Computational Methods for Document Analysis 02

Mirco Dietrich
Maurice-Roman Isele

25. April 2016

Exercise 2

Problem 1:
People don't always use a white-space after a punctuation mark or use multiple marks. Replacing each punctuation mark by an empty string will glue the last and first word of sentences together. So we needed to replace each punctuation mark by a whitespace and then reduce multiple whitespaces to one.

Problem 2:
' Should be removed around words if it is used as a quotation mark, but shouldn't be removed when it is used as in a abbreviation(e.g. *"you're"*). Since there are not that many quotes with ' we decided to not remove it. There would be a solution to remove the' from quotes although it would be a bit of a hack:
You could read in a file where the String[] with the tokens was saved with Arrays.toString. Then replace ', (which is always at the end of a quote), *""*(which is always at the beginning of a quote) and all other no-characters with whitespaces, split at whitespaces and save the new tokens in the Review object.

Problem 3:
Removing ',' , '.' and '/' splits numbers separated by them(e.g. decimals, fractions). We don't have a solution for that.

Exercise 3

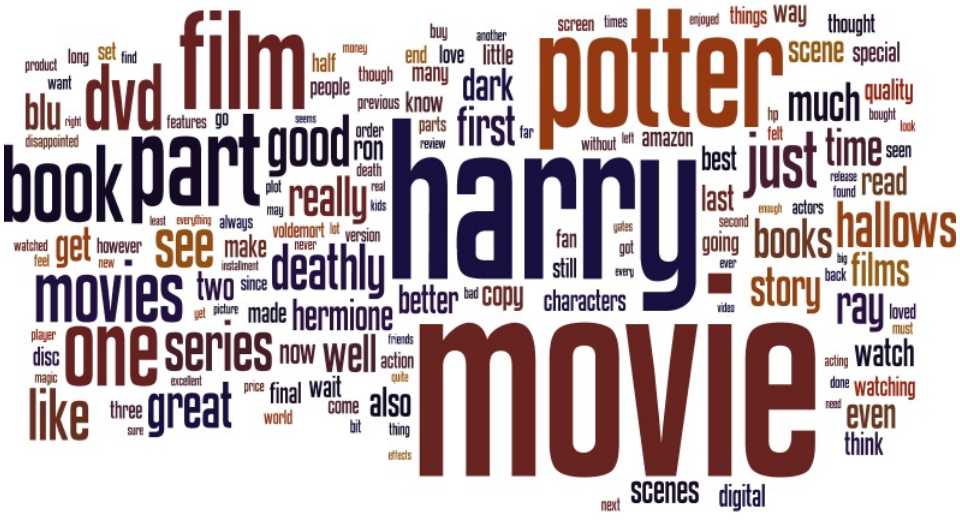
One of the obvious problems was to identify the abbreviations used by the people reviewing. We solved this by looking at the data and furthermore considering more common abbreviations used in the English language. Additionally, any name abbreviations are possible, e.g. *"Mr. K."*, *"T.J."*, and so on.

A more complex problem was to decide whether an abbreviation is the end of the sentence or not. For example, in *"[...] mental insanity, suicide, etc. However, if you don't want [...]"* the abbreviation is the end of a sentence, but in *[...] acting, storyline, etc. of this were good [...]* it is not. You can't decide this by only considering punctuation and a list of all abbreviations. You could say that an upper-case letter helps you identify the beginning of the next sentence in this scenario, but you have to consider 1) you always write names in upper-case (like New York) and 2) you can't count on the user to use correct spelling and grammar. That's why we never use dots in abbreviations to split the string.

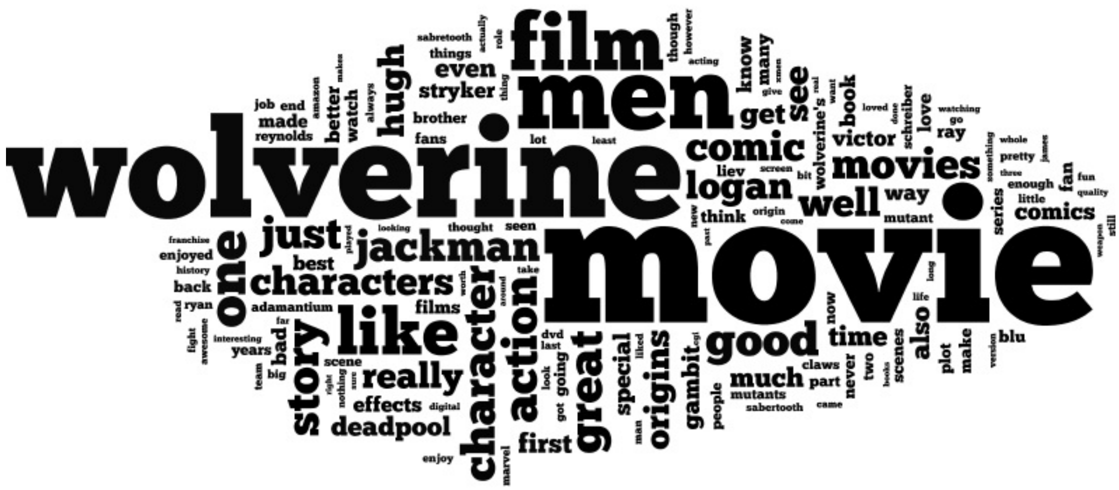
The last problem was dealing with multiple dots/question marks/exclamation marks, and again deciding whether the sentence ended after multiple dots, which leads to the same problem. With our current knowledge in text analysis (no grammar etc.), this seems to be the best we can do.

Exercise 4

For Harry Potter and the Deathly Hallows Part 1 (B004WO6BPS):



For X-Men Origins: Wolverine(B002P8ZOUQ);


wolverine
movie
men
film
comic
logan
movies
well
jackman
just
one
characters
like
story
character
action
great
origins
good
time
really
deadpool
first
special
gambit
much
hugh
even
stryker
see
get
victor
know
book
comics
series
also
make
plot
effects
adamantium
best
back
enjoyed
made
reynolds
watch
better
brother
fans
films
years
bad
scene
think
way
love
ray
mutant
fan
pretty
enough
little
fun
blu
scenes
claws
part
never
two
mutants
people
sabertooth
sabretooth
things
job
end
ryan
lot
seen
thought
origin
liev
schreiber
wolverine's
many
though
however
acting
always
amazon
franchise
history
read
fight
awesome
interesting
team
big
digital
enjoy
marvel
dvd
last
look
going
got
man
liked
now
life
long
came
whole
still
take
around
screen
bit
loved
watching
go
least
new
come
quality
version
worth