Moontrip: Testes `wordnet`

Para compilar este documento, correr em R:

```
library(knitr)
knit2pdf('wordnet.Rmd')
```

Para instalar `wordnet`, correr:

- no Ubuntu: `sudo apt-get install wordnet-base`
- depois no R: `install.packages('wordnet')`

Vamos definir a seguinte função para, dada uma palavra e um tipo de relação que queremos encontrar, pesquisar por quais palavras estão relacionadas dessas forma:

```
relacionadas <- function(palavra, tipo='NOUN', relacao='&')
{
    library('wordnet')
    filter <- getTermFilter('ExactMatchFilter', palavra, TRUE)
    terms <- getIndexTerms(tipo, 1, filter)
    synsets <- getSynsets(terms[[1]])
    related <- tryCatch(
        getRelatedSynsets(synsets[[1]], relacao),
        error=function(condition) NULL
    )
    if (is.null(related))
        return(NULL)
    sapply(related, getWord)
}
```

Para tipo usar um dos:

- `NOUN`
- `VERB`
- `ADJECTIVE`
- `ADVERB`

Para as relações possíveis ver na documentação do `wordnet`, secção WordNet Searches

Para ganharmos alguma sensibilidade sobre qual o tipo de relações que estão disponíveis, vamos definir a seguinte função:

```
s <- c('!','@',' ','*','&','#m','#s','#p','%m','%s','%p','%','#','>','<','^','\\','\\=','$','+',';','-'

st <- c('antonyms','hypernyms','hyponyms','entailment','similar','member meronym','substance meronym','

relacoes <- function(palavra, tipo='NOUN')
{
    for(i in 1:length(s)) {
        ws <- relacionadas(palavra, tipo, s[i])
        if(!is.null(ws)) {
            print('')
```

```
            print(paste('**', toupper(st[i])))
            print(str(ws))
        }
    }
}
```

Podemos começar assim a fazer alguns testes:

```
relacoes('surf')
```

```
## [1] ""
## [1] "** HYPERNYMS"
##  chr [1:2, 1] "wave" "moving ridge"
## NULL
## [1] ""
## [1] "** DERIVATIONALLY RELATED FORM"
## List of 2
##  $ : chr "break"
##  $ : chr [1:2] "surfboard" "surf"
## NULL
```

Temos portanto hiperónimos e surfboard como palavra relacionada. Talvez interessante para dizer ao utilizador: será que quer surfboard, de forma a ajudar a ser mais específico. Mas nada que agrupe como desporto, por exemplo.

```
relacoes('sport')
```

```
## [1] ""
## [1] "** HYPERNYMS"
##  chr [1:2, 1] "diversion" "recreation"
## NULL
## [1] ""
## [1] "** DERIVATIONALLY RELATED FORM"
## List of 2
##  $ : chr [1:3] "acrobatic" "athletic" "gymnastic"
##  $ : chr [1:12] "frolic" "lark" "rollick" "skylark" ...
## NULL
## [1] ""
## [1] "** SHOW DOMAIN TERMS FOR TOPIC"
## List of 117
##  $ : chr "in play(p)"
##  $ : chr "out of play(p)"
##  $ : chr [1:2] "one-on-one" "man-to-man"
##  $ : chr "loose"
##  $ : chr "legal"
##  $ : chr "disqualified"
##  $ : chr "home(a)"
##  $ : chr "away"
##  $ : chr "most-valuable"
##  $ : chr "ineligible"
##  $ : chr "defending"
##  $ : chr "onside"
```

```
##  $ : chr [1:2] "offside" "offsides"
##  $ : chr [1:3] "underhand" "underhanded" "underarm"
##  $ : chr [1:3] "overhand" "overhanded" "overarm"
##  $ : chr "upfield"
##  $ : chr "downfield"
##  $ : chr "downfield"
##  $ : chr "at home"
##  $ : chr "offside"
##  $ : chr "wipeout"
##  $ : chr [1:3] "pass" "toss" "flip"
##  $ : chr "daisy cutter"
##  $ : chr "call"
##  $ : chr [1:2] "birling" "logrolling"
##  $ : chr [1:2] "stroke" "shot"
##  $ : chr "position"
##  $ : chr "foul"
##  $ : chr "personal foul"
##  $ : chr "possession"
##  $ : chr "save"
##  $ : chr "press box"
##  $ : chr "tuck"
##  $ : chr "game plan"
##  $ : chr "won-lost record"
##  $ : chr [1:2] "English" "side"
##  $ : chr "series"
##  $ : chr "trial"
##  $ : chr [1:3] "defense" "defence" "defending team"
##  $ : chr "bench warmer"
##  $ : chr [1:3] "coach" "manager" "handler"
##  $ : chr "free agent"
##  $ : chr [1:2] "iron man" "ironman"
##  $ : chr [1:2] "referee" "ref"
##  $ : chr [1:2] "scout" "talent scout"
##  $ : chr "shooter"
##  $ : chr [1:2] "timekeeper" "timer"
##  $ : chr "deficit"
##  $ : chr "lead"
##  $ : chr "average"
##  $ : chr "free agency"
##  $ : chr "regulation time"
##  $ : chr "sudden death"
##  $ : chr [1:3] "turn" "bout" "round"
##  $ : chr "surge"
##  $ : chr "seed"
##  $ : chr "outclass"
##  $ : chr "call"
##  $ : chr "curl"
##  $ : chr "start"
##  $ : chr "field"
##  $ : chr "shoot"
##  $ : chr [1:2] "referee" "umpire"
##  $ : chr "drop"
##  $ : chr "down"
##  $ : chr "bandy"
```

```
##  $ : chr "double-team"
##  $ : chr "submarine"
##  $ : chr "kick"
##  $ : chr "punt"
##  $ : chr "follow through"
##  $ : chr "kill"
##  $ : chr "kill"
##  $ : chr "drive"
##  $ : chr "racket"
##  $ : chr [1:2] "dribble" "carry"
##  $ : chr "cut"
##  $ : chr "box"
##  $ : chr "spar"
##  $ : chr "spar"
##  $ : chr "prizefight"
##  $ : chr "shadowbox"
##  $ : chr "tramp"
##  $ : chr "hike"
##  $ : chr "mountaineer"
##  $ : chr [1:3] "rappel" "abseil" "rope down"
##  $ : chr [1:2] "backpack" "pack"
##  $ : chr "run"
##  $ : chr "jog"
##  $ : chr "skate"
##  $ : chr "spread-eagle"
##  $ : chr "ice skate"
##  $ : chr "figure skate"
##  $ : chr "roller skate"
##  $ : chr "skateboard"
##  $ : chr "Rollerblade"
##  $ : chr "speed skate"
##  $ : chr "ski"
##  $ : chr "schuss"
##   [list output truncated]
## NULL
```

