Zalando Textmining

4. Meilenstein - Abschlusspräsentation

Benjamin Fischer
Sebastian Krawczyk
Sebastian Kasanzew
Raul Vinh Khoa Nguyen

konkrete Problemstellung: Identifizieren neuer themenrelevanter Wörter

Welche Wörter sind fashion relevant?

Verschiedene Blog-Einträge sollen auf “Fashion-Relevanz” untersucht werden:

"The adidas YEEZY is the first *Kanye West adidas* sneaker."

"*Adidas*-Schuh von *Kanye West*: Verkauf für Yeezy Boost 350 startet am..."

Welche Wörter sind fashion relevant?

Fashion-relevante Begriffe sollen markiert werden:

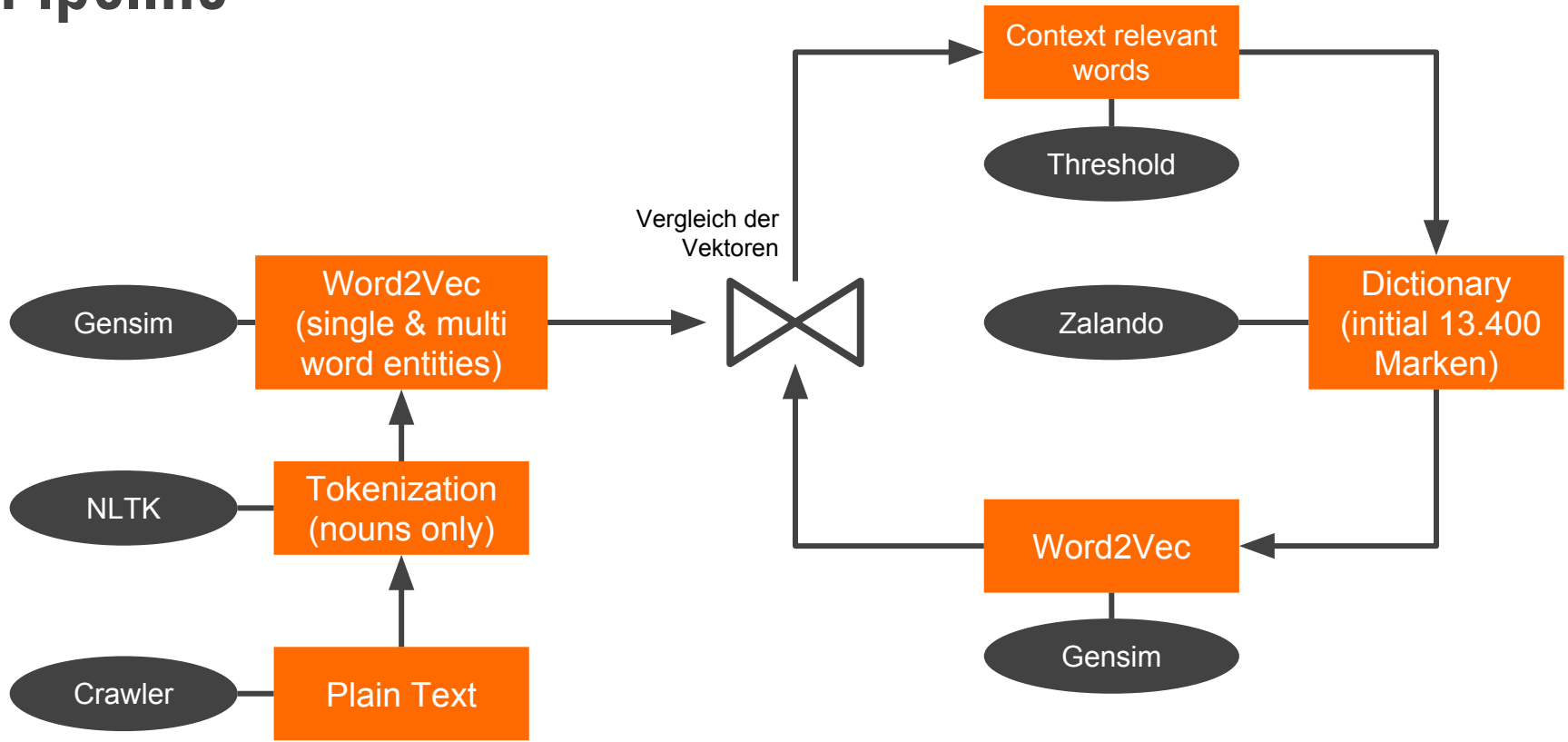
"The **adidas YEEZY** is the first ***Kanye West adidas* sneaker**."

