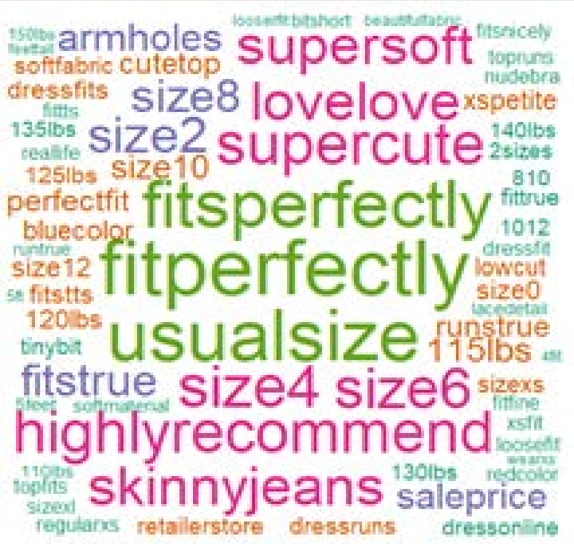
DEVELOPING A BIGRAM

UNDERSTANDING THE CONTEXT RELATED TO SPECIFIC WORDS:

- A FREQUENCY WORD CLOUD with the 2 most common words is called a Bigram. A BIGRAM is used to provide more context on the words that are being used consecutively or in conjunction with the words that appear the most.
- To demonstrate, in the "simplified" example, R Studio is used to analyse approximately 23,000 text only reviews obtained for a Women's Ecommerce clothing company (data was sourced from Kaggle – a Machine Learning and Data Science Community website). The 23,000 reviews covered 20 product categories with multiple variations of each product – resulting in 1,206 discrete products.
- The first step in the analysis is looking at a high-level view of the approximately 23,000 reviews by creating a FREQUENCY WORD CLOUD. This is done by condensing each of the reviews into a string of text and counting the times a single word appears
- To provide context related to the words that appear most frequent because of the frequency word cloud, a BIGRAM can be used. This is done by *TOKENISING consecutive sequences of words using after cleaning the text of 'STOP-WORDS'
- Interestingly the resulting CLOUD shown below indicates the word fit is mostly associated with PERFECT and reviewers 'HIGHLY RECOMMEND' the product.

BIGRAM WORD CLOUD



DEVELOPING A BIGRAM:

As can be seen, the BIGRAM suggests a very positive experience with associated words such as 'SUPER CUTE' and 'LOVE LOVE' and 'HIGHLY RECOMMEND' appearing. It is notable that the word FIT is used frequently, and with the positive spin of the majority being FITS PERFECTLY. With SIZE8 and SIZE2 being mentioned indicates some smaller sizes are more popular and/or some

CLEANING THE DATA

- The Women's Clothing Data Set – in a .csv file format, contained approximately 23,000 reviews covering 1,206 products.
- Before commencing the analysis, a review of the data set identifies a total number of reviews (23,000) across 5 departments (Bottoms, Dresses, Intimates, Jackets, Tops) with a total of 20 Product Categories within the departments and 1,206 Product sub-categories. Each of the 1,206 Products has their individual Stock Keeping Unit identifications or SKUs which provides specific information on the silhouette (style of the product) and/or colour.
- Note that the quantity of customer reviews has a direct correlation to the quantity of sales – i.e., you purchase the item then review it.

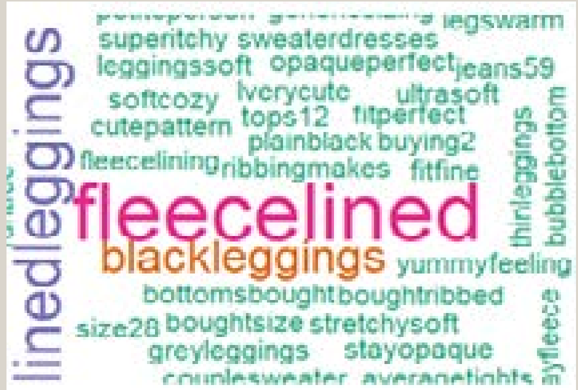
Feedback text can contain other words, symbols, numbers, punctuation, stemming (e.g., argue vs argues vs arguing) co joining words (e.g. for, and, nor, but, or, yet, and so) and stop words (ie words that occur frequently but don't provide a lot of insights (such as the, I, she, either etc)). Other specific words which are removed to avoid the data being skewed, for example a name of a product or business., these all have to be removed to provide greater accuracy to the insight. These can be removed using a R package (tm package).

