7. Übung

Prokhar Navitski, 818431

February 1, 2024

1 Distributionelle Semantik

***Ich musste den Wert für **window size** auf 5 reduzieren, da mein Code sonst die Fehlermeldung **Process finished with exit code 137 (interrupted by signal 9: SIGKILL)** zurückgab. Ich denke, das liegt an den Beschränkungen meines Systems.

1.1 Aufgabe a

1.2 Aufgabe b

Nach mehreren Durchläufen des Programms mit nur Standard-Stoppwörtern aus der NLTK-Bibliothek erhielt ich verschiedene Listen der 10 häufigsten Wörter und auf der Suche nach Symbolen oder funktionalen Partikeln in ihnen erweiterte ich die Liste der Stoppwörter um folgende: additional$_s$*topwords* =″ ....′,′ −−′,′″″ ,′″″ ,″ *s*″,′ *said*′,′ *mr*′,′ *mrs*′,′ ,′″ ′.″,′ ;″ ,′ ?″,′ !″,′″ ′*one*′,′ *would*′,′)″,′ (″,′ :″ ,′ *new*′.

Aber auch wenn die Zeichen "." und "," in der Liste der Stoppwörter enthalten sind, werden sie vom Programm bei der Ausführung der Funktion "find-most-common-words" berücksichtigt. Ich denke, das liegt daran, dass die Methode **for word in brown.words()** diese beiden Zeichen nicht als Wörter betrachtet und sie daher nicht ignoriert, obwohl sie in der erweiterten Stoppwortliste enthalten sind.

1.3 Aufgabe c

"Window-Size" = 5

"good-evil" 10 most common words INCLUDING stopwords: ['the', ',', '.', 'of', 'and', 'to', 'a', 'in', 'that', 'is'] The similarity between "good" and "evil" INCLUDING stopwords is 0.8231494666515363. 10 most common words WITHOUT stopwords: [',', '.', 'could', 'time', 'two', 'may', 'first', 'like', 'man', 'even'] The similarity between "good" and "evil" WITHOUT stopwords is 0.8231494666515363.

'state', 'country' 10 most common words INCLUDING stopwords: ['the', ',', '.', 'of', 'and', 'to', 'a', 'in', 'that', 'is'] The similarity between "state" and "country" INCLUDING stopwords is 0.9617716459409003. 10 most common

words WITHOUT stopwords: [',', '.', 'could', 'time', 'two', 'may', 'first', 'like', 'man', 'even'] The similarity between "state" and "country" WITHOUT stopwords is 0.9617716459409003.

'fire', 'water' 10 most common words INCLUDING stopwords: ['the', ',', '.', 'of', 'and', 'to', 'a', 'in', 'that', 'is'] The similarity between "fire" and "water" INCLUDING stopwords is 0.949119563328333. 10 most common words WITHOUT stopwords: [',', '.', 'could', 'time', 'two', 'may', 'first', 'like', 'man', 'even'] The similarity between "fire" and "water" WITHOUT stopwords is 0.949119563328333.

"Window-Size" = 3

"good-evil" 10 most common words INCLUDING stopwords: ['the', ',', '.', 'of', 'and', 'to', 'a', 'in', 'that', 'is'] The similarity between "good" and "evil" INCLUDING stopwords is 0.7344583771801991. 10 most common words WITHOUT stopwords: [',', '.', 'could', 'time', 'two', 'may', 'first', 'like', 'man', 'even'] The similarity between "good" and "evil" WITHOUT stopwords is 0.7344583771801991.

'state', 'country' 10 most common words INCLUDING stopwords: ['the', ',', '.', 'of', 'and', 'to', 'a', 'in', 'that', 'is'] The similarity between "state" and "country" INCLUDING stopwords is 0.9414808848482041. 10 most common words WITHOUT stopwords: [',', '.', 'could', 'time', 'two', 'may', 'first', 'like', 'man', 'even'] The similarity between "state" and "country" WITHOUT stopwords is 0.9414808848482041.

'fire', 'water' 10 most common words INCLUDING stopwords: ['the', ',', '.', 'of', 'and', 'to', 'a', 'in', 'that', 'is'] The similarity between "fire" and "water" INCLUDING stopwords is 0.9168388452341972. 10 most common words WITHOUT stopwords: [',', '.', 'could', 'time', 'two', 'may', 'first', 'like', 'man', 'even'] The similarity between "fire" and "water" WITHOUT stopwords is 0.9168388452341972.

Leider kann ich aufgrund der begrenzten Rechenleistung meines Rechners nicht alle Kombinationen von Wortpaaren und Fenstergrößen, die von Interesse sind, effizient und schnell testen. Aus den obigen Ergebnissen geht jedoch klar hervor, dass Wörter, die polare Antonyme zueinander sind, einen geringeren Grad an Kosinusgleichheit aufweisen. Dies ist höchstwahrscheinlich darauf zurückzuführen, dass sie von annähernd denselben Wörtern umgeben sind, so dass der Winkel des Auftretens auf dem Graphen zwischen den Wörtern "good" und "evil" größer ist als z. B. bei den Paaren "state" und "country", die eher als Synonyme in Bezug aufeinander verwendet werden. Und da der Winkel größer ist, wird der Kosinus (der Wert der Gleichheit) dieser Wörter kleiner sein.

Die Tatsache, dass die Wörter "good" und "evil" je nach ihrer Position im Satz die Rolle verschiedener Wortteile spielen können, während die Wörter "state" und "country" eher auf die Rolle von Substantiven beschränkt sind, kann ebenfalls Einfluss auf die Verringerung der Bedeutung von "gleich" haben.