"***Adidas*-Schuh** von ***Kanye West***: Verkauf für **Yeezy Boost 350** startet am..."

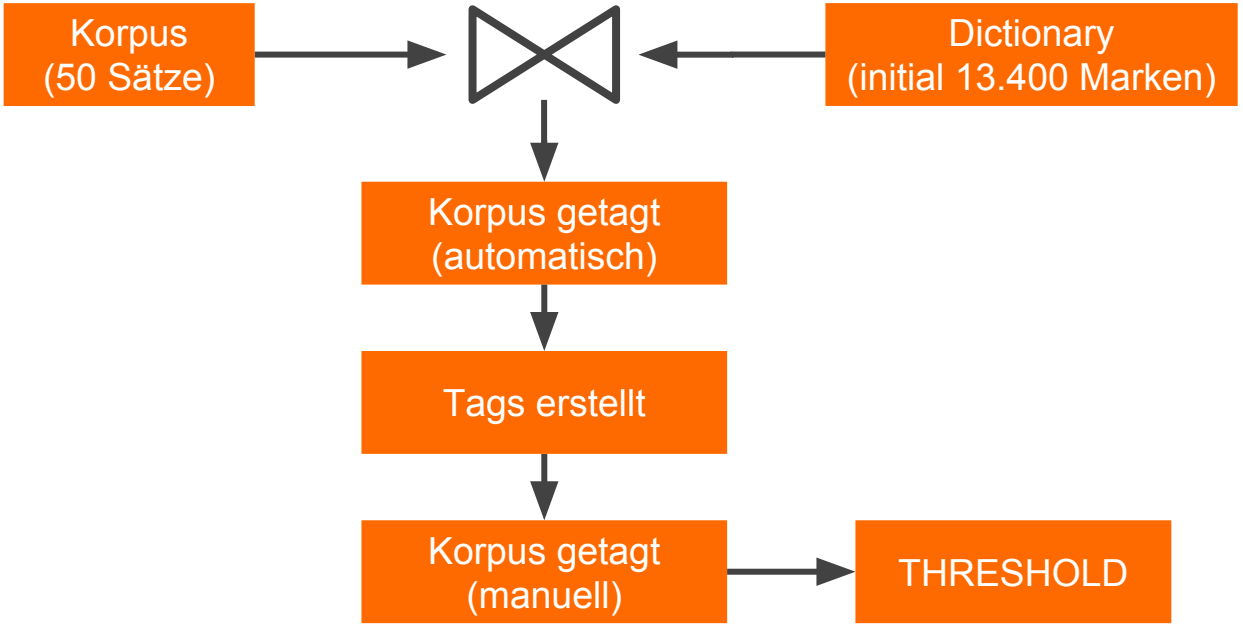
zalando

Pipeline

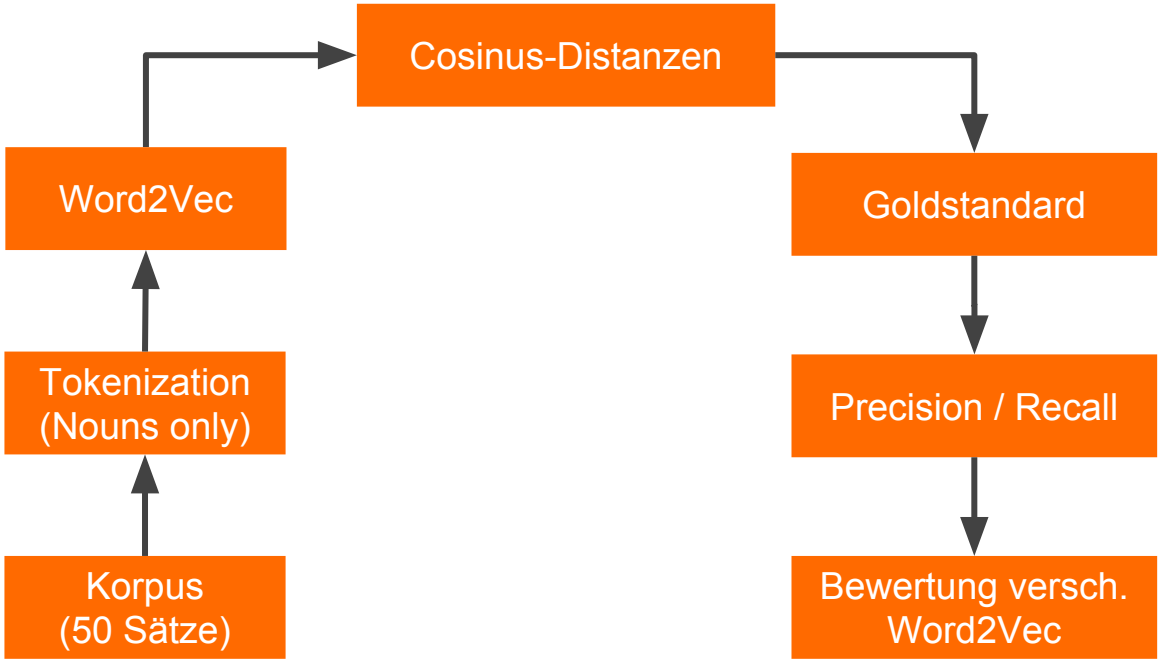
Pipeline


Context relevant words
Threshold
Vergleich der Vektoren
Word2Vec (single & multi word entities)
Gensim
Zalando
Dictionary (initial 13.400 Marken)
NLTK
Tokenization (nouns only)
Word2Vec
Crawler
Plain Text
Gensim

Geeigneten Threshold über Goldstandard finden

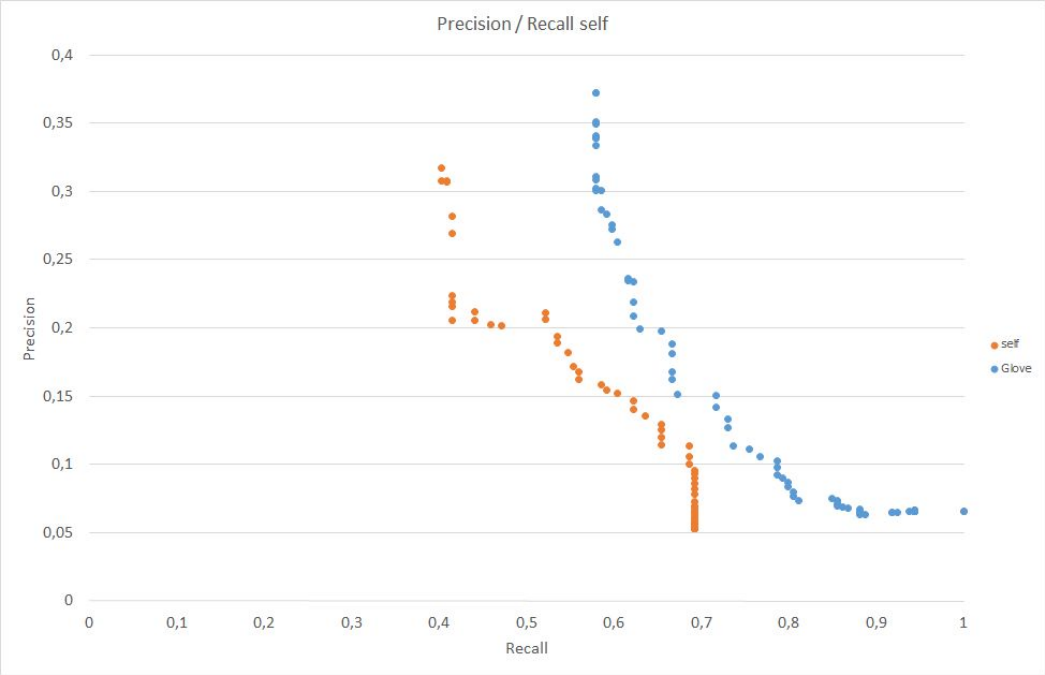


Bewertung der Word2Vec (self-trained & Glove)



Blick auf Ergebnisse

Vergleich von selftrained Word2Vec und Glove



Vergleich von selftrained Word2Vec und Glove

Visualisierung Ergebnisse: Goldstandard

Document ID: 559be2c065f2886d5687b9cd

Aleali May H&M Life presents their special segment called "Everyday Icon". The segment includes interviews as well as photos of women from different cities with amazing style! Honored to be one of these women, Adam Katz of Le21eme, shot these photos of me in my hometown, Los Angeles. Look 1: Y-3 Flannel, Maison Martin Margiela Leather Leggings, Camper Sneakers, Gucci Mask Sunglasses Look 2: Bathing Ape Button Down, Ksubi Denim, Vans Vault Leather Sk8 Hi & Givenchy Bag Look 3: Army Surplus Zip Up, Acne Studios Denim, Fendi Sunglasses, Patrick Mohr Sneakers, Hermes CDC Bracelet Look 4: Homme Boy Graphic Tee, Acne Jeans, Jordan 1 OG "Royal"

