Benjamin Schlegel

27. September 2017

Dieser Beitrag erklärt, wie man einen Text in seine Sätze aufspalten kann in R. Der hier angegebene Code funktioniert in vielen Fällen recht gut, ist aber bei weitem noch nicht perfekt.

Als erstes wird der Text eingelesen.

```
txt = "Politische Philosophie ist deshalb ferner ein Gebiet der praktischen Philosophie, in welchem normative Fragen untersucht werden. Dabei werden Gesichtspunkte der Moralphilosophie und der angewandten Ethik mit der politischen Theorie verbunden, wobei in der Regel eine Reflexion auf die politische Ideengeschichte erfolgt. Wie eingangs bereits angesprochen, besteht ihre Aufgabe daher in der Kritik, der Sinngebung und der Wegweisung des politischen, d.h. im weiteren Sinne, des menschlichen Handelns schlechthin."
```

Anschliessend wird der Text in seine Wörter aufgetrennt.

```
words = unlist(strsplit(txt," "))
```

Mit folgender Funkton werden die Wörter in Sätzen zusammengefügt. Die Abkürzungsliste ist nicht vollständig, kann aber problemlos erweitert werden. Für andere Sprachen als Deutsch, müsste die Abkürzungsliste vollständig neu geschrieben werden.

```
get.sentences.de = function(word.list){
  abbr.list = c("a.a.O.", "a.","Abb.","Abf.","Abk.", "Abs.", "Abt.", "abzgl.", "a.D.", "a.D.", "Adr.", "a.M.", "am.",
                "amtl.", "amtl.", "Anh.","Ank.","Anl.", "Anm.", "a.Rh.", "A.T.", "Aufl.", "b.", "B.", "Bd.", "beil.",
                "bes.","Best.-Nr.", "Betr.", "Bez.", "Bhf.", "b.w.", "bzgl.", "bzw.", "ca.", "Chr.", "d.Ä.", "dgl.",
                "d.h.", "Dipl.-Ing.", "Dipl.-Kfm.", "Dir.", "d.J.", "Dr.", "dt.", "Dtzd.", "e.h.", "ehem.", "eigtl.",
                "einschl.", "entspr.", "erb.", "erw.", "Erw.", "ev.", "e.V.", "evtl.", "e.Wz.", "exkl.", "f.", "Fa.",
                "Fam.", "F.f.", "Ffm.", "Forts.", "Fr.", "Frl.", "frz.", "geb.", "Gebr.", "gedr.", "gegr.", "gek.",
                "Ges.", "gesch.", "gest.", "gez.", "ggf.", "ggfs.", "Hbf.", "hpts.", "Hptst.", "Hr.", "Hrn.", "Hrsg.",
                "i.A.", "i.b.", "i.B.", "i.H.", "i.J.", "Ing.", "Inh.", "inkl.", "i.R.", "i.V.", "jew.", "Jh.",
                "jhrl.", "Kap.", "kath.", "Kfm.", "kfm.", "kgl.", "k.o.", "K.o.", "k.u.k.", "l.", "led.", "m.E.", "med.",
                "Mio.", "möbl.", "Mrd.", "Msp.", "mtl.", "m.", "M.", "m.W.", "MwSt.", "MWSt.", "näml.", "nat.", "n.Chr.", "Nr.", "n.u.Z.",
```

```
                "o.", "o.A.", "o.B.", "Obb.", "od.", "o.g.", "österr.", "p.Adr.",
"phil.", "Pfd.", "Pl.", "r.", "Reg.-Bez.", "rer.",
                "r.k.", "r.-k.", "röm.", "röm.-kath.", "S.", "s.", "s.a.", "Sa.",
"schles.", "schwäb.", "schweiz.", "s.o.",
                "So.", "sog.",  "St.", "Str.", "StR.", "s.u.", "südd.", "tägl.", "u.",
"ü.", "u.a.", "u.ä.", "u.Ä.", "u.a.m.",
                "u.A.w.g.", "usw.", "u.v.a.", "u.U.", "V.", "v.Chr.", "Verf.", "verh.",
"verw.", "vgl.", "v.H.", "vorm.",
                "v.R.w.", "v.T.", "v.u.Z.", "z.", "z.B.", "z.Hd.", "Zi.", "zur.",
"zus.", "z.T.", "Ztr.", "zzgl.", "z.Z.")
  result = c()
  sentence = ""
  for(w in words){
    sentence = paste0(sentence, " ", w)
    if(!(w %in% abbr.list) && grepl("\\.|\\?|\\!", w) && !grepl("[0-9]+\\.", w)){
      result = c(result, trimws(sentence))
      sentence = ""
    }
  }
  result
}
```

Nun kann die Funktion verwendet werden, um die Sätze zu erhalten.

```
sentences = get.sentences.de(words)
sentences
```

```
[1] "Politische Philosophie ist deshalb ferner ein Gebiet der praktischen Philosophie,
in welchem normative Fragen untersucht werden."

[2] "Dabei werden Gesichtspunkte der Moralphilosophie und der angewandten Ethik mit der
politischen Theorie verbunden, wobei in der Regel eine Reflexion auf die politische
Ideengeschichte erfolgt."
[3] "Wie eingangs bereits angesprochen, besteht ihre Aufgabe daher in der Kritik, der
Sinngebung und der Wegweisung des politischen, d.h. im weiteren Sinne, des menschlichen
Handelns schlechthin."
```

Hast du Ideen für Verbesserungen der Funktion? Poste deine Vorschläge in die Kommentarspalte.