshop Session One: Text Analysis in R	2
1.1	Load and Preprocess a File	3
1.2	Tokens, Types and Word Frequency	5
1.3	Exploring Word Frequency Across a Novel	6
1.4	Handling Multiple Files and Key Word in Context	7
1.5	Using the 'tm' Package	9
2	Workshop Session Two: Word Vectors in R	12
2.1	What is a Vector Space Model and How is it Created?	12
2.2	Create, or Load, a Vector Space Model	13
2.3	Prepare Texts for KWIC Analysis	14
2.4	Exploring a Vector Space Model	15
2.5	Using the w2v_analysis Function	16
2.6	Some Resources on using word2vec with Asian Texts	18
3	Contact Details	19

Overview

This workshop will provide an introduction to textual analysis and word vector analysis in R. R is a free and open source piece of software traditionally used for statistical analysis. Packages for R can be installed from the CRAN website, via R Studio, or directly from a source website.

The workshop aims to introduce textual analysis and word vectors using R and provide a starting point for further exploration. The materials used will be examples from Jane Austen's works. Session 1 will explore some basic text analysis using R, in particular frequency analysis and Key Word in Context. While these are not necessary for using word vectors, they are helpful for putting the results of the vector space model into context and allow exploration at text level.

To apply the analysis in Workshop Session One to Japanese the chosen texts will need to be split into words using a package like **RMeCab** (which can be found [here](#)) as the function used to separate the texts into word lists relies upon a regular expression and gaps between words. Unfortunately I can't give further instructions on the use of this package as the website is in Japanese. Another useful package is **Nippon** which includes a variety of utilities for handling Japanese text, including converting text from Katakana to Hiragana. A short article, [Basic Techniques in Text Mining Using Open-source Tools](#) gives a brief outline of some of the processes and tools for working with Japanese Texts.

For more advanced exploration of texts there is **Stylo**, a package for stylometric analysis, which has options for use with Chinese and Japanese text. For further information about the **Stylo** package see [Stylo](#), [Stylometry with R](#), or [Using R for Stylometric Analysis](#).

1 Workshop Session One: Text Analysis in R

Session 1, subsections 1–4 are adapted from Matthew L. Jockers (2014) *Text Analysis with R for Students of Literature*.

Session 1, subsection 5 is adapted from [A gentle introduction to text mining using R](#).

Word vectors

You will need to go to <https://sarajkerr.com/workshops/> and download the resources you need for the workshop. When creating folder names it is important that they do not contain any gaps as this can cause problems when accessing them from R Studio. Your Workshop folder should have the following structure:

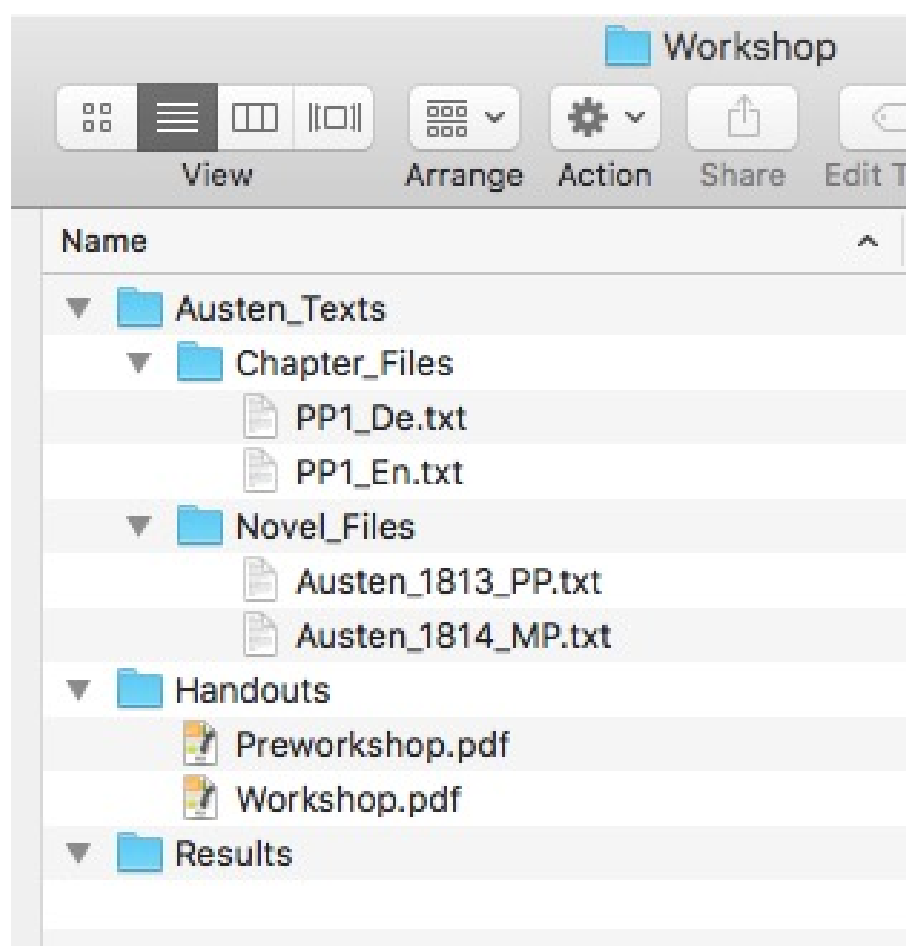


Figure 1: The Workshop Folder

Open R Studio.