*orange markierte Wörter befinden sich bereits im Dictionary, grüne noch nicht

Visualisierung Ergebnisse: self-trained Word2Vec

Threshold: 0,75

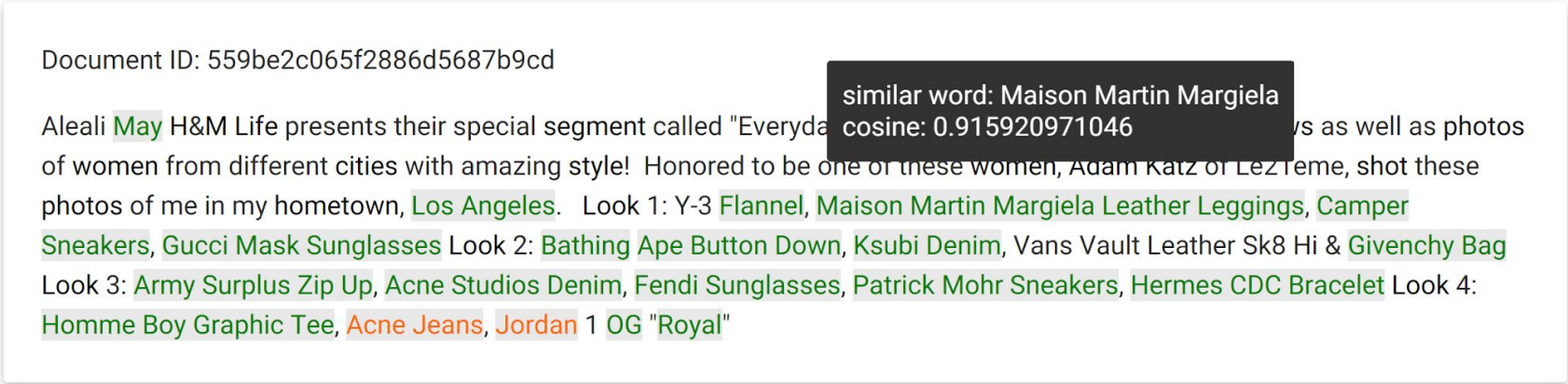
Document ID: 559be2c065f2886d5687b9cd

Aleali May H&M Life presents their special segment called "Everyday Icon". The segment includes interviews as well as photos of women from different cities with amazing style! Honored to be one of these women, Adam Katz of Le21eme, shot these photos of me in my hometown, Los Angeles. Look 1: Y-3 Flannel, Maison Martin Margiela Leather Leggings, Camper Sneakers, Gucci Mask Sunglasses Look 2: Bathing Ape Button Down, Ksubi Denim, Vans Vault Leather Sk8 Hi & Givenchy Bag Look 3: Army Surplus Zip Up, Acne Studios Denim, Fendi Sunglasses, Patrick Mohr Sneakers, Hermes CDC Bracelet Look 4: Homme Boy Graphic Tee, Acne Jeans, Jordan 1 OG "Royal"

*orange markierte Wörter befinden sich bereits im Dictionary, grüne noch nicht

Visualisierung Ergebnisse: self-trained Word2Vec

Threshold: 0,75



Beim Hovern über markierte Wörter, werden Cosinusdistanz und das entsprechende Wort aus dem Dictionary angezeigt

*orange markierte Wörter befinden sich bereits im Dictionary, grüne noch nicht

Abschlussbetrachtung

Erreichte Ziele

- Identifizieren themenrelevanter Wörter mittels verschiedener Word2Vec
- Ermittlung von Glove als besser geeignetes Word2Vec (im Vergleich zum self-trained)
- Multithreading
- Visualisierung der Ergebnisse mittels generierter HTML-Seiten
- automatisiertes Erweitern des Dictionary
- Integration der Ergebnisse des Crawler-Teams

Ausblick für die nächsten 6 Monate

- Anlernen besser geeigneter Word2Vec
- Optimieren des Threshold
- Erweitern des Dictionaries auf mehr als nur Marken
- Direkter Vergleich mit SPIED (Stanford Pattern-based Information Extraction and Diagnostics)
- Patternbased learning Algorithmen einfügen wie in SPIED http://nlp.stanford.edu/software/patternslearning.shtml
- Daten des Crawler-Teams in SPIED integrieren

Entwicklungsaufwand

- 4 Personen
- durchschnittlich jeweils 7h / Woche
- 16 Wochen
- entspricht 4 x 7h x 16W = 448PS = 56PT

* PS - Personenstunden
* PT - Personentage

zalando

Vielen Dank für die Aufmerksamkeit