Mario Fischer und Tobias Aubele

DIE SEO CAMPIXX 2017

Wie jedes Jahr trafen sich im März im Hotel am Müggelsee in Berlin Suchmaschinenoptimierer und -optimiererinnen zum zwanglosen und vor allem offenen Know.how-Austausch. Bei den über hundert Vorträgen war für jede Vorbildung etwas dabei. Einsteiger konnten ihr Wissen verbreitern und Experten fanden wie immer ebenfalls diverse Anregungen zum Einstieg in noch tiefere Datengefilde oder neue Tools.
Website Boosting hat sich für Sie auf der SEO Campixx umgesehen und Ihnen einige Highlights, Learnings und Tipps mitgebracht.

SEO = Cash & Run

So der Titel des Vortrags von Thomas Mindnich. Er empfahl, als SEO mehr zu testen. Ohne Hypothesen zu bilden, bringen Experimente bzw. Studien nichts, weil sie dann weder verifiziert noch falsifiziert werden können.

Mindnich weiß, wovon er spricht, denn er führt selbst viele eigene datengetriebene Tests von Textinhalten durch. Seiner Meinung nach beginnt SEO mit dem Content. Bringen die meisten vermutlich nur für Suchmaschinen verfassten Texte wirklich Vorteile im Ranking (Abbildung 1)? Statt dem üblichen „der Mitbewerber hat das so, also mach ich das auch“, plädiert er für Split-Tests. Nicht selten wird bei Tests das Ergebnis dann abgelehnt oder uminterpretiert, weil nicht das Erwartete herauskam.

Aktuell explodiert die Nachfrage nach sog. holistischen Inhalten. Wer sich noch an die „Was ist was?“-Bücher aus der Kindheit erinnern kann – da war alles zu einem Thema drin. Mit einem Schmunzeln wies Mindnich darauf hin, dass es dieses holistische Prinzip also schon sehr viel früher gab – offline und als Print. Für Webseiten wird leider oft übersehen, dass gerade für Long-Tail-Suchen (solche mit mehreren Worten) bzw. Treffer die Entfernung der Terme einer der Hauptrankingfaktoren ist, die sog. Long-Tail-Proximity.

Zudem muss man sauber auswählen, welches Tool man auf welche Quellen loslässt. Es kommt ja immer auch auf die Intention eines Textes an. Als Beispiel zeigt Mindnich einen Text der Spiegeltochter bento.de (Abbildung 3). Dieser ist wahrscheinlich eher auf Klickbaits (Köder) in sozialen Medien ausgerichtet und nicht auf Holistizität. Liegen solche Texte dann im Sammelkorb automatisierter Tools, bekommen man z. T. völlig falsche Signale. Er vertrat die Meinung, dass bei transaktionalen, E-commerce-orientierten Suchanfragen SEO-Textbausteine fast immer fehl am Platz sind.

Tipps zur Google Search Console

Auch Sebastian Erlhöfer betonte in seinem Vortrag die Vorteile einer datengetriebenen Vorgehensweise für SEO. Als einen möglichen Datenlieferanten stellte er die Google Search Console (vormals „Webmaster Console“) näher vor. Gleich zu Anfang räumte er mit einer häufigen Fehlinterpretation einiger Zahlen auf. Zeigt die Search Console unter Crawling an, dass pro Tag durchschnittlich z. B. 3.500 Seiten geholt werden (Abbildung 4) und man hat ca. 9.000 Seiten auf der Site, bedeutet das mitnichten, dass die Site etwas alle drei Tage einmal durchgecrawlt wird. Diese Zahlenangabe sagt zwar aus, dass 3.500 Seiten abgerufen wurden – aber eben nicht 3.500 verschiedene. Google crawlt einzelne Dokumente nach einem eigenen Rhythmus und holt vielleicht die Startseite oder „News“ mehrmals pro Tag. So kann es passieren

DER AUTOR

Mario Fischer ist Professor für E-Commerce an der Hochschule Würzburg-Schweinfurt und Herausgeber der Website Boosting. Er beschäftigt sich u. a. seit 1996 intensiv mit dem Thema Suchmaschinen.