Before you start you need to set your ‘working directory’, this is the folder where the texts and results are stored. Go to *Session - Set Working Directory - Choose Directory* then browse to the folder *Workshop* on your desktop.

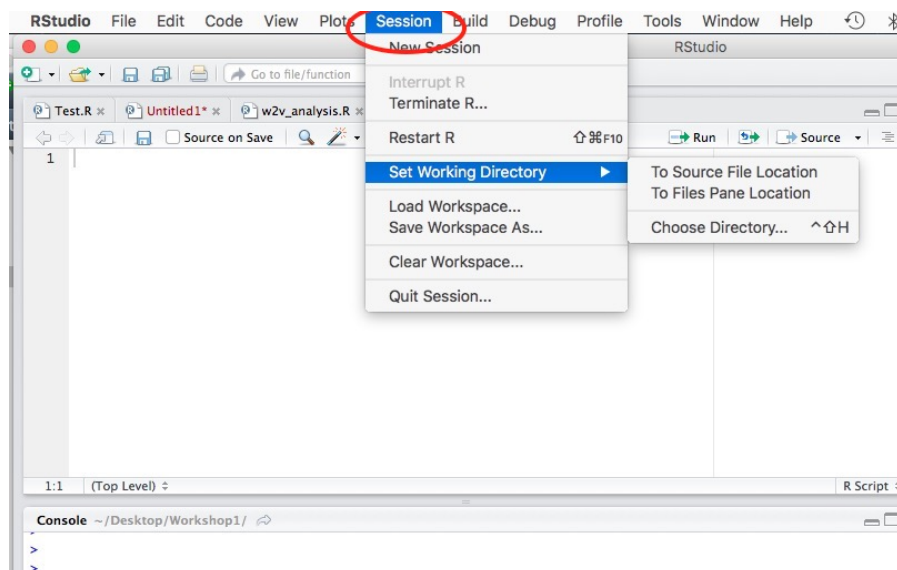


Figure 2: Setting the Working Directory

You will notice that a piece of code has appeared in the R console:

```
1 setwd("pathway to the working directory")
2
3 # This is the function that sets the working directory
```

This first section focuses on frequency analysis using base R.

1.1 Load and Preprocess a File

To load the first file and view the first 6 elements:

```
1 ja_ch <- scan("Austen_Texts/Chapter_Files/PP1_En.txt", what = "character",
2             sep = "\n")
3
4 # This can also be done using
5 mytext <- scan(file.choose(), what = "character", sep = "\n")
6 # which allows you to search for the file you want
7
8 # A character file has been created with 80 elements.
9
10 head(ja_ch)
```

The output will show the first 6 lines of the document - in this case the title and first 4 lines of Chapter 1 of Jane Austen's *Pride and Prejudice*. Each line is enclosed in speech marks as they are character strings. You will also notice that a value **ja_ch** is now in the environment pane, each time you create something it will appear in this pane. This is a character (chr) vector which has 80 elements. To access an individual element or a selection of elements:

```
1 ja_ch[1]
2 ja_ch[3:4]
```

Now create a new vector which only contains the chapter text:

```
1 ja_ch1 <- ja_ch[3:80]
2 # Check the first 6 lines
3 head(ja_ch1)
```

Load the second chapter **PP1_De.txt** into a vector called **ja_ch_d**. Create a vector which removes the first two lines and saves the chapter text into a new vector called **ja_ch1d**. You should now have a character vector which contains a bad German translation of the first chapter of *Pride and Prejudice* (thanks to Google Translate).

Remove line breaks from each file (if you wish to start over in the same session rerun these two lines of code to recreate the unprocessed text). It is a good idea to keep a vector of the text in an unprocessed form, if you make a mistake you won't have to start from the beginning.

```
1 text1 <- paste(ja_ch1, collapse = " ")
2 text2 <- paste(ja_ch1d, collapse = " ")
```

The chapters are each now a single character string.

