University of South Wales
Prifysgol De Cymru

MS4S09

Text Mining – Part 2

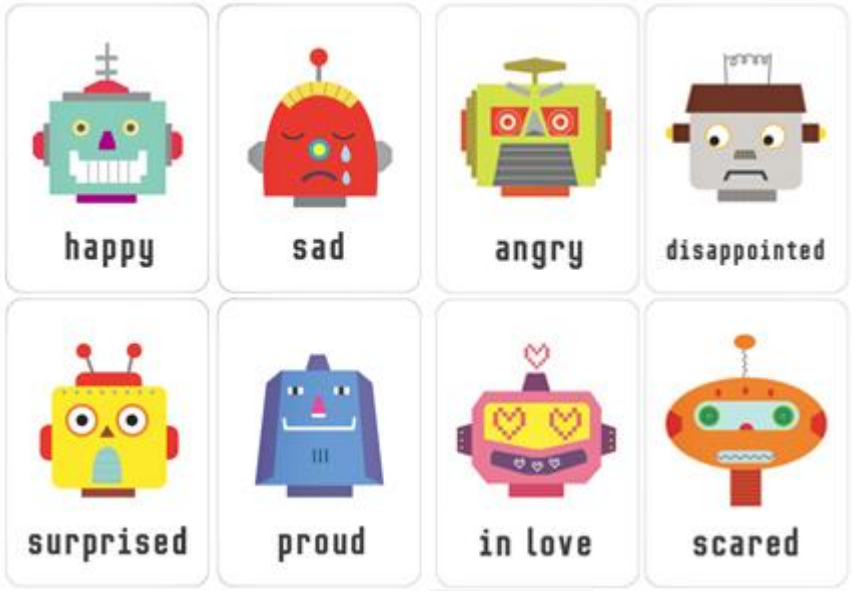
University of South Wales
Prifysgol De Cymru

Natural Language Processing

The human language is complex.

Teaching a machine to analyse the various grammatical nuances, cultural variations, slang and misspellings that occur in online mentions is a difficult process.

Teaching a machine to understand how context can affect tone is even more difficult.



University of South Wales
Prifysgol De Cymru

Sentiment Analysis

A recent trend is to analyse people's feelings, opinions and orientation about facts and brands: this is done by exploiting Sentiment Analysis techniques, whose goal is to classify the polarity of a piece of text according to the opinion of the writer.

These systems extract the following attributes of the expression:

- Polarity: if the speaker express a positive or negative opinion,
- Subject: the thing that is being talked about,
- Opinion holder: the person, or entity that expresses the opinion.

University of South Wales
Prifysgol De Cymru

Sentiment Analysis

Exercise

Take the following three sentences:

- I love flying British Airways because they have the best food.
- That Fiat Punto is the ugliest car I've ever seen.
- I love this phone but wouldn't recommend it to my friends.

State the polarity, subject and opinion.

University of South Wales
Prifysgol De Cymru

Sentiment Analysis

You had to read each sentence manually and determine the sentiment, whereas sentiment analysis, on the other hand, can scan and categorise these sentences for you as positive, negative, or neutral.

Like humans, sentiment analysis looks at sentence structure, adjectives, adverbs, magnitude, keywords, and more to determine the opinion expressed in the text.

The advantage of sentiment analysis is that it’s much, much quicker.

University of South Wales
Prifysgol De Cymru

Sentiment Analysis

Exercise

What is more difficult about this text?

Penny stopped at Costa Coffee on her way home.
She thought a coffee was good every few hours.
But it turned out to be too expensive there.