DER AUTOR

Tobias Aubele ist Professor für E-Commerce, insbesondere Conversion-Optimierung und Usability an der Hochschule Würzburg-Schweinfurt. Darüber hinaus berät er KMU im Bereich Webanalytics & Website-Optimierung.

INFO KÖLN

Die Geschichte der berühmten Stadt am Rhein, welche zu den ältesten Deutschlands zählt, begann bereits vor rund 2000 Jahren mit den Römern, von denen die Stadt auch ihren Namen erhielt.

Köln ist nicht nur für sein Kölnisch Wasser „4711", den rheinischen Frohsinn und den Kölner Zoo bekannt, sondern vor allem für seinen Dom. Hier ist neben zahlreichen Kunstobjekten auch der Schrein der Heiligen Drei Könige zu besichtigen. Wer den Kölner Dom besucht und die 509 Stufen erklommen hat, darf den schönsten Ausblick über das Stadtgebiet genießen. Am Abend ist der Blick auf das beleuchtete Rheinufer ein weiteres Highlight.

Die lebensfrohe Metropole zählt zu den größten deutschen Universitätsstädten und belegt in verschiedenen Fachbereichen immer wieder Top-

Abb.1: Ob dieser Text wirklich für Menschen geschrieben wurde, die in Köln einen Anwalt suchen? *(Quelle: www.anwalt.de)*

Abb. 2: Den folgenden Text haben wir für Google geschrieben ...

Abb. 3: Solche Texte sind für linguistisch-maschinelle Analysen eher ungeeignet *(Quelle: bento.de)*

bzw. ist durchaus normal, dass einige Seiten wochen- oder gar monatelang gar nicht überprüft werden.

Wer mit SEO auf Unterseiten anfängt und sich wundert, warum er keine Änderungen im Ranking beobachten kann, sollte daher auch immer prüfen, ob die modifizierten Seiten auch wirklich schon abgeholt und damit neu indexiert und bewertet wurden. Als Werkzeug dazu empfahl Erlhöfer, z. B. den Screaming Frog LogfileAnalyzer in Kombination mit einer (aktuellen) XML-Sitemap einzusetzen. Damit lässt sich relativ einfach gegenüberstellen, welche Seiten und welche eben nicht in einem bestimmten Zeitraum gecrawlt wurden.

Auch beim Entfernen von Seiten aus dem Index via Search Console muss man Vorsicht walten lassen. Der Ausschluss beträgt nur „ungefähr 90 Tage", wie Google dazu angibt. Danach können die Seiten also jederzeit wieder auftauchen. Wer Seiten wegen rechtlicher Probleme auf diese Art entfernen möchte, sollte sie tatsächlich auf dem Webserver löschen oder das Metatag „noindex" im Head der Seite verwenden. Wer es zu gut meint und die Seite dann auch aus der robots.txt-Datei ausschließt, schießt sich damit erst recht ins Knie. Er verbietet den Robots, die Seite aufzurufen – wodurch diese gar keine Kenntnis von dem „noindex"-Tag bekommen können. Und der Ausschluss in der Search Console wirkt wie erwähnt nur 90 Tage. Zunächst sieht also alles gut aus und nach Ablauf der Frist erscheint die Seite plötzlich wieder in den Suchergebnissen. Dies kann einen gegnerischen Anwalt bei Urheberrechtsverletzungen durchaus verzücken.

> „Interpretationen töten den Test"
> – Thomas Mindnich

Ebenso häufig werden die Daten der Impressions falsch verstanden. Es kommt jeweils darauf an, was man sich anzeigen lässt. URL-basierte Auswertungen zählen das Erscheinen für jede URL, die im Zeitraum in Suchergebnissen mit ausgegeben wurde. Bei domainorientierten Auswertungen wird das Erscheinen nur einmal pro Suchergebnisseite gezählt – unabhängig

> „Sei schlau, nutze die API“ – Sebastian Erlhöfer

Abb. 4: Die Daten der Google Search Console können missverstanden werden

davon, wie viele einzelne URLs einer Domain dort vorhanden waren. Abbildung 5 zeigt den Unterschied. Für die Domain *www.campixx-week.de* wurden acht unterschiedliche URLs ausgespielt. Auf die Domain bezogen wird aber nur eine Impression ausgewiesen.

