lab2_w2w

Julia

10 June 2016

Необходимые библиотеки:

```
require(wordVectors)
require(magrittr)
require(parallel)
require(readr)
require(R2HTML)
```

```
setwd(paste0(getwd(), "/karpov/cl/"))
threads = detectCores() - 1
data_path <- "data/20news-18828/alt.atheism/"
```

Подготовим данные для модели (prep_word2vec) и натренируем модель (train_word2vec).

```
combined <- prep_word2vec(data_path, dest = "output/w2w.rdata", lowercase = T)
model <- train_word2vec("output/w2w.rdata", threads = threads,
                        vectors = 300, window = 10, min_count = 10, cbow = 1)
vectors <- read.vectors("output/w2w.bin")
```

Раз уж в корпус затесалось загадочное слово “koresh”, посмотрим, кто же наш “koresh”.

```
nearest_to(vectors, vectors[["koresh"]])
```

```
##       koresh         poor    obviously        crazy          mad
## 3.330669e-16 6.135415e-01 6.164470e-01 6.203700e-01 6.220130e-01
##       minded    perfectly      painful        drawn    matthew's
## 6.288322e-01 6.421016e-01 6.489997e-01 6.527582e-01 6.531338e-01
```

Попробуем сделать нечто более осмысленное и получить нечто близкое к каноническому king:man::queen:woman.

```
nearest_to(vectors, vectors[["god"]] - vectors[["man"]] + vectors[["woman"]])
```

```
##          god        woman       common     disbelief homosexuality
##    0.3032727    0.4319748    0.6002480    0.6171955    0.6464328
##     property        equal     interact        imply        bobby
##    0.6481558    0.6501676    0.6502079    0.6567215    0.6569752
```

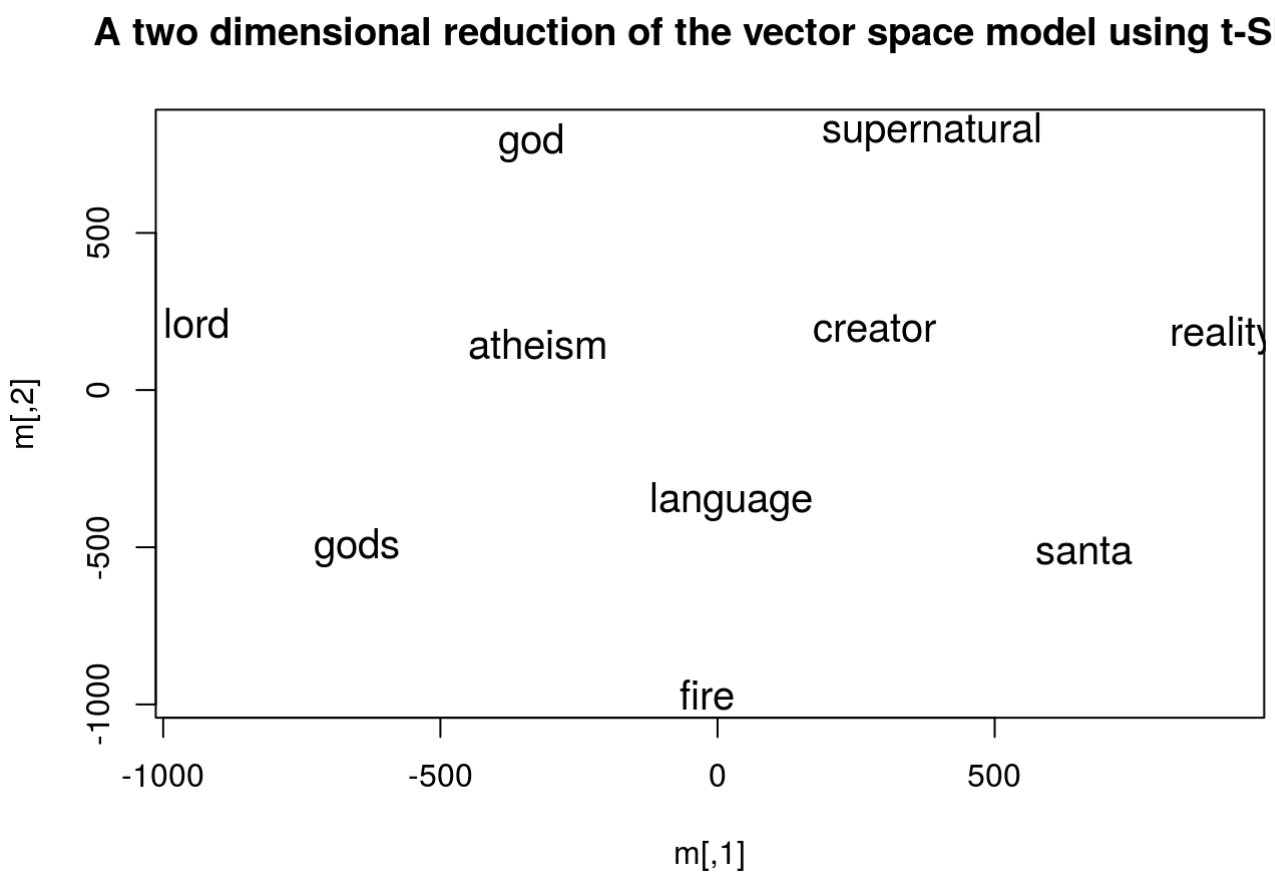
```
nearest_to(vectors, vectors[["god"]] - vectors[["he"]] + vectors[["she"]])
```

```
##          god          she    omniscient  supernatural        allah
##    0.2965296    0.3951148    0.6964717    0.6999147    0.7083979
## contradiction    damnation       present    dichotomy       choice
##    0.7289382    0.7297303    0.7348208    0.7359931    0.7393596
```

Видимо, в атеистских текстах богини не фигурируют.

Найдем максимально близкие к Богу слова.

```
god_names = nearest_to(vectors, vectors[[c("god","gods","lord","master")]], 10)
plot(filter_to_rownames(vectors, names(god_names)))
```



Санта?

Попробуем удалить всяческую семантическую близость с полом.

```
genderless = vectors[["god"]] %>%
  reject(vectors[["he"]]) %>%
  reject(vectors[["she"]])
vectors %>% nearest_to(genderless)
```

```
##          god  supernatural contradiction     explained          full
##    0.08593046    0.62207435    0.68210764    0.68499622    0.69370634
##         lines        virtue          type        favour        choice
##    0.70111381    0.70326472    0.70416501    0.70821222    0.71279868
```

…и получим ответ “ну, это уже фантастика!” (supernatural).

И напоследок попробуем отфильтровать зло =)

```
evilless = vectors[["god"]] %>%
  reject(vectors[["satan"]]) %>%
  reject(vectors[["evil"]]) %>%
  reject(vectors[["she"]])
vectors %>% nearest_to(evilless)
```

```
##          god     explained  supernatural        favour         bobby
##    0.1353126     0.6857409     0.6903467     0.6947599     0.6958194
##       prison        prayer        virtue         heart contradiction
##    0.7110499     0.7120191     0.7162337     0.7252499     0.7308400
```

Остается только сила и колледж! И болезни.