Aqui está cortado, mas este já tem como "termos de domínio por tópico", o "surf". Às tantas, se quisermos por alguma razão saber o grupo maior duma palavra teremos que fazer alguma ginástica, começando com algumas categorias raíz e tentando derivá-la daí.

```
relacoes('tennis')
```

```
## [1] ""
## [1] "** HYPERNYMS"
##  chr "court game"
## NULL
## [1] ""
## [1] "** PART HOLONYM"
##  chr [1:5] "footfault" "return" "service break" "advantage" ...
## NULL
## [1] ""
## [1] "** SHOW DOMAIN TERMS FOR TOPIC"
## List of 13
```

```
##  $ : chr "double fault"
##  $ : chr [1:2] "break" "break of serve"
##  $ : chr [1:2] "cut" "undercut"
##  $ : chr "drive"
##  $ : chr [1:3] "forehand" "forehand stroke" "forehand shot"
##  $ : chr "forehand drive"
##  $ : chr [1:2] "serve" "service"
##  $ : chr "fault"
##  $ : chr [1:2] "rally" "exchange"
##  $ : chr "match point"
##  $ : chr "game"
##  $ : chr "ace"
##  $ : chr "drop one's serve"
## NULL
```

Nada sobre ping-pong. :P

```
relacoes('ski')
```

```
## [1] ""
## [1] "** HYPERNYMS"
##  chr "runner"
## NULL
## [1] ""
## [1] "** DERIVATIONALLY RELATED FORM"
##  chr "ski"
## NULL
```

Muito pobre.

```
relacoes('wine')
```

```
## [1] ""
## [1] "** HYPERNYMS"
##  chr [1:5, 1] "alcohol" "alcoholic drink" "alcoholic beverage" ...
## NULL
## [1] ""
## [1] "** SUBSTANCE MERONYM"
##  chr [1:2] "grape" "negus"
## NULL
## [1] ""
## [1] "** DERIVATIONALLY RELATED FORM"
## List of 7
##  $ : chr [1:2] "winy" "winey"
##  $ : chr [1:2] "winy" "winey"
##  $ : chr [1:2] "vinous" "vinaceous"
##  $ : chr [1:2] "vinous" "vinaceous"
##  $ : chr "wine"
##  $ : chr "wine"
##  $ : chr "vinify"
## NULL
```

```
relacoes('food')
```

```
## [1] ""
## [1] "** HYPERNYMS"
##  chr "substance"
## NULL
## [1] ""
## [1] "** PART HOLONYM"
##  chr [1:2, 1] "food" "solid food"
## NULL
## [1] ""
## [1] "** DERIVATIONALLY RELATED FORM"
## List of 2
##  $ : chr [1:6] "alimentary" "alimental" "nourishing" "nutrient" ...
##  $ : chr [1:3] "nutrify" "aliment" "nourish"
## NULL
```

```
relacoes('sex')
```

```
## [1] ""
## [1] "** HYPERNYMS"
##  chr [1:4, 1] "bodily process" "body process" "bodily function" ...
## NULL
## [1] ""
## [1] "** DERIVATIONALLY RELATED FORM"
## List of 2
##  $ : chr "sexual"
##  $ : chr [1:5] "arouse" "sex" "excite" "turn on" ...
## NULL
```

Nada que defina tipos tipos de *sex*, nomeadamente:

```
relacoes('masochism')
```

```
## [1] ""
## [1] "** HYPERNYMS"
##  chr "sexual pleasure"
## NULL
## [1] ""
## [1] "** DERIVATIONALLY RELATED FORM"
##  chr [1:2] "masochistic" "masochist"
## NULL
```

Mais uma vez, não agrupa na categoria pai de "sex".

Outros pensamentos

Caso o "stemming" também nos funcione mal, podemos considerar o wordnet para isso. Também poderá talvez ser usado para inferir emoções da página, uma vez que apenas queremos destinos "positivos". Por exemplo, uma página pode referir muitas vezes a palavra *gay*, mas de forma negativa, e não a queremos então recomendar.