Die Weboberfläche der Search Console liefert nach Erlhöfer meist nur 1.000 Datensätze. Wer mehr Daten braucht oder haben möchte, sollte daher auf die API (Datenschnittstelle) zurückgreifen. Außerdem bekommt man über diese Schnittstelle tatsächlich auch mehr inhaltliche Daten, wie z. B. das Datum des ersten und des letzten Crawls für eine URL. Google hat umfangreiche Hilfstools ins Netz gestellt. Die unterschiedlichen Testtools für die APIs der Search Console finden Sie unter *http://einfach.st/gscapis*. Dort kann man interaktiv Abfragen zusammensetzen und auch ausführen (Abbildung 7). Da die Anzahl der Datensätze pro Zeitraum für die automatisierte Abfrage begrenzt ist, sollte man bei größeren Sites ggf. mehrere Properties anlegen, z. B. eines für jedes Verzeichnis. Somit lassen sich nicht nur mehr Daten holen, sondern dies geht auch durch die Option der gleichzeitigen Abfrage schneller.

Erlhöfer zeigte auch eine gute Möglichkeit, die API anzuzapfen, ohne programmieren zu müssen oder zu können. Hierzu kann man die in der Website Boosting schon mehrfach beschriebenen SEO-Tools für Excel von Niels Bosma sehr gut einsetzen. Dort ist eine Schnittstelle verbaut, die über ein Formular die entsprechenden Abfragen

Abb. 5: Bei der Interpretation der Impressions in der Search Console ist Vorsicht angebracht

Abb. 6: Sebastian Erlhöfer zeigte den richtigen Umgang mit der Google Search Console

zusammenstellt und die Daten in die Zellen von Excel fließen lässt.

Google stellt unter *http://einfach.st/anasheets* auch ein kostenloses Browser-Plug-in „Search Analytics for Sheets“ zur Verfügung, mit dem man Daten online direkt in Google Sheets holen kann (über Add-ons in der Menüzeile).

Auch einige Anwendungsfälle blieb Erlhöfer nicht schuldig.

» **Attraktive Schwellenkeywords finden**
Daten nach Position 11-15 filtern und vielversprechende Rankings in die Top 10 optimieren.

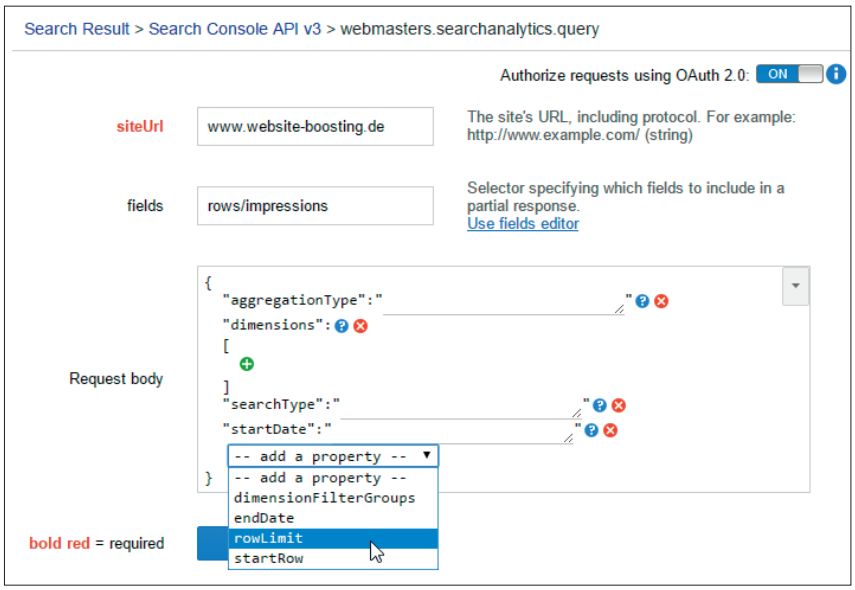

Abb. 7: Mit etwas Programmierwissen lassen sich Daten aus der Search Console interaktiv abfragen

