Semantische Beziehungen in Texten mit Word2Vec

B A C H E L O R A R B E I T

im Studiengang
Medieninformatik (MI7)
an der Hochschule der Medien in Stuttgart
vorgelegt von Ruben Müller

im Juli 2015

Erstprüfer: Prof. Dr-Ing. Johannes Maucher,
Hochschule der Medien, Stuttgart

Zweitprüfer: M.Sc. Andreas Stiegler,
Hochschule der Medien, Stuttgart

Erklärung

Hiermit versichere ich, Ruben Müller, an Eides Statt, dass ich die vorliegende Bachelorarbeit mit dem Titel: SSemantische Beziehungen in Texten mit Word2Vecßelbständig und ohne fremde Hilfe verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen wurden, sind in jedem Fall unter Angabe der Quelle kenntlich gemacht. Die Arbeit ist noch nicht veröffentlicht oder in anderer Form als Prüfungsleistung vorgelegt worden.
Ich habe die Bedeutung der eidesstattlichen Versicherung und die prüfungsrechtlichen Folgen (§ 23 Abs. 2 Bachelor-SPO (7 Semester) der HdM) sowie die strafrechtlichen Folgen (gem. § 156 StGB) einer unrichtigen oder unvollständigen eidesstattlichen Versicherung zur Kenntnis genommen.

Filderstadt, den XX. Juli 2015

Ruben Müller

Kurzfassung

Abstract

Inhaltsverzeichnis

Begriffsverzeichnis

Begriff	Erklärung
Ähnliche Worte	Im Word2Vec-Modell mit der Methode *most_similar*() erhaltene Worte.
SVM	Support Vector Machine
NBC	Naive Bayes Classifier

Kapitel 1

Einleitung

1.1 Motivation

1.2 Problemstellung

1.3 Aufbau der Arbeit

Kapitel 2

Daten und Vorverarbeitung

2.1 Datenbasis

2.2 Externe Programme und Hilfsmittel

2.3 Vorverarbeitung

Kapitel 3

Word2Vec

3.1 Parameter

3.2 CBOW

3.3 Skip-gram

3.4 Negative sampling

3.5 Hierarchical softmax

3.6 Distanz zwischen Vektoren im Word2Vec Modell

Kapitel 4

Wikipedia-Korpus

4.1 Gesamtkorpus

4.2 Teilkorpus

4.3 Testdaten

4.3.1 Vergleich und Analyse

Kapitel 5

Experimente

In diesem Kapitel sollen die unterschiedlichen Korpora (Gesamtkorpus[1] und Techkorpus[2]) untersucht werden. Dies soll durch ausgewählte Fragestellungen realisiert werden.
Die Fragestellungen beziehen sich immer auf die Ergebnisse, die aus den Tastdaten[3] erhaltenen ähnlichen Worten.
Jedes Experiment ist in drei Teile aufgeteilt Beschreibung, Durchführung und Interpretation/Ergebnis.

5.1 Synnonymsuche durch Rekursion

5.1.1 Beschreibung

Es soll untersucht werden, ob man Synonyme zum Testwort erhält, wenn man die ähnlichen Worte dieses Testwortes erneut im Model mittels der Methode *most_similar*() sucht.

5.1.2 Durchführung

.

5.1.3 Interpretation/Ergebnis

.

[1] vgl. 4.1
[2] vgl. 4.2
[3] vgl. 7.1

5.2 Konkretisierungen

5.2.1 Beschreibung

.

5.2.2 Durchführung

.

5.2.3 Interpretation/Ergebnis

.

5.3 Verallgemeinerungen

5.3.1 Beschreibung

.

5.3.2 Durchführung

.

5.3.3 Interpretation/Ergebnis

.

5.4 Unterschiedliche Beziehungen

5.4.1 Beschreibung

.

5.4.2 Durchführung

.

5.4.3 Interpretation/Ergebnis

.

5.5 Mehrdeutigkeit

5.5.1 Beschreibung

.

5.5.2 Durchführung

.

5.5.3 Interpretation/Ergebnis

.

Kapitel 6

Fazit und Ausblick

6.1 Fazit

6.2 Ausblick

Kapitel 7

Anhang

7.1 Testdaten

3d	3ds	3g	4chan
4g	acer	acta	activision
adobe	amazon	android	anonymous
aol	apple	app	augmented
arcade	architecture	arpanet	asus
auto	automobile	battlefield	bing
biometrics	bitcoin	bittorrent	blackberry
blizzard	blogging	blog	bluray
broadband	browser	casual	chatroulette
chrome	chromebook	cispa	computing
console	cookies	craigslist	crowdfunding
crowdsourcing	cryptocurrency	cybercrime	cyberwar
darknet	data	dell	diablo
doodle	dotcom	drone	dropbox
e3	ebay	email	emoji
encryption	energy	engine	engineering
ereader	events	facebook	fat
filesharing	firefox	flickr	foursquare
gadget	game	gameplay	gamergate
games	gaming	ghz	gmail
google	googlemail	gps	groupon
gta	hacking	halo	handheld
hardware	hashtag	hd	heartbleed
htc	html5	i	ibm
icloud	ie	imac	indie

instagram	intel	internet	ios
ipad	iphone	ipod	isp
itunes	keyboard	kickstarter	kindle
kinect	laptop	lenovo	lg
limewire	link	linkedin	linux
live	machinima	macintosh	macworld
malware	mario	megaupload	microsoft
minecraft	mmorpg	mobile	monitor
motoring	mouse	mozilla	myspace
nes	net	netbook	nfs
nintendo	nokia	oracle	ouya
p2p	paypal	pc	phablet
phishing	photography	photoshop	pi
pinterest	piracy	pirate	platform
playback	playstation	pokemon	power
processor	programming	ps	ps2
ps3	ps4	psp	python
raider	ram	raspberry	rayman
recommendation	reddit	retro	robot
rpg	rts	safari	samsung
search	security	seo	skype
smartphone	smartphones	smartwatch	smartwatches
software	sonic	sony	sopa
spam	spotify	steam	stream
starcraft	stuxnet	sun	surface
symbian	tablet	technology	technophile
ted	telecom	television	tetris
titanfall	tomb	trojan	tumblr
twitch	twitter	viber	vine
virus	warcraft	web	whatsapp
wheel	wifi	wii	wikipedia
windows	windows7	wireless	worms
wow	xbox	xp	y2k
yahoo	youtube	zelda	zynga