Before word frequencies can be calculated, the text needs to be processed. The level of processing will depend on the text and its language. For the texts we are using we will change the text to lower case, and extract the words saving them as a vector .

```
1 # Change text to lower case
2 text1 <- tolower(text1)
3 # Check in the environment pane - the capital letters have now been removed
4
5 # select words only
6 text1 <- strsplit(text1, "\\W") # The second argument is a regular expression
7
8 # If we look in the environment pane text1 is now a 'List of 1'
9 text1 <- unlist(text1) # simplify to a vector
10 text1
```

Most of the chapter is now displayed (R has a 'maximum print' limit), but where there was punctuation there are now "" indicating an empty space. Our next step is to remove these.

```
1 text1 <- text1[which(text1 != "")] # '!=' means 'does not equal'
2 # This subsets the vector by removing elements which are not spaces
3 text1
```

text1 is now a character vector where each word is an element. If we look at the environment pane we can see the length of the vector in square brackets - 852. This gives us the number of tokens (words) in the text. It is also possible to calculate the length of a vector directly in R:

```
1 length(text1)
```

Now do the same for **text2**.

1.2 Tokens, Types and Word Frequency

You can search for a word by its index number - R starts indexing at 1 - or by using the **which()** function. The **which()** function will return the index numbers of all occurrences of the selected word. By using the **length()** function we can identify how many times a word appears in the text. We can also use the **unique()** function to identify the word types in the text.

```
1 text1[42]
2 text2[723]
3
4 which(text1 == "neighbourhood")
5 which(text2 == "nachbarschaft")
6
7 # number of times a word appears in the text
8 length(text1[which(text1 == "neighbourhood")])
9
10 # number of types in the text
11 length(unique(text2))
```

Now we will use R to create a raw frequency table, sort it by most frequent words and view the top 10 as a list and as a simple plot.

```
1 # create a table of frequencies
2 text1_freqs <- table(text1)
3
4 # sort by frequency
5 sorted_text1 <- sort(text1_freqs, decreasing = TRUE)
6
7 # View the top 10 most frequent words in the text
8 sorted_text1[1:10]
9
10 # View as a plot
11 plot(sorted_text1[1:10])
12 # If all 10 words don't appear click on the Zoom option in the plot pane.
```

Although it can be interesting to view the raw frequencies, calculating the relative frequencies allows for comparisons to be made across and between texts. The code below calculates the relative frequencies and plots the top ten words. This time we are going to add a title to the plot and custom x and y-axis titles.

```
1 # To calculate the relative frequencies
2 rel_freqs_text1 <- 100 * (sorted_text1 / sum(sorted_text1))
3
4 # Plot top 10 relative frequencies
5 plot(rel_freqs_text1[1:10], type = "b", main= "Relative frequencies in PP Ch 1",
6       xlab = "Top Ten Words", ylab = "Percentage of Chapter")
7 # If all 10 words don't appear click on the Zoom option in the plot pane.
```

1.3 Exploring Word Frequency Across a Novel

So far we have focused on a single chapter - now we will briefly explore a full novel, Jane Austen's *Mansfield Park*. Rather than calculating the relative frequencies for the whole novel, which we can do using the code we used for the individual chapter, we are going to calculate the relative frequencies per chapter using a regular expression.

```

1  # Load the text into R
2  text_MP <- scan("Austen_Texts/Novel_Files/Austen_1814_MP.txt",
3                 what = "character", sep = "\n")
4
5  # View the first 6 lines of the file
6  head(text_MP)
7
8  # Using regular expressions to identify chapter index positions
9  chapters <- grep("^CHAPTER \\d", text_MP)
10
11 text_MP[chapters]
12
13 # Add an end point to the novel and add to chapters
14 text_MP <- c(text_MP, "END") # the length of text_MP is increased by 1
15 finish <- length(text_MP)
16 chapters <- c(chapters, finish)
17
18 # Calculating the word frequencies by chapter using a for loop
19
20 ch_raw_freqs <- list() # Creating empty lists to be filled by the loop
21 ch_rel_freqs <- list()
22
23 for(i in 1:length(chapters)) {
24   if(i != length(chapters)){
25     ch_title <- text_MP[chapters[i]]
26     start <- chapters[i] + 1
27     end <- chapters[i + 1] - 1
28     ch_lines <- text_MP[start:end]
29     ch_text <- tolower(paste(ch_lines, collapse = " "))
30     ch_text <- strsplit(ch_text, "\\W")
31     ch_text <- unlist(ch_text)
32     ch_text <- ch_text[which(ch_text != "")]
33     ch_freqs <- table(ch_text)
34     ch_raw_freqs[[ch_title]] <- ch_freqs
35     rel_ch_freqs <- 100 * (ch_freqs / sum(ch_freqs))
36     ch_rel_freqs[[ch_title]] <- rel_ch_freqs
37   }
38 }
39
40 # To find out the frequency of a word across each chapter
41 marriage <- lapply(ch_rel_freqs, '[', "marriage")
42
43 head(marriage) # shows the relative frequency per chapter
44
45 marriage_com <- do.call(rbind, marriage) # Combines the results by row
46 # This creates a matrix
47
48 head(marriage_com)

```

The matrix created in R now appears in the environment pane under **data**. There is a small spreadsheet icon on the right hand side of the entry, if you click this the matrix will open in the document editor pane.

