Universität Potsdam
Computerlinguistische Techniken
Dozent:
Prof. Dr. David Schlangen
Wintersemester 2020/21

Bericht

Terminologie Extraktion

Name: Katja Konermann
Martikelnummer: 802658
Email: katja.konermann@uni-potsdam.de

Inhaltsverzeichnis

1 Kandidatenauswahl **1**

2 Korpus **1**

3 Anwendung **2**

4 Ergebnisse und Bewertung **2**

1 Kandidatenauswahl

Zur Auswahl der Kandidaten wurden verschiedene Parameter genutzt. So werden alle Bigramme herausgefiltert, in denen eines oder beider der Token in einer Menge von Stoppwörtern vorkommt. Ich habe dafür die englischen Stoppwörter von *nltk* verwendet. Im Projekt sind sie in der Datei *data/stops_en.txt* zu finden.
Weil Fachbegriffe zumeist aus Nomen bestehen sollten, habe ich zudem Tagging verwendet. Dabei muss ein Bigramm aus zumindest einem relevanten Tag bestehen, um als Kandidat in Frage zu kommen. Mithilfe von *nltk*s *Averaged Perceptron Tagger* werden so alle Bigramme getaggt und die Bigramme, die keinerlei relevante Tags enthalten, herausgefiltert.
Da das *acl* Korpus durch *optical character recognition* erstellt wurde, sind viele Zeichen vorhanden, die keinerlei Bedeutung haben. Um diese Bigramme herauszufiltern, wird bei der Kandidatenauswahl zusätzlich getestet, ob ein Bigramm aus Token besteht, die ausschließlich alphabetisch sind.
Die Anzahl der Kandidaten kann außerdem reduziert oder erhöht werden, indem ein Minimum für die absolute Häufigkeit eines Bigramms festgelegt wird.
Die Kandidaten, die schließlich für die Terminologieextraktion genutzt wurde, sind in der Datei *data/candidates.txt* gespeichert. Um diese Liste von Termen zu reproduzieren, können folgende Argumente bei der Ausführung von *main.py* an die Kommandozeile übergegeben werden:

```
candidates --stops data/stops_en.txt --min_count 2 acl_texts/ <file> NN NNS NNP
```

Dabei sollte <*file*> durch den gewünschten Namen der Ausgabedatei ersetzt werden. Für genauere Information zu den einzelnen Argumenten siehe Abschnitt *Anwendung* oder die *README*.

2 Korpus

Dokumente	10 922
Sätze	1 575 233
Token	42 482 606
Types	520 446
Kandidaten (Bigramme)	341 517

Tabelle 1: Korpus Statistiken

Datei	α	θ	Fachbegriffe
output/output1.csv	0.5	2	20 404
output/output2.csv	0.75	1.75	4 957
output/output3.csv	0.25	2.75	13 495
output/output4.csv	0.9	1.25	9 965
output/output5.csv	0.1	3	15 818

Tabelle 2: Anzahl der Fachbegriffe

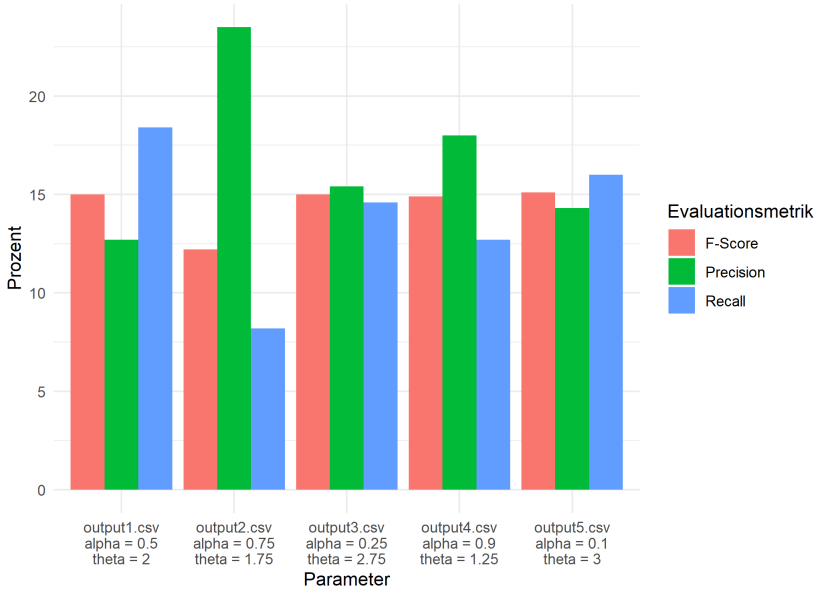

Abbildung 1: Accuracy verschiedener Parameter

3 Anwendung

4 Ergebnisse und Bewertung