University of South Wales
Prifysgol De Cymru

Sentiment Analysis

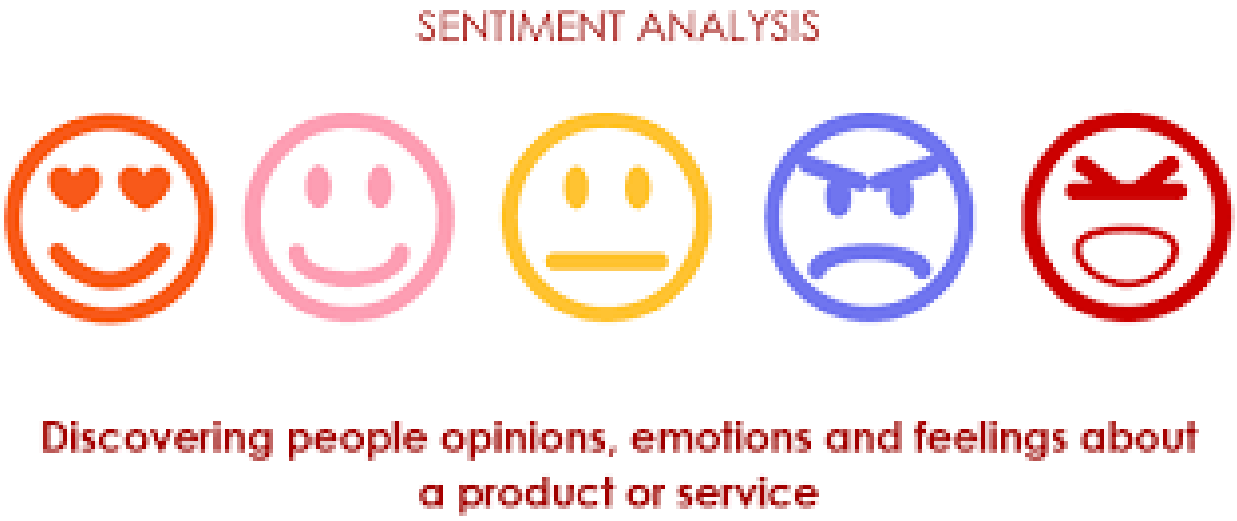
There are three main approaches to sentiment analysis. These are:

- Knowledge-based - categorise text based on unambiguous ‘affect words’ like love, like, hate, happy, sad, angry, and so on.
- Statistical methods – using advanced techniques the computer detects the holder of a sentiment (the person with the opinion) and the target (the product or service) in a sentence.
- Hybrid approaches - a combination of the above with the addition of an advanced understanding of language in order to detect semantics.

University of South Wales
Prifysgol De Cymru

Sentiment Analysis

Once we have cleaned up our text and performed some basic word frequency analysis, the next step is to understand the opinion or emotion in the text.



University of South Wales
Prifysgol De Cymru

Lexicon-based Sentiment Analysis

Lexicon based Sentiment Analysis uses a predefined dictionary of positive and negative words and calculates the sentiment score based on the number of matches of words in text with each of the dictionaries.

Sentiment is calculated as follows:

positive_matches -
negative_matches

University of South Wales
Prifysgol De Cymru

Steps to follow

Create or find a list of words (lexicon) associated with strongly positive or negative sentiment.

Count the number of positive and negative words in the text.

Analyse the mix of positive to negative words.

Many positive words and few negative words indicates positive sentiment, while many negative words and few positive words indicates negative sentiment.

University of South Wales
Prifysgol De Cymru

Lexicon-based Sentiment Analysis

The NRC Emotion Lexicon is a list of English words and their associations with eight basic emotions:
anger, fear, anticipation, trust, surprise, sadness, joy, and disgust
and two sentiments:
negative and positive

The annotations were manually completed by crowdsourcing.

https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

University of South Wales
Prifysgol De Cymru

Lexicon-based Sentiment Analysis

word <chr>	sentiment <chr>
abacus	trust
abandon	fear
abandon	negative
abandon	sadness
abandoned	anger
abandoned	fear
abandoned	negative
abandoned	sadness
abandonment	anger
abandonment	fear

University of South Wales
Prifysgol De Cymru

Lexicon-based Sentiment Analysis

The Bing lexicon categorises words in a binary fashion into positive and negative categories.

word <chr>	sentiment <chr>
2-faced	negative
2-faces	negative
a+	positive
abnormal	negative
abolish	negative
abominable	negative
abominably	negative
abominate	negative
abomination	negative
abort	negative

https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

University of South Wales
Prifysgol De Cymru

Lexicon-based Sentiment Analysis

These were constructed via either crowdsourcing (using, for example, Amazon Mechanical Turk) or by the labor of one of the authors, and were validated using some combination of crowdsourcing again, restaurant or movie reviews, or Twitter data.