However, it is easier to view the different frequencies of words by plotting them next to each other. In this next example we will compare the relative frequency of 'he' and 'she' in the novel by plotting them next to each other:

```

1  # Compare instances of 'he' and 'she' in the novel
2  he <- lapply(ch_rel_freqs, '[', "he")
3  he_com <- do.call(rbind, he)
4  she <- lapply(ch_rel_freqs, '[', "she")
5  she_com <- do.call(rbind, she)
6
7  # Extract the relative frequencies and combine in columns
8  he_rf <- he_com[, 1]
9  she_rf <- she_com[, 1]
10 # This extracts all rows of the 1st column - to extract data from a matrix
11 # you use nameofmatrix[rows, columns]
12 # Combine the frequencies into two columns
13 rf_he_she <- cbind(he_rf, she_rf)
14
15 head(rf_he_she)
16
17 # Create a plot comparing the frequencies side by side
18 barplot(rf_he_she, beside = TRUE, col = "red")

```

Once you have created a relative frequencies list you may wish to compare a variety of different words. Rather than rewriting the code each time we can create a function that allows us to do the comparison with a single line of code.

```

1  # The arguments for the function are the chapter relative frequencies and
2  # your two chosen words - the words need to be in speech marks
3  rel_freq_comp <- function(ch_rel_freqs, word1, word2) {
4      worda <- lapply(ch_rel_freqs, '[', word1)
5      worda_com <- do.call(rbind, worda)
6      wordb <- lapply(ch_rel_freqs, '[', word2)
7      wordb_com <- do.call(rbind, wordb)
8      worda_rf <- worda_com[, 1]
9      wordb_rf <- wordb_com[, 1]
10     rf_a_b <- cbind(worda_rf, wordb_rf)
11     # This changes the column names to your chosen words
12     colnames(rf_a_b) <- c(word1, word2)
13     barplot(rf_a_b, beside = TRUE, col = "red",
14             main = "Relative Frequencies By Chapter")
15 }
16
17 # To use the function to compare the names of the two main characters
18 rel_freq_comp(ch_rel_freqs, "fanny", "edmund")

```

1.4 Handling Multiple Files and Key Word in Context

If we want to look at a corpus of texts we need to have a way of identifying the files we wish to examine. To do this, we will create functions to list the files in a folder and to

create a word list from each of the files. The time this takes will depend upon the number of files in the folder and their size.

```

1  # Create a path to the files
2  input.dir <- "Austen_Texts/Novel_Files"
3  # Read the name of all .txt files
4  files <- dir(input.dir, "\\*.txt")
5
6  # Create a function to list the files
7  show_files <- function(files){
8      for(i in 1:length(files)){
9          cat(i, files[i], "\n", sep = " ")
10     }
11 }
12
13 # To use the function:
14 show_files(files) # the files are listed and numbered
15
16 # Create a function to create a list of words from each file, you will recognise
17 # the code from the previous sections
18
19 make_word_list <- function(files, input.dir) {
20     # create an empty list for the results
21     word_list <- list()
22     # read in the files and process them
23     for(i in 1:length(files)) {
24         text <- scan(paste(input.dir, files[i], sep = "/"),
25                     what = "character", sep = "\n")
26         text <- paste(text, collapse = " ")
27         text_lower <- tolower(text)
28         text_words <- strsplit(text_lower, "\\W")
29         text_words <- unlist(text_words)
30         text_words <- text_words[which(text_words != "")]
31         word_list[[files[i]]] <- text_words
32     }
33     return(word_list)
34 }
35
36 # To use the function:
37 my_corpus <- make_word_list(files, input.dir)

```

If you scroll down to the bottom of the environment pane you will see the heading **functions** and the three functions we have created will be listed.

To explore individual words within the context they appear in, we can use Key Word in Context. We will create a **kwic** function which uses the output from the **make_word_list()** function and uses the **show_files()** function. This means that before you run the **kwic()** function you will need to check that **make_word_list()** and **show_files()** functions are listed.

```

1  # Creating a Key Word in Context (KWIC) function - this uses the output from
2  # the make_word_list and includes the show_files function
3
4  kwic <- function(my_corpus) {
5      show_files(names(my_corpus))
6      # identify the chosen file