Abb. 8: In einigen Vorträgen wurden statt Folien echte Live-Beispiele gezeigt – so wie hier nützliche Excel-Makros für SEO von Thomas Grübel

» **Mehrfachrankings finden**
Wenn eine Domain für eine Suchabfrage mehrfach an verschiedenen Positionen auftaucht, sollte man die die letzte(n) URL(s) so modifizieren (oder deindexieren), dass sie für diese Suche nicht mehr auftauchen. Das schiebt den ersten Treffer in der Regel spürbar nach oben. Ausnahme: Platz eins und zwei.

» **Keyword Mappings prüfen**
Eine Abfrage der angezeigten URLs für ein Suchwort gibt die Möglichkeit zu prüfen, ob auch wirklich die richtigen URLs ranken. Bsp: Für einen Shop rankt für ein generisches Keyword eine Produkt- statt der Kategorieseite.

» **Suchergebnisse mit schlechter Klickrate finden**
Die von der Search Console übermittelten Klickraten können mit Bezug auf die durchschnittliche Position Hinweise geben, wo die Klickraten deutlich niedriger sind, als es zu erwarten wäre. Hier prüft man in der Regel Title und Description, ob sie wirklich zur Suchintention passen und klickaffin getextet wurden.

Lügen-Tracker Analytics – enttarnt und verbessert

Google Analytics zählt zu den meistverbreiteten Webanalysesystemen, ist jedoch in vielen Fällen unzureichend integriert und liefert demnach falsche Ergebnisse. So misst Google Analytics im Standard die Sitzungsdauer als Differenz zwischen zwei Pageviews. Bei Blogs wird teilweise nur eine Seite intensiv betrachtet, was zu einer Absprungrate von 100 % und einer Sitzungsdauer von 0 Sekunden führt. Weiterhin ist die Sitzungszeit aufgrund des „Tab-Surfens“ zu hoch (Seite nicht immer im aktiven Tab) bzw. führt fehlendes Cross-Domain-Tracking bei einem externen Check-out zu entsprechend falschen Absprungraten. Andi Petzoldt beleuchtete weiterhin die Restriktionen von Analytics und gab konkrete Verbesserungshinweise zum Tracking (inkl. eines eigenen Tracking-Codes). Konkrete Empfehlungen sind bspw. die Integration eines Absprungraten-Faktors, welcher eines eigenen Sekundenzählers bedarf (zählt, wenn der Tab den Fokus hat). Analytics sollte anschließend ein Event senden, wenn der Nutzer die Seite eine entsprechende Zeit im aktiven Tab hatte (bspw. 15 Sekunden), eine entsprechende Scrolltiefe erfüllte (bspw. doppelter Viewport) oder eine relevante Aktion (bspw. Videoaufruf) durchführte. Weiterhin ist ein Besuchszeiten-Faktor nützlich, welcher die Textlänge der Seite betrachtet (Abbildung 9) und dadurch einen Hinweis auf Verbesserungspotenzial liefert.

Das Problem fehlender Sessions in Analytics kann vielfältiger Natur sein. Neben fehlendem Analytics-Code (Überprüfung bspw. mit Screaming Frog) sowie zu spätem Laden des Codes aufgrund ungünstiger Positionierung im

Code sind es teilweise gegebene Limits von Analytics:

» 10 Millionen Hits/Monat pro Property
» 200.000 Hits/User pro Tag
» 500 Hits pro Session
» Jedes Tracker-Objekt startet mit 20 Hits, welche mit 2 Hits pro Sekunde „nachgefüllt" werden
» 4 Millionen URLs