Given this information, we may hesitate to apply these sentiment lexicons to styles of text dramatically different from what they were validated on, such as narrative fiction from 200 years ago.

University of South Wales
Prifysgol De Cymru

Lexicon-based Sentiment Analysis

Not every English word is in the lexicons because many English words are pretty neutral. It is important to keep in mind that these methods do not take into account qualifiers before a word, such as in “no good” or “not true”; a lexicon-based method like this is based on unigrams only. For many kinds of text.

One last caveat is that the size of the chunk of text that we use to add up unigram sentiment scores can have an effect on an analysis. A text the size of many paragraphs can often have positive and negative sentiment averaged out to about zero, while sentence-sized or paragraph-sized text often works better.

University of South Wales
Prifysgol De Cymru

Lexicon-based Sentiment Analysis

Let's use the text of Jane Austen's 6 completed, published novels from the janeaustenr package (Silge 2016), and transform them into a tidy format.

The janeaustenr package provides these texts in a one-row-per-line format, where a line is this context is analogous to a literal printed line in a physical book.

While it is true that using these sentiment lexicons may give us less accurate results than with tweets sent by a contemporary writer, we still can measure the sentiment content for words that are shared across the lexicon and the text.

https://www.tidytextmining.com/sentiment.html

University of South Wales
Prifysgol De Cymru

Introduction to tidytext

Tidy data has a specific structure:

- Each variable is a column
- Each observation is a row
- Each type of observational unit is a table

We thus define the tidy text format as being a table with one-token-per-row.

A token is a meaningful unit of text, such as a word, that we are interested in using for analysis, and tokenisation is the process of splitting text into tokens.

To work with this as a tidy dataset, we need to restructure it in the one-token-per-row format, which is done with the unnest_tokens() function.

University of South Wales
Prifysgol De Cymru

Introduction to tidytext

We can remove stop words (kept in the tidytext dataset stop_words) with an anti_join().