```

```

7   file_id <- as.numeric(
8     readline("Which file would you like to examine? Enter a number: \n"))
9   # identify the number of words each side of the keyword
10  context <- as.numeric(
11    readline("How much context do you want? Enter a number: \n"))
12  # identify the focus word
13  keyword <- tolower(readline("Enter a keyword: \n"))
14  # create the KWIC readout
15  hits <- which(my_corpus[[file_id]] == keyword)
16  if(length(hits) > 0) {
17    result <- NULL
18    for(j in 1:length(hits)) {
19      start <- hits[j] - context
20      if(start < 1) {
21        start <- 1
22      }
23      end <- hits[j] + context
24      cat("\n-----", j, "-----\n")
25      cat(my_corpus[[file_id]][start:(hits[j] - 1)], sep = " ")
26      cat("[", my_corpus[[file_id]][hits[j]], "]", sep = " ")
27      cat(my_corpus[[file_id]][(hits[j] + 1): end], sep = " ")
28      myrow <- cbind(hits[j],
29                    paste(my_corpus[[file_id]][start: (hits[j] - 1)],
30                          collapse = " "),
31                    paste(my_corpus[[file_id]][hits[j]],
32                          collapse = " "),
33                    paste(my_corpus[[file_id]][(hits[j] + 1): end],
34                          collapse = " "))
35      result <- rbind(result, myrow)
36    }
37    colnames(result) <- c("position", "left", "keyword", "right")
38    return(result)
39  } else {
40    cat("YOUR KEYWORD WAS NOT FOUND\n")
41  }
42 }
43
44 # To use the function:
45 results <- kwic(my_corpus)
46 # select the file, context and keyword when prompted
47 # For example - choose text 1, 7 for context and work as the keyword
48
49 # If you wish to save the results as a .csv file
50 write.csv(results, "Results/yourkeyword_text.csv")
51 # e.g. if the chosen word was work in Pride and Prejudice
52 write.csv(results, "Results/work_PP.csv")

```

1.5 Using the 'tm' Package

An alternative to working in base R is to use the 'tm' text mining package. The package is designed specifically for text mining a corpus and offers a variety of tools. An introduction to the 'tm' package can be found [here](#).

Begin by clearing your workspace, this can be done by clicking on the broom in the environment pane or choosing *Session - Clear Workspace*. This will remove all the objects

created during the previous session.

We start by telling R where the files we are using are stored, in this case the *Novel_Files* folder.

```
1 # Tell the computer where the texts are located
2 dname <- file.path("Austen_Texts/Novel_Files") # Tells R the path to the files
3 dname
4 dir(dname) # Lists the files in the directory/folder
```

To load the packages you need to tell R you want to use them. We are going to use *tm*, *ggplot2* and *wordcloud*.

```
1 library(tm) # tm is a text mining package
2 library(ggplot2) # ggplot2 is a graph plotting package
3 library(wordcloud) # lets you create wordclouds
```

Now we have loaded the packages we want to use we need to load the corpus into R using the **tm** package. This will create a V or volatile corpus which means that it only exists in the current session of R. Next we will process the corpus removing punctuation, numbers, uppercase letters and white spaces. We could also remove stopwords, but we won't do that at the moment.

```
1 # load the corpus
2 docs <- Corpus(DirSource(dname))
3
4 # Preprocess the text
5 docs <- tm_map(docs, removePunctuation) # Remove punctuation
6 docs <- tm_map(docs, removeNumbers)     # Remove numbers
7 docs <- tm_map(docs, tolower)           # Convert to lowercase
8 # docs <- tm_map(docs, removeWords, stopwords("english")) # To remove stopwords
9 docs <- tm_map(docs, stripWhitespace)   # Strip whitespace
10 docs <- tm_map(docs, PlainTextDocument)
11 # This is the end of the preprocessing stage.
```

Once the corpus has been processed a Document Term Matrix (DTM) can be created. The DTM is a matrix which describes the frequency of terms in the corpus where each row is a document and each column is a word or term from the corpus.

```
1 # Create a Document Term Matrix
2 dtm <- DocumentTermMatrix(docs)
3
4 # To view or export the DTM
5 corpus_dtm <- as.matrix(dtm)
6 rownames(corpus_dtm) <- c("PP", "MP")
7
8 # It is also possible to create a Term Document Matrix
9 tdm <- TermDocumentMatrix(docs)
10 corpus_tdm <- as.matrix(tdm)
11 colnames(corpus_tdm) <- c("PP", "MP")
12
13 # Write files to a folder
14 write.csv(corpus_dtm, "Results/dtm.csv")
15 write.csv(corpus_tdm, "Results/tdm.csv")
16
17 # Explore your data - this gives the raw frequency for each word in the corpus
```

```
18 freq <- colSums(as.matrix(dtm))
19 length(freq) # the number of unique words in the corpus
20 head(freq) # view the top 6 elements
```

It can be useful to visualise the word frequencies. Below are two examples, a frequency graph showing words which appear more than 500 times, and word clouds showing 100 words and words with a minimum frequency of 100.

```
1 # Create a frequency graph of the words which appear more than 500 times
2
3 word_freq <- data.frame(word=names(freq), freq=freq)
4 p <- ggplot(subset(word_freq, freq>500), aes(word, freq))
5 p <- p + geom_bar(stat="identity")
6 p <- p + theme(axis.text.x=element_text(angle=45, hjust=1))
7 p
8
9 # Create a wordcloud
10
11 set.seed(142) # set.seed is used to ensure replicability
12 wordcloud(names(freq), freq, max.words=100)
13
14 set.seed(142)
15 wordcloud(names(freq), freq, min.freq=100, scale=c(5, .1),
16           colors=brewer.pal(8, "Set1")) # This section determines the colours
```

You could rerun the code on page 10 removing the `#` at the start of line 8 to remove the stopwords to see the difference it makes to the visualisations.