Die Limits der kostenlosen Version von Analytics erscheinen auf den ersten Blick sehr hoch, durch die Vielzahl an empfehlenswerten Events können diese jedoch schnell erreicht werden (Tracking des Scrollverhaltens, Tab-Timer, PDF-Downloads, Klicks auf externe Links, Tracking von geklickten E-Mail-Adressen, Formular-Abbruch inkl. Messung des letzten aktiven Feldes, 404-Fehler, Ad-Block-Nutzung, Kopieren von Text auf Seite etc.). Sofern Session-IDs Teil der URL werden und diese nicht ausgeschlossen werden, kann auch die Grenze von vier Mio. URLs schnell erreicht werden. Daher sollte die eigene Analytics-Integration kritisch überprüft und modifiziert werden, denn: „Richtig messen ist wichtig!" Tipp: Auf www.tracking-garten.de entsteht ein modifizierter Tracking-Code, welcher erweiterte Trackingmöglichkeiten bietet sowie diverse Unzulänglichkeiten wie bspw. eingebauten Spamschutz sowie Opt-out-Rate optimiert.

SEO für informationsorientierte Suchanfragen

Wenn Jonas Weber über SEO spricht, hören viele Teilnehmer sehr aufmerksam zu. Weber war schließlich lange Zeit Mitglied des Spamfighter-Teams bei Google in Dublin. Warum lohnt SEO? 2015 lag der Kanalanteil für Online-Shops laut einer Studie von sem-deutschland.de bei über 85 % mit weiterhin steigender Tendenz. Umgekehrt sinkt der Anteil an AdWords leicht von 7,47 % auf knapp über 5 % für

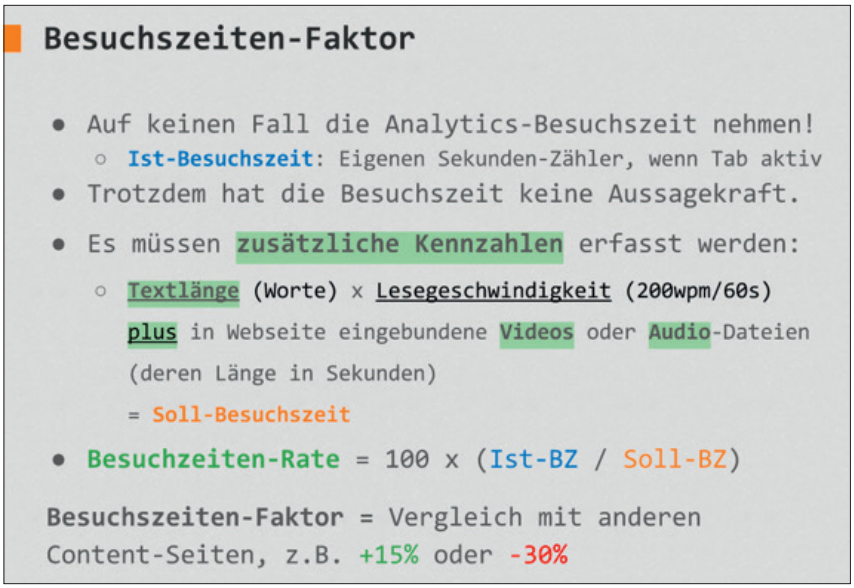

Abb. 9: Definition Besuchszeiten-Faktor *(Quelle: Andi Petzoldt)*

2016. In der Regel hat kein Kanal auf Dauer einen so hohen Return on Investment (ROI) wie organischer Traffic über Suchmaschinen, sprich Google.

Natürlich kann man einen hohen Aufwand betreiben, um für umkämpfte Suchbegriffe im organischen Bereich ganz oben zu ranken. Es geht aber auch anders. Weber führte das an einem Beispiel an. Die Suchabfrage „es juckt am ganzen Körper" bringt keine AdWords-Werbung über den Suchergebnissen. „Medikament gegen Jucken" allerdings schon. Dort werden oben Google-Shopping-Ergebnisse und mehrere AdWords über die unbezahlten Ergebnisse gestellt. Informative Suchanfragen sind oft recht nachhaltig. „Ständig juckende Haut/Hände/Beine etc." oder „immer wieder jucken im Ohr/im Intimbereich/in der Nase etc." zeigen als Suchanfrage schon in sich, dass dafür rankende Seiten dauerhaft eine hohe Besuchsfrequenz haben können.

