CBD2204: Week 11

T. Zourntos

unstructured data

The majority of data analysis occurs on structured data, or data that can be readily converted into data frames

However, there is a substantial amount of useful information that can be gleaned from unstructured data, such as articles, video or audio files, medical reports, speeches, financial statements, witness accounts, etc.

Text mining is a process by which prose (unstructured text) is processed for useful information; we often convert text into a kind of structured form, and then perform sentiment analysis, word correlations (N-grams),

text mining

tokenization

the process of parsing out text into meaningful units (think: strings separated by white space, for example) that are pertinent to our analysis

We can “*tidy*” text in this manner, allowing a **data frame** to be created from the text.

corpus

a group of unstructured text files (containing strings), often accompanied by metadata https://www.tidytextmining.com/

text mining cont’d

document-term matrix

a sparse matrix that describes a corpus with one row for each document and one column for each term

word counts and frequencies

a common form of text analysis which allows a straightforward quantification of unstructured text

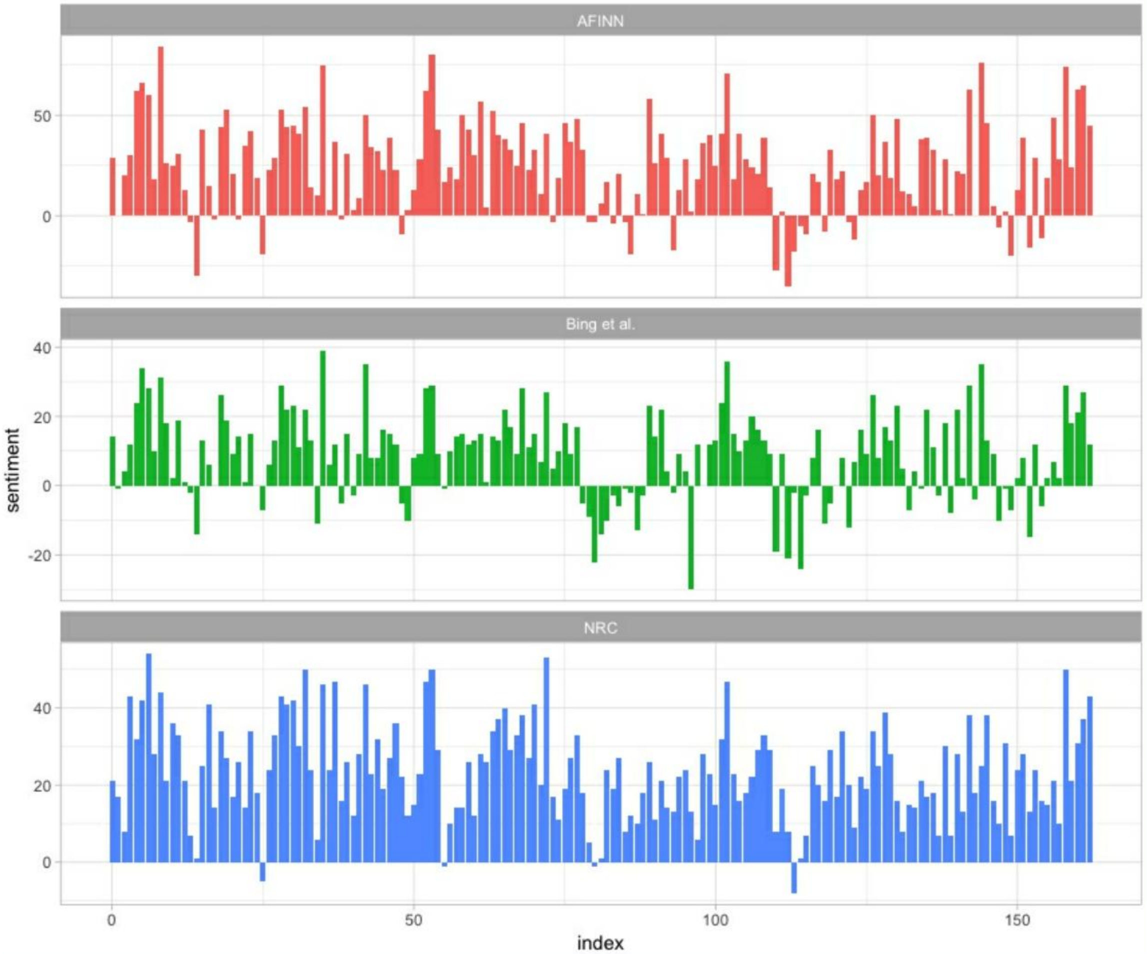


Figure 2-3. Comparing three sentiment lexicons using Pride and Prejudice

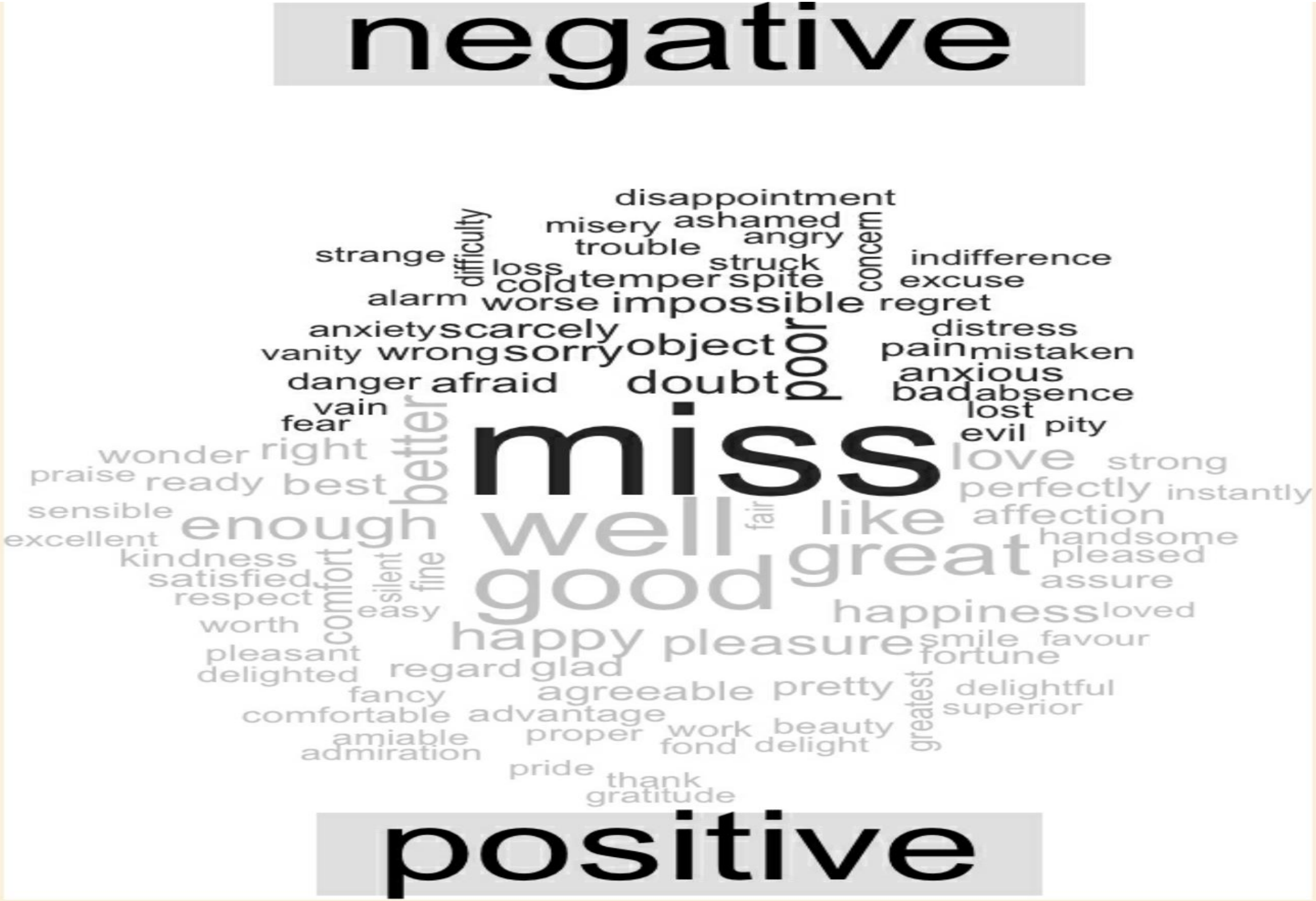
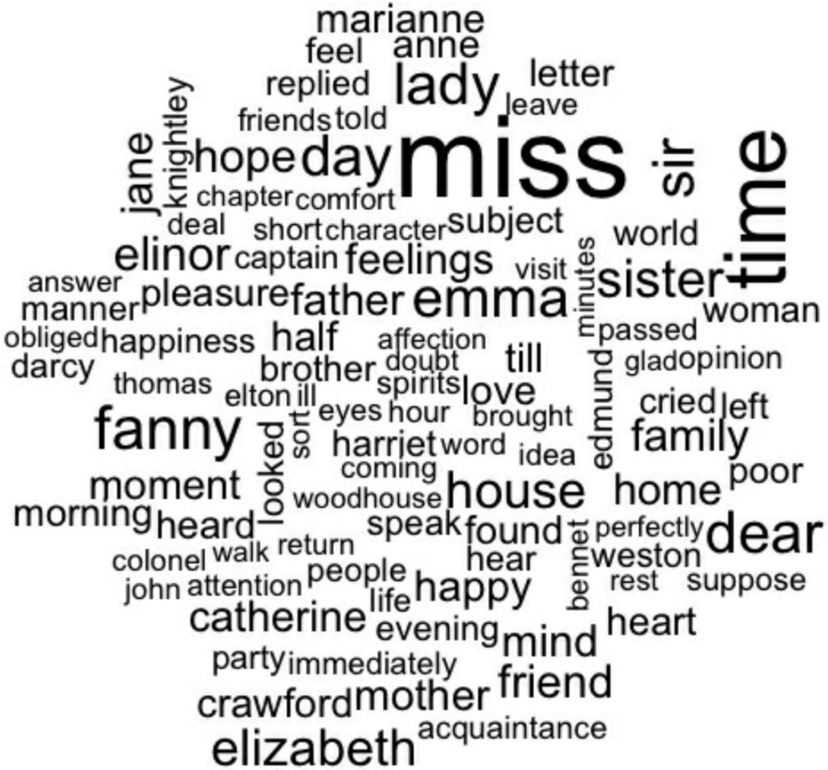


Figure 2-6. Most common positive and negative words in Jane Austen's novels


marianne
feel
anne
replied
lady
letter
leave
friends
told
knightley
jane
hope
day
miss
sir
time
chapter
comfort
deal
short
character
subject
world
elinor
captain
feelings
visit
minutes
sister
answer
pleasure
father
emma
manner
woman
obliged
happiness
half
affection
passed
darcy
brother
doubt
till
glad
opinion
thomas
spirits
love
edmund
elton
ill
cried
left
fanny
sort
eyes
hour
brought
looked
harriet
word
family
idea
coming
moment
house
home
poor
woodhouse
morning
heard
speak
found
perfectly
dear
bennet
colonel
walk
return
hear
weston
people
john
attention
happy
rest
suppose
life
catherine
evening
heart
mind
party
immediately
friend
crawford
mother
acquaintance
elizabeth