2 Workshop Session Two: Word Vectors in R

Session 2 is adapted from Benjamin Schmidt's (2015) blog post [Word Embeddings](#) and my doctoral research 'Jane Austen in Vector Space' presented at JADH (2016), which can be found [here](#).

2.1 What is a Vector Space Model and How is it Created?

This description from [TensorFlow](#) explains briefly what a vector space model is and the difference between the skip-gram and the continuous bag of words model. The website has an extended, more technical explanation for those interested.

Vector space models (VSMs) represent (embed) words in a continuous vector space where semantically similar words are mapped to nearby points ('are embedded nearby each other'). VSMs have a long, rich history in NLP, but all methods depend in some way or another on the Distributional Hypothesis, which states that words that appear in the same contexts share semantic meaning. The different approaches that leverage this principle can be divided into two categories: count-based methods (e.g. Latent Semantic Analysis), and predictive methods (e.g. neural probabilistic language models).

This distinction is elaborated in much more detail by [Baroni et al.](#), but in a nutshell: Count-based methods compute the statistics of how often some word co-occurs with its neighbor words in a large text corpus, and then map these count-statistics down to a small, dense vector for each word. Predictive models directly try to predict a word from its neighbors in terms of learned small, dense embedding vectors (considered parameters of the model).

Word2vec is a particularly computationally-efficient predictive model for learning word embeddings from raw text. It comes in two flavors, the Continuous Bag-of-Words model (CBOW) and the Skip-Gram model (Section 3.1 and 3.2 in [Mikolov et al.](#)). Algorithmically, these models are similar, except that CBOW predicts target words (e.g. 'mat') from source context words ('the cat sits on the'), while the skip-gram does the inverse and predicts source context-words from the target words. This inversion might seem like an arbitrary choice, but statistically it has the effect that CBOW smoothes over a lot of the distributional information (by treating an entire context as one observation). For the most part, this turns out to be a useful thing for smaller datasets. However, skip-gram treats each context-target pair as a new observation, and this tends to do better when we have larger datasets.

The skip-gram model is the default for the **wordVectors** package - although it is possible to change this, the skip-gram model works slightly better with less frequent words so it makes sense to keep the default settings unless there is a particularly good reason to change them.

Begin by clearing your workspace, this can be done by clicking on the broom in the

environment pane or choosing *Session - Clear Workspace*. This will remove all the objects created during the previous session.

2.2 Create, or Load, a Vector Space Model

The **wordVectors** package in R allows you to create a vector space model or to load a previously created model. The packages needed for this session should already have been loaded (see the Pre-workshop materials for details).

```
1 library(wordVectors)
2 library(tsne)
3 library(Rtsne)
4 library(ggplot2)
5 library(ggrepel)
6 library(stringi)
```

Before a vector space model can be created the .txt files need to be processed. The **prep_word2vec** function takes in a folder of .txt files and outputs a single .txt file which combines the texts in one document removes punctuation and converts all words to lower case.

The model then needs to be trained. The created model will appear in the environment pane in R Studio, a copy called **Novel.bin** will be saved to the *Results* folder. This can be loaded into future R sessions saving processing time. It is important to remember that the larger the corpus the longer it will take to process and load.

```
1 # If a prepared text file has not already been created follow this step
2
3 prep_word2vec("Austen_Texts/Novel_Files",
4             "Austen_Texts/Novel_Files/Novel_corpus.txt", lowercase = T)
5
6 # This can take some time depending on the size of the files
7
8 # train_word2vec takes several parameters - input prepared .txt file, an
9 # output file, vectors are the number of dimensions the default is 100, and
10 # window is the number of words either side of the context word, by default
11 # the function uses the skip-gram method
12
13 ja <- train_word2vec("Austen_Texts/Novel_Files/Novel_corpus.txt",
14                   output = "Results/Novel.bin", threads = 3,
15                   vectors = 100, window = 15)
16
17 # This will take a little while to process
18 # A Vector Space Model has been created
19
20 # If a model has previously been created and saved it can be loaded using
21 ja <- read.vectors("Results/Novel.bin")
```

2.3 Prepare Texts for KWIC Analysis

While the **wordVectors** package will allow a detailed exploration of the words from the corpus, it does not show the specific context in which the word originally appeared in the novel. To enable contextualisation of any words of interest, we will use the **kwic()** function we created in Session 1.

```

1  # Create a path to the files
2  input.dir <- "Austen_Texts/Novel_Files"
3  # Read the name of all .txt files
4  files <- dir(input.dir, "\\*.txt")
5
6  # Create a function to list the files
7  show_files <- function(files){
8      for(i in 1:length(files)){
9          cat(i, files[i], "\n", sep = " ")
10     }
11 }
12
13 # To use the function:
14 show_files(files) # the files are listed and numbered
15
16 # Create a function to create a list of words from each file, you will recognise
17 # the code from the previous sections
18
19 make_word_list <- function(files, input.dir) {
20     # create an empty list for the results
21     word_list <- list()
22     # read in the files and process them
23     for(i in 1:length(files)) {
24         text <- scan(paste(input.dir, files[i], sep = "/"),
25                     what = "character", sep = "\n")
26         text <- paste(text, collapse = " ")
27         text_lower <- tolower(text)
28         text_words <- strsplit(text_lower, "\\W")
29         text_words <- unlist(text_words)
30         text_words <- text_words[which(text_words != "")]
31         word_list[[files[i]]] <- text_words
32     }
33     return(word_list)
34 }
35
36
37 # To use the function:
38 my_corpus <- make_word_list(files, input.dir)
39
40 # The Novel_corpus.txt file will be read in as 2 items as it has already been
41 # processed - this doesn't have an impact on using kwic() later.
42
43 # Creating a Key Word in Context (KWIC) function - this uses the output from
44 # the make_word_list and includes the show_files function
45
46 kwic <- function(my_corpus) {
47     show_files(names(my_corpus))
48     # identify the chosen file
49     file_id <- as.numeric(

```

```

50     readline("Which file would you like to examine? Enter a number: \n"))
51   # identify the number of words each side of the keyword
52   context <- as.numeric(
53     readline("How much context do you want? Enter a number: \n"))
54   # identify the focus word
55   keyword <- tolower(readline("Enter a keyword: \n"))
56   # create the KWIC readout
57   hits <- which(my_corpus[[file_id]] == keyword)
58   if(length(hits) > 0) {
59     result <- NULL
60     for(j in 1:length(hits)) {
61       start <- hits[j] - context
62       if(start < 1) {
63         start <- 1
64       }
65       end <- hits[j] + context
66       cat("\n-----", j, "-----\n")
67       cat(my_corpus[[file_id]][start:(hits[j] - 1)], sep = " ")
68       cat("[", my_corpus[[file_id]][hits[j]], "]", sep = " ")
69       cat(my_corpus[[file_id]][(hits[j] + 1): end], sep = " ")
70       myrow <- cbind(hits[j],
71                     paste(my_corpus[[file_id]][start: (hits[j] - 1)],
72                           collapse = " "),
73                     paste(my_corpus[[file_id]][hits[j]],
74                           collapse = " "),
75                     paste(my_corpus[[file_id]][(hits[j] + 1): end],
76                           collapse = " "))
77       result <- rbind(result, myrow)
78     }
79     colnames(result) <- c("position", "left", "keyword", "right")
80     return(result)
81   } else {
82     cat("YOUR KEYWORD WAS NOT FOUND\n")
83   }
84 }
85
86 # To use the function, this time there will be a choice of three files,
87 # Pride and Prejudice, Mansfield Park, and a text combining the two:
88 results <- kwic(my_corpus)

```

2.4 Exploring a Vector Space Model

There are a number of functions built into the **wordVectors** package which allow the vector space model to be explored. We will look at the **nearest_to()** function. Details of other functions can be found on the GitHub page for the **wordVectors** package - [here](#).

The **nearest_to()** function allows you to search the vector space model for words which appear near to a chosen word or set of words, both semantically and contextually. The selected words can be viewed as a list or a plot. It is useful at this point to explore the specific context any words of interest appear in by using the **kwic()** function.


```

1  # The nearest_to() function uses cosine similarity
2  nearest_to(ja, ja[["marriage"]], 50) # 50 nearest words to marriage
3
4  marriage <- nearest_to(ja, ja[["marriage"]], 100)
5  # This creates a vector of distances for the 100 words nearest to marriage
6
7  # Create a plot of the 100 words nearest to marriage
8  plot(ja[[names(marriage), average = F]])
9  # This takes a little time to run
10
11 # You can also search for words nearest to a semantic field
12 wealth <- nearest_to(ja, ja[[c("establishment", "income", "fortune",
13                               "privilege", "wealth", "property", "affluence")]], 300)
14
15 wealth <- names(wealth)
16
17 sample(wealth, 100)
18 # This allows a random sample of 100 words to be taken from the vector
19
20 # An alternative is to use the 'magrittr' package which acts as a pipeline
21 # enabling short cuts
22 # It works in a similar manner to the vector for marriage but extracts the names
23 install.packages("magrittr")
24 library(magrittr)
25 wealth1 <- ja %>% nearest_to(ja[[c("establishment", "income", "fortune",
26                                   "privilege", "wealth", "property", "affluence")]], 300) %>% names
27 sample(wealth1, 100)
28
29 # Plot the chosen segment
30 plot(ja[[wealth[1:100], average = F]])
31
32 # using the kwic() function explore the context of some of the words in the
33 # plot
34 results1 <- kwic(my_corpus)
35
36 # If you wish to save your results amend this piece of code:
37 # If you wish to save the results as a .csv file
38 write.csv(results, "Results/yourkeyword_text.csv")