DEVELOPING A BIGRAM

APPLYING A BIGRAM ANALYSIS TO A SUBSET OF THE DATA

- Investigate a subset of products (SKUS)
- Takings a subset of products, either they are new products and want to provide a high level context to how customer view the products, product subset could be for new products, poor performing, or high performing product.
- The BIGRAM below for the products selected show they are winter garments, but have some qualities that need to be investigated related to the material such as some customer describe them as "super itchy" . "thin leggings", "stay opaque

BIGRAM WORD CLOUD



BARPLOT OF A BIGRAM

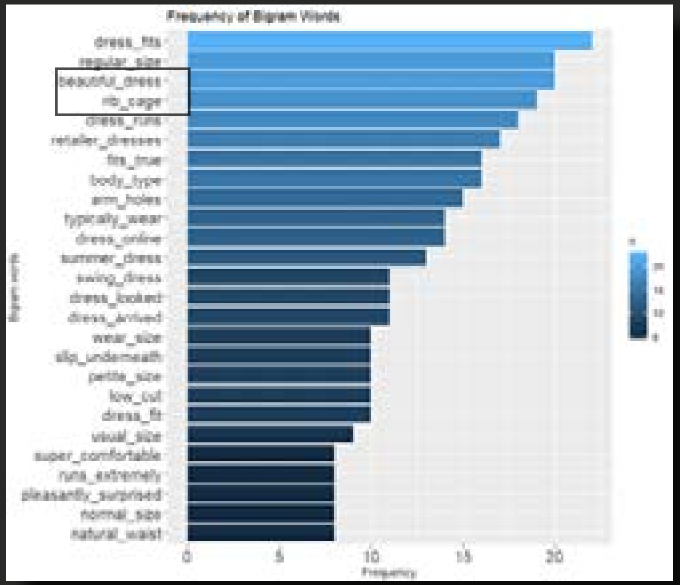
Plotting the bigram in a bar plot can more clearly identify how many customers are saying the same thing.

Dress Product Category

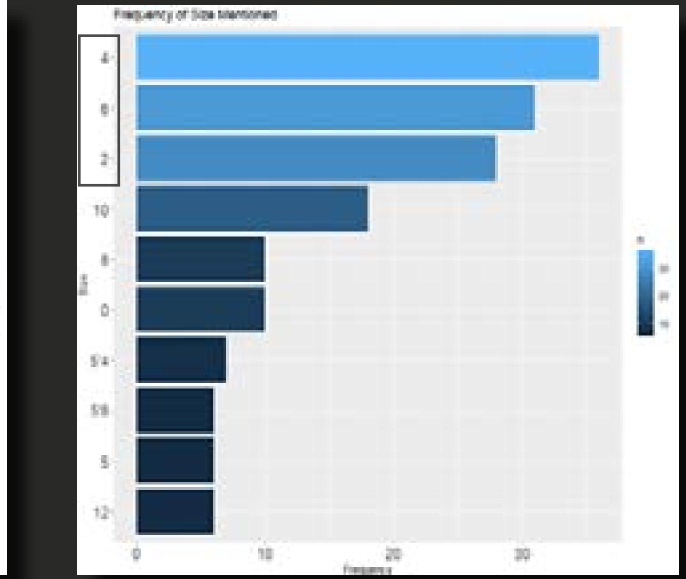
- As shown below , for a products category "dresses" it can be seen the fit runs tight/small with ~ 18 customers mentioning "rib cage" , "dress runs" and "petite size".
- To get a feel for what sizes are being mentioned the most. The tokenised bigrams can be filtered for the first or second word to include size. The bar plot below shows the frequency of sizes mentioned in the reviews. It can be shown that size 4 is mentioned the most, depending on the target customer, this would suggest that customers needing to mention the size that it runs small/large.

DEVELOPING A BIGRAM

BIGRAM PAR PLOT



DRESS SIZE BIGRAM BAR PLOT



OUTCOMES FROM A BIGRAM

Condenses 23,000 written reviews into a high-level overview of how customers respond to the brand/company.

- Can be used for induvial products, departments.
- Identify categories that may need to be identified within the data set. In this case categorize reviews into categories such as
  - o Fabric Reviews – reviews related to fabric
  - o Fit Reviews – reviews related to fit
  - o Colour Reviews – reviews related to colour
  - o Order Reviews – reviews related to the order
  - o Where further analysis can be performed on these categories.
- Combined with other data to determine aspects related to high sales volume products

APPENDIX 1 – FURHTUR READING

- Part 1 - DEVELOPING A FREQUENCY WORD CLOUD
- Part 2 – DEVELOPING A BIGRAM
- Part 3 – Text Analytics – Sentiment Analysis
- Part 4 - Text Analytics – Emotional Classification
- Part 5 –Full case Study – Text Analysis on written reviews from a Women's Online fashion company using RStudio

APPENDIX 2 – R PACKAGES

While performing the analysis using R, the analyst can call on specialised packages to perform detailed analysis of the data.

Packages used to perform detailed analysis.

```
library(tidyverse)
library("tm")
library("SnowballC")
library("wordcloud")
library("RColorBrewer")
library("ggplot2")
library("tidytext")
library("glue")
library(DT)
library(tidytext)
library(dplyr)
library(stringr)
library(readr)
library(wordcloud)
library(reticulate)
library(crfsuite)
```

SCOPE DATA

VISIT

www.scopedata.net

INFO@SCOPEDATA.NET