Ein weiteres Beispiel: „Joggingschuhe" bringt vier Anzeigen, „Laufstrecke" dagegen keine. Der Klickpreis ist bei den Joggingschuhen fünfmal so hoch. Marketer geben oft viel Geld für bezahlte Klicks aus, um eine Vermarktungskette anzustoßen. Sie kaufen ihren Traffic nicht selten teuer ein und haben oft nennenswerte Streuverluste. Mittels intelligenten Einsatzes von SEO geht das hingegen fast kostenlos bzw. mit einer einmaligen Investition, die sich sehr schnell rechnet, wie Weber zeigte.

„Brustvergrößerung München" kostet bei AdWords pro Klick ca. 8,80 €. Die Ergebnisseite ist voll mit Werbung. *„Kleiner Busen"* hat fast das doppelte Suchvolumen, kostet aber nur 15 ct. pro Klick. Im ersten Fall kosten 1.000 Besucher 8.800.- €, im zweiten die gleiche Anzahl nur 150.- €. Aber auch über Facebook lässt sich oft kostengünstig per Push Traffic für Contentseiten abholen. Noch immer sind Agenturen und Werbetreibende wohl wenig kreativ und gehen mehr nach Gefühl als datengetrieben vor.

Im Kern ging es Weber darum zu zeigen, wie man die Vermarktungskette via Retargeting-Pixel (bei AdWords: Remarketing) deutlich günstiger starten kann. Dies gelingt z. B., indem man für informationsorientierte Suchintentionen suchmaschinenoptimierte Seiten baut und auf diesen dann die Pixel ausspielt und damit die Besucher markiert. Später können diese beim Surfen auf anderen Sites gezielt per Werbung angesprochen werden. Idealerweise fällt man dabei nicht gleich mit der Tür ins Haus, sondern bietet vielleicht zunächst weitere Informationen an wie z. B. ein kostenloses Webinar. Kann man

Abb. 10: Jonas Weber erklärte, wie man mittels SEO Retargeting-Pixel günstig ausspielen kan

den Besucher dort überzeugen, lässt er sich leicht(er) in einen echten Kunden umwandeln.

Was ist nun besser? Streut man interessenbasierte Retargeting-Pixel bezahlt oder besser über SEO? Weber hatte zum Test 40 suchmaschinenoptimierte und nutzerfreundliche Beiträge geschrieben. Die Seiten sind seit sechs Monaten live und bringen mittlerweile über 100.000 Besucher im Monat direkt aus der gewünschten Zielgruppe. Es funktioniert also – wenn man es richtig anpackt.

Ein Vektor sagt mehr als 1.000 Worte

Im letzten Zeitslot am Sonntag hielt Stefan Fischerländer seinen Vortrag über den Einsatz von Machine Learning. Wer die Heimreise schon vorzeitig angetreten hatte, darf und sollte sich intensiv darüber ärgern, einen der sicherlich besten Vorträge des letzten und dieses Jahres verpasst zu haben. Fischerländer machte das Publikum erst warm, indem er erklärte, wie Clustering funktioniert und wie man die sog. euklidische Distanz dazu einsetzen kann. Klassifizierungen und Mustererkennung zählen tatsächlich zu den wichtigen Bausteinen einer Suchmaschine.

Im Prinzip kann man Worte mittels Vektoren im n-dimensionalen Raum beschreiben. Fischerländer erläuterte das anhand des folgenden Beispiels:

king - man + woman = queen

Abbildung 13 zeigt schematisch, wie man mithilfe von Vektoren vom nicht männlichen König (king - man) zur weiblichen Bezeichnung (+ woman) „Königin“ gelangt. Vom Vektor king (1) wird zunächst der Vektor man (2) abgezogen und anschließend der Vektor women (3) addiert. Dieser neue Vektor landet im Raum in der Nähe des Vektors „queen“, also Königin. Daraus kann eine Maschine in einem Modell dann automatisch ableiten, dass die Bezeichnung Königin höchstwahrscheinlich etwas mit dem weiblichen Pendant eines Königs zu tun haben muss.

