SIRCSS-CTA

Lab 1: Topic models

1 Topic Models

This assignment uses the R package `uuml` with data and functionality to simplify coding. To install the packages just run the following:

```
install.packages("remotes")
remotes::install_github("MansMeg/IntroML", subdir = "rpackage")
install.packages("tidytext")
install.packages("topicmodels")
```

We will now analyze the classical book Pride and Prejudice by Jane Austen using a probabilistic topic model. If you have not read the book, **here** you can read up on the story.

For this part of the assignment, Griffiths and Steyvers (2004) is the primary reference. I would also recommend reading Blei (2012) before starting with the assignment.

We will use a Gibbs sampler to estimate ten different topics occurring in Pride and Prejudice and study where they occur. A tokenized version of the book and a `data.frame` with stopwords can be loaded as follows:

```
library(uuml)
library(dplyr)

##
## Attaching package:  'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(tidytext)
data("pride_and_prejudice")
data("stopwords")
```

1. As a first step, we will remove stopwords (common English words without much semantic information):

   ```
   pap <- pride_and_prejudice
   pap <- anti_join(pap, y = stopwords[stopwords$lexicon == "snowball",])

   ## Joining with `by = join_by(word)`
   ```

2. Then we will remove rare words. Here we remove words that occur less than five times.

```r
word_freq <- table(pap$word)
rare_words <- data.frame(word = names(word_freq[word_freq <= 5]), stringsAsFactors = FALSE)
pap <- anti_join(pap, y = rare_words)

## Joining with `by = join_by(word)`
```

3. Now we have a corpus we can used to implement a probabilistic topic model. We do this by using the `topicmodels` R package. As a first step we will compute a document term matrix using the `tm` package, where we treat each paragraph as a document. How many documents and terms (word types) do you have?

```r
library(tm)
crp <- aggregate(pap$word, by = list(pap$paragraph), FUN = paste0, collapse = " ")
names(crp) <- c("paragraph", "text")
s <- SimpleCorpus(VectorSource(crp$text))
m <- DocumentTermMatrix(s)
```

4. To compute a topic model with ten topics, we use a Gibbs sampling algorithm. Below is an example of how we can run a Gibbs sampler for 2000 iterations. Run your topic model for 2000 iterations.

```r
library(topicmodels)
K <- 10
# Note: delta is beta in Griffith and Steyvers (2004) notation.
control <- list(keep = 1, delta = 0.1, alpha = 1, iter = 2000)
tm <- LDA(m, k = K, method = "Gibbs", control)
```

5. In the `uuml` R package you have three convenience functions to extract Θ, Φ and the log-likelihood values at each iteration. This is the parameter notation used in Griffiths and Steyvers (2004).

```r
library(uuml)
lls <- extract_log_liks(tm)
theta <- extract_theta(tm)
phi <- extract_phi(tm)
```

6. As a first step, check that the model has converged by visualizing the log-likelihood over epochs/iterations. Does it seem like the model have converged?

7. Extract the 20 top words for each topic (i.e. the words with the highest probability in each topic). Choose two topics you find coherent/best (the top words seem to belong together). Interpret these two topics based on the storyline of the book. What have these two topics captured?

8. Visualize these two topics evolve over the paragraphs in the books by plotting the θ parameters for that topic over time (paragraphs) in the book. Think of this as the time-line of the book. On the y-axis, you should plot θ_i for your chosen topic i and the x-axis should be the paragraph number (first paragraph has number 1 and so forth).

9. How do these two chosen topics evolve over the course in the book? If you want, you can take a rolling mean of the theta parameters to more easily show the changes in the topic over the book. *Hint!* Here `zoo::rollmean()` might be a good function to use.

10. Test to change the number of topics and do your own analysis of the novel when you feel you have a good number of topics.

References

Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.