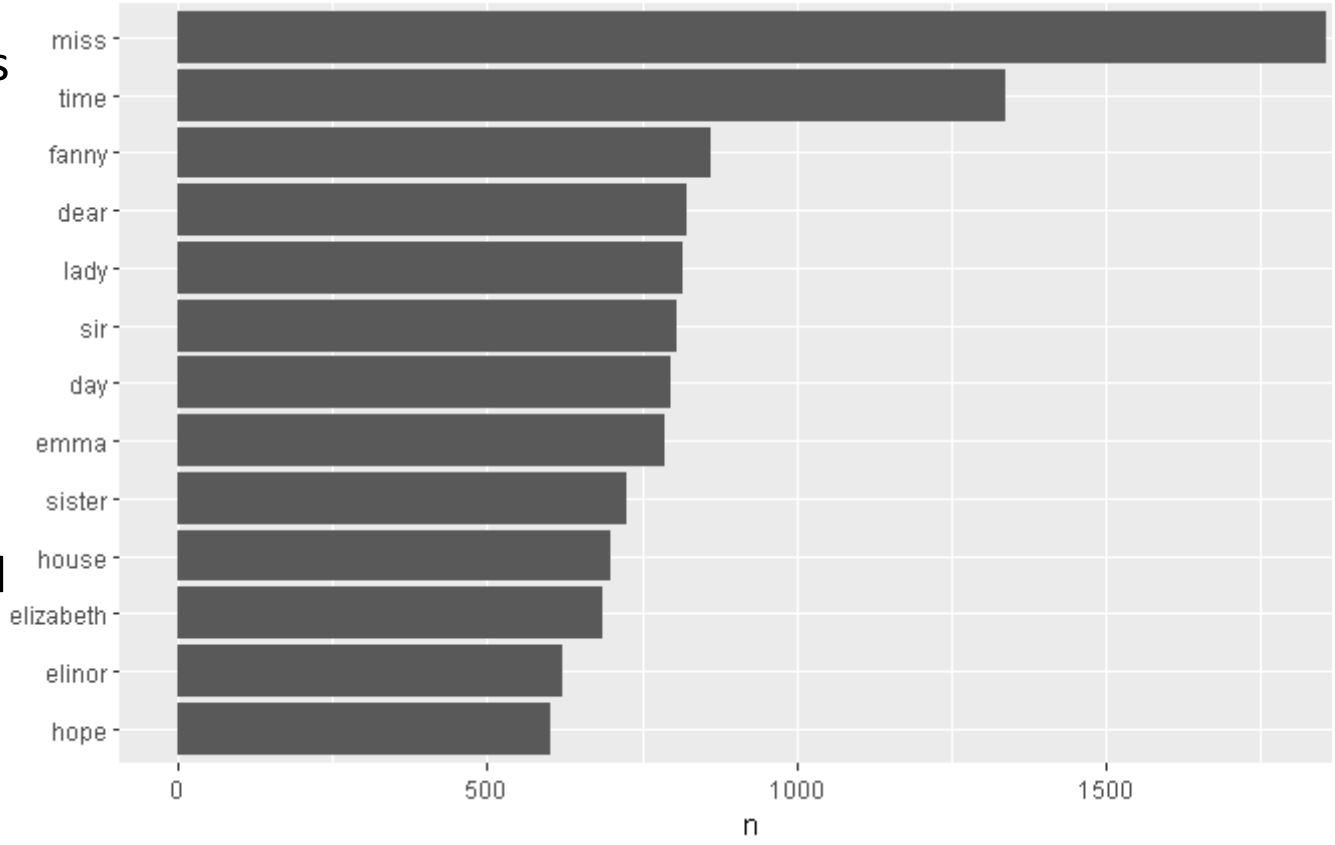
The stop_words dataset in the tidytext package contains stop words from three lexicons. We can use them all together, as we have here, or filter() to only use one set of stop words if that is more appropriate for a certain analysis.

University of South Wales
Prifysgol De Cymru

Introduction to tidytext

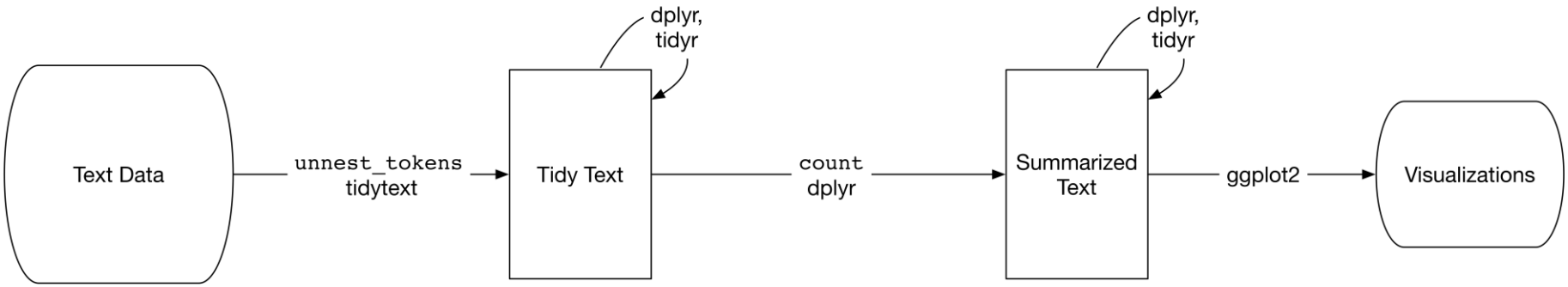
We can also use dplyr's count() to find the most common words in all the books as a whole.

Because we've been using tidy tools, our word counts are stored in a tidy data frame. This allows us to pipe this directly to the ggplot2 package.