Je näher die beiden Vektoren am Ende zusammenliegen, desto stärker ist die begrifflich inhaltliche Übereinstimmung.

Soweit die Theorie. Um das in der Praxis auch einmal auszuprobieren, lud sich Fischerländer zunächst das Textkorpus der gesamten deutschen Wikipedia-Site mit ca. 5 GB Volumen herunter und erzeugte mit dem von Google als Open Source veröffentlichten Framework „word2vec“ *(http://einfach.st/word2vec)* auf einem normalen Linux-PC in ca. fünf Stunden ein Modell mit 200 Dimensionen. Der Entwickler von word2vec ist übrigens auch der Hauptautor des von Google veröffentlichten Papers über RankBrain.

Er lud dieses Modell in sein Notebook und zeigte unter Zuhilfenahme der dafür entwickelten Python-Bibliothek Gensim *(http://einfach.st/gensim)* live, was man mit solchen Vektoren machen kann.

Nach Eingabe von Positiv- oder Negativvektoren für Begriffe warf das Modell faszinierend gute Ergebnisse aus. Abbildung 14 zeigt die bereits erklärte Abfrage mit den realen Ausgabewerten des Modells. Auf Abbildung 15 und Abbildung 16 sind weitere Beispiele zu sehen. Die Ergebnisse der Abfrage „model.most_similar(positive=[´BVB´], negative=[´Dortmund´] lassen wir aus Gründen sportlicher Diskretion hier aus. Die Teilnehmer hatten natürlich sofort Feuer gefangen und so machte Fischerländer viele weitere aufschlussreiche Abfragen. Natürlich kam irgendwann die Bitte, „Merkel“ mit „Mann“ zu verknüpfen. Und siehe da, im benachbarten Vektorenraum konnte man u. a. „Kohl“ und „Frank-Walter“ lesen.

Man muss sich nochmals ins Gedächtnis rufen: Die Texte stammen nur aus der deutschen Wikipedia und wurden nur auf einem einfachen Homecomputer modelliert. Und bereits dies ermöglicht es zu prüfen, welche Begriffe einen inneren Zusammenhang

$$d(x,y) = \|x-y\|_2 = \sqrt{(x_1-y_1)^2+\cdots+(x_n-y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

Abb. 11: Die euklidische Distanz

> „Machine Learning ist keine künstliche Intelligenz, es sind statistische Verfahren“;
> – Stefan Fischerländer

haben, und mit Worten zu „rechnen“.

Fischerländer hatte aber noch mehr im Gepäck. Er ließ 150.000 Tweets von 60 deutschen SEO-Accounts mit ca. 14 MB und für einen „nur“ 20-dimensionalen Raum auf einem MacBook durch das word2vec-Modell berechnen. Die Dauer? 30 Sekunden. Anschließend zeigte er, wie das Modell die in den Tweets enthaltenen Berichte korrekt klassifizieren kann. Auf einzelne Abfragen („Campixx“ -> seokomm, seocampixx, omx, semseo u. a/„Mediadonis“ [Markus Tandler, Anm. d. Red] -> cemper, seonaut, pascalfantou, boeserseo, nerdinskirt, tobiasfox u. a.) hin zeigte sich erneut, dass bereits mit einem kleinen Textkorpus solche trainierten Modelle bemerkenswert gut Ähnlichkeiten finden können.