```

2.5 Using the w2v_analysis Function

To speed up the exploration of the vector space model I created a function called **w2v_analysis** using the **wordVectors** package but plotting in **ggplot2** which allows more modifications. This function uses the **Rtsne** package which uses the Barnes-Hut implementation of T-sne and is somewhat quicker than the **tsne** package which is used in the **plot()** function in **wordVectors**.

```

1  # The function takes 4 arguments:
2  # vsm - a vector space model
3  # words - a character vector of focus words
4  # seed - an integer
5  # path - the path to the folder you want files saved to
6  # ref_name - the reference name for the exported files
7

```

```

8  # The function will create a vector which is the average of the words input and
9  # will output a wordlist of the 500 nearest words, a csv of the words and their
10 # positions, and a plot of the 2D reduction of the vector space
11 # model using the Barnes-Hut implementation of t-SNE. The points for each word
12 # are marked in red so the labels can be moved by ggrepel for ease of reading.
13 # set.seed is used to ensure replicability
14
15 w2v_analysis <- function(vsm, words, seed, path, ref_name) {
16   # Set the seed
17   if (!missing(seed))
18     set.seed(seed)
19
20   # Identify the nearest 10 words to the average vector of search terms
21   ten <- nearest_to(vsm, vsm[[words]])
22
23   # Identify the nearest 500 words to the average vector of search terms and
24   # save as a .txt file
25   main <- nearest_to(vsm, vsm[[words]], 500)
26   wordlist <- names(main)
27   filepath <- paste0(path, ref_name)
28   write(wordlist, paste0(filepath, ".txt"))
29
30
31   # Create a subset vector space model
32   new_model <- vsm[[wordlist, average = F]]
33
34   # Run Rtsne to reduce new VSM to 2D (Barnes-Hut)
35   reduction <- Rtsne(as.matrix(new_model), dims = 2, initial_dims = 50,
36                     perplexity = 30, theta = 0.5, check_duplicates = F,
37                     pca = F, max_iter = 1000, verbose = F,
38                     is_distance = F, Y_init = NULL)
39
40   # Extract Y (positions for plot) as a dataframe and add row names
41   df <- as.data.frame(reduction[,Y])
42   rows <- rownames(new_model)
43   rownames(df) <- rows
44
45   # Save dataframe as .csv file
46   write.csv(df, paste0(filepath, ".csv"))
47
48   # Create t-SNE plot and save as jpeg
49   ggplot(df) +
50     geom_point(aes(x = V1, y = V2), color = "red") +
51     geom_text_repel(aes(x = V1, y = V2, label = rownames(df))) +
52     xlab("Dimension 1") +
53     ylab("Dimension 2 ") +
54     # geom_text(fontface = 2, alpha = .8) +
55     theme_bw(base_size = 12) +
56     theme(legend.position = "none") +
57     ggtitle(paste0("2D reduction of VSM ", ref_name, " using t_SNE"))
58
59   ggsave(paste0(ref_name, ".jpeg"), path = path, width = 24,
60         height = 18, dpi = 100)
61
62   new_list <- list("Ten nearest" = ten, "Status" = "Analysis Complete")
63   return(new_list)
64

```

```
65 }  
66  
67 # To use the function a single word or a vector of connected words can be used  
68 status <- c("status", "rank", "position", "superior", "aristocracy", "gentry")  
69  
70 # The function is called as below:  
71 # w2v_analysis(VectorSpaceModel, Word, Number, "Path/", "Reference_Name")  
72 # E.g.  
73 w2v_analysis(ja, status, 42, "Results/", "Status")
```

This will take a short time to run. The console will return the 10 nearest words and include the status "Analysis Complete". If you look in the Results folder there will be 3 new files a .txt file containing the word list of the 500 nearest words, a csv containing the words and their position in the plot, and a .jpeg file containing the plot.

Take some time to explore the model.

2.6 Some Resources on using word2vec with Asian Texts

Here are some resources and articles on using word2vec with Asian texts.

- [Vector Similarity of Related Words and Synonyms in the Japanese Word-Net.](#)
- [Japanese-words-to-vectors: Word2vec \(word to vectors\) approach for Japanese language using Gensim.](#)
- [Bilingual Word Embeddings for Phrase-Based Machine Translation.](#)

3 Contact Details

If you wish to contact me my details are:

- **Email:** sarajkerr@icloud.com
- **Web:** <http://sarajkerr.com>
- **Twitter:** @data_fiend
- **GitHub:** <https://github.com/SaraJKerr>