University of South Wales
Prifysgol De Cymru

Introduction to tidytext



University of South Wales
Prifysgol De Cymru

Introduction to tidytext

There are a variety of methods and dictionaries that exist for evaluating the opinion or emotion in text.

The tidytext package contains several sentiment lexicons in the sentiments dataset.

abacus	trust	nrc	NA
abandon	fear	nrc	NA
abandon	negative	nrc	NA
abandon	sadness	nrc	NA
abandoned	anger	nrc	NA
abandoned	fear	nrc	NA
abandoned	negative	nrc	NA
abandoned	sadness	nrc	NA
abandonment	anger	nrc	NA

University of South Wales
Prifysgol De Cymru

Introduction to tidytext

With data in a tidy format, sentiment analysis can be done as an inner join.

First, let's use the NRC lexicon and filter() for the joy words.

Next, let's filter() the data frame with the text from the books for the words from Emma and then use inner_join() to perform the sentiment analysis.

What are the most common joy words in Emma?

University of South Wales
Prifysgol De Cymru

Introduction to tidytext

We can also examine how sentiment changes throughout each novel.

First, we find a sentiment score for each word using the Bing lexicon and inner_join().

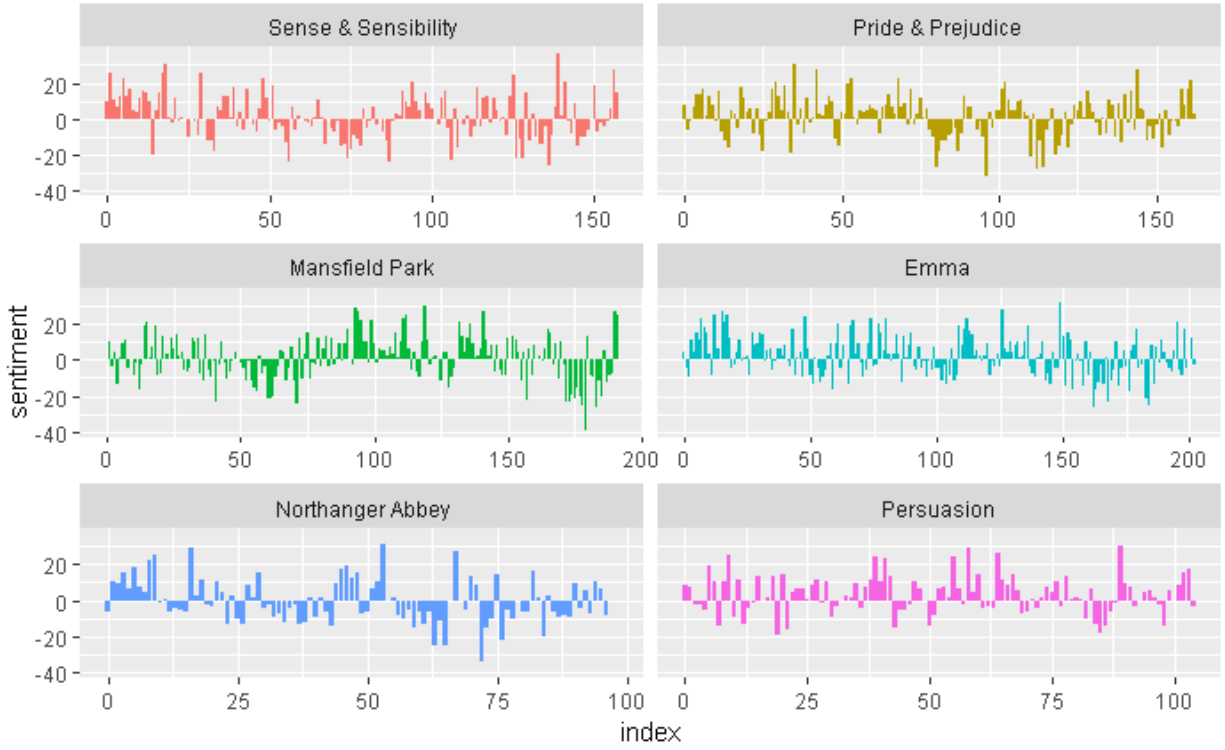
Next, we count up how many positive and negative words there are in defined sections of each book.