Abbildung 16 zeigt etwas anderes, nämlich einen weiteren Erkenntniswert für SEO. Über das Modell lässt sich prüfen, welche Worte nicht zusammenpassen. Das Beispiel ist sehr einfach und so würde es natürlich noch nicht dazu taugen, die Qualität eines Textes wirklich valide zu prüfen. Allerdings arbeitet Google auch nicht nur mit einer vergleichsweise so kleinen Datenbasis, sondern hat über das Web über 80 Trillionen Dokumente als Korpus zur Verfügung. Nicht zu vergessen die vielen Datenbanken und Entitäten, die mittlerweile aufgebaut wurden (siehe Website Boosting #36, Google RankBrain). Und natürlich verwendet man dort nicht nur einen PC dafür. Experten vermuten, dass mittlerweile weltweit jeder zweite Webserver bei Google steht.

Abb. 12: Hielt einen der bemerkenswertesten Vorträge: Stefan Fischerländer

Aber ist das jetzt schon Intelligenz?

Was „intelligent“ ist, darüber lässt sich sicher trefflich streiten. Wer die Vergangenheit aufmerksam beobachtet hat, erkennt, dass wir Menschen auch dazu neigen, die Schwelle, ab wann wir etwas so nennen, beständig zu verschieben. Meist immer dann, wenn uns ein Computer in einer Domäne geschlagen hat, die wir für uns reklamiert hatten (Backgammon, Dame, Schach, Jeopardy, Go und zuletzt Poker). Fischerländer wählte dazu einen einfachen Wald- und-Wiesen-„Intelligenz“-Test auf Brigitte.de aus und ließ die zehn Fragen durch das Modell laufen. Und dieses war tatsächlich in der Lage, sechs der zehn Fragen korrekt zu beantworten. Die Wahrscheinlichkeit, dass dies zufällig passierte, lag rechnerisch bei nur zwei Prozent. Klar, das ist alles andere als ein wissenschaftlicher Rahmen. Erstaunlich aber allemal.

Im entsprechenden Forschungspaper „Efficient Estimation of Word Representations in Vector Space“ von vier Google Engineers (PDF unter http://einfach.st/w2vpaper) steht dazu bei den Schlussfolgerungen:

„Our ongoing work shows that the word vectors can be successfully applied to automatic extension of facts in Knowledge Bases, and also for verification of correctness of existing facts. Results from machine translation experiments also look very promising.“

Das war Stand 2013! Und wurde damals wie immer in der Szene nicht ganz so ernst genommen, wie es nötig gewesen wäre. Der Sinn einer solchen Faktenerkennung liegt wie oben angedeutet klar auf der Hand. Die Qualität von Sätzen, Aussagen und Texten lässt sich maschinell sehr viel besser einschätzen. Fischerländer stellte rhetorisch die Frage, wie seriös wohl der folgende Satz auf einer Webseite auf einen Menschen wirkt.

„Die besten Online-Shops für Mode sind Zalando, Otto, MeinTollerModeShop24 und Bonprix.“

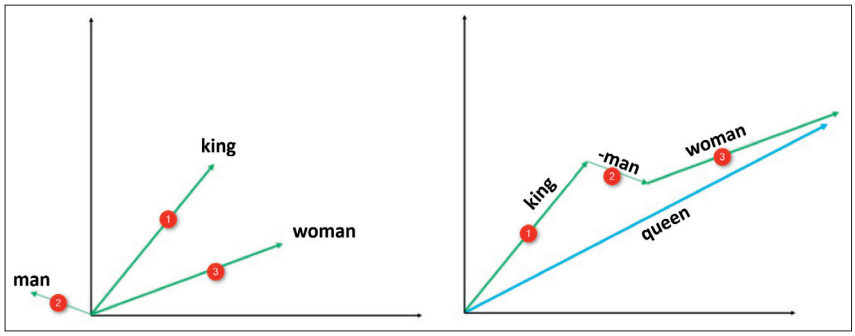

Abb. 13: Per Vektordistanz wird aus dem König, der nicht Mann, sondern Frau ist, eine Königin *(Quelle: Stefan Fischerländer)*

Natürlich würde uns auffallen, dass hier etwas nicht stimmt. Zumindest das kann das einfache Vektorenmodell von Fischerländer aber auch. Und bereits sein kleines Vektorenmodell wäre in der Lage, eine solche Aussage als unