Creative Component

Taikgun Song

The Trip Advisor reviews are scrapped and read into `R` (R Core Team 2014) using the `RCurl` (Lang 2015) package. Then the following `R` packages were utilized to conduct `Latent Dirichlet Allocation`(Blei, Ng, and Jordan 2003) method: `openNLP`(Hornik 2016a), `NLP` (Hornik 2016b), `topicmodels` (Bettina Grün 2016), `tm` (Ingo Feinerer 2015), `RTextTools` (Timothy P. Jurka 2014), and `SnowballC` (Bouchet-Valat 2014) packages.

1 September 9, 2016

1. Last week's Meeting ??? Setting up a GitHub repository. https://github.com/jeffsong9/creative Upload data file ??? Xkcd ??? Run LDA with different n-grams

2. This week

- Built a desktop PC
- Installed Linux
- Parsey McParseface
- Google Blog: https://research.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html
- Github: https://github.com/tensorflow/models/tree/master/syntaxnet

2 September 16, 2016

1. Last meeting

- List of stop words for English.
- Stemming
- What are k-grams? Try {1, 2, 3}-grams
- Change number of topics for LDA
- Document above in steps.
- Read LDA paper by Blei.
- Track R packages

2. This week

1) Stop words. "Stop words" are words frequently used in language, but with insignificant meanings. For example, POS such as articles or preposition is a good example of "stop words". Since LDA method estimates parameters in the model based on observed documents, meaningless "stop words" should be filtered out before applying text mining techniques. List of "stop words" in R package `tm` was used.

e.g.

```
## [1] "dtm <- DocumentTermMatrix(vdc,control=list(stopwords=T))"
```

`tm` has different list of "stop words", therefore I need to find out which one works the best for our research. The list could be selected by changing the "kind" variable.

e.g.

```
## [1] "stopwords(kind = "en"), stopwords(kind = "SMART")"
```

2) Stemming Stemming is the process of reducing inflected words to their word stem, base or root form. Since words with the same stem could be written in different POS, we could reduce the word dimension of a corpus by stemming with minimal loss of information. "stemDocument" function in "tm" package will be used.

e.g.

```
## [1] "stemDocument(vdc[[1]])"
```

3. Next Week

1) Read and summarize `Latent Dirichlet Allocation`(Blei, Ng, and Jordan 2003) by Blei, Ng, and Jordan.
2) Get familiar with the "tm" package by running toy dataset. Q: Is there any way that I could see the function written in packages? I have tried "?LDA" Q2: Any suggestion on how to compare different methods? E.g. Two way table?

References

Bettina Grün, Kurt Hornik. 2016. *Topic Models*. https://cran.r-project.org/web/packages/topicmodels/topicmodels.pdf.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (Jan): 993–1022.

Bouchet-Valat, Milan. 2014. *Snowball Stemmers Based on the c Libstemmer UTF-8 Library*. https://cran.r-project.org/web/packages/SnowballC/SnowballC.pdf.

Hornik, Kurt. 2016a. *Apache OpenNLP Tools Interface*. https://cran.r-project.org/web/packages/openNLP/openNLP.pdf.

———. 2016b. *Natural Language Processing Infrastructure*. https://cran.r-project.org/web/packages/NLP/NLP.pdf.

Ingo Feinerer, Artifex Software, Kurt Hornik. 2015. *Text Mining Package*. https://cran.r-project.org/web/packages/tm/tm.pdf.

Lang, Ducan Temple. 2015. *General Network Client Interface for R*. http://cran.r-project.org/web/packages/RCurl/RCurl.pdf.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Timothy P. Jurka, Amber E. Boydstun, Loren Collingwood. 2014. *Automatic Text Classification via Supervised Learning*. https://cran.r-project.org/web/packages/RTextTools/RTextTools.pdf.