We define an index here to keep track of where we are in the narrative; this index counts up sections of 80 lines of text.

University of South Wales
Prifysgol De Cymru

Introduction to tidytext

Now we can plot these sentiment scores across the plot trajectory of each novel.



University of South Wales
Prifysgol De Cymru

Introduction to tidytext

Let's use all three sentiment lexicons and examine how the sentiment changes across the narrative arc of Pride and Prejudice.

First, let's use filter() to choose only the words from the one novel we are interested in.

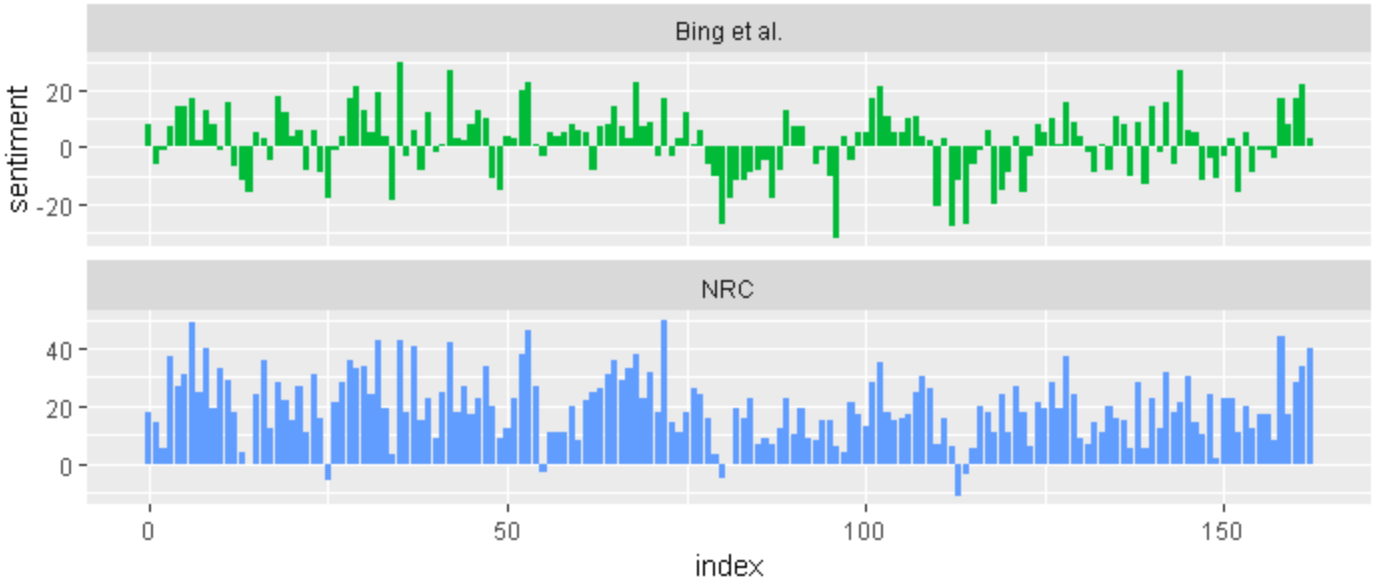
The AFINN lexicon measures sentiment with a numeric score between -5 and 5, while NCR and Bing lexicons categorise words in a binary fashion, either positive or negative.

To find a sentiment score in chunks of text throughout the novel, we will need to use a different pattern for the AFINN lexicon than for the other two.

University of South Wales
Prifysgol De Cymru

Introduction to tidytext

We now have an estimate of the net sentiment (positive - negative) in each chunk of the novel text for each sentiment lexicon.



University of South Wales
Prifysgol De Cymru

Introduction to tidytext

Exercise

1. Why are there differences in the sentiment scores for each lexicon over time?

2. What are the most common positive and negative words?

3. Identify how to add customer stop words.

4. Look to produce a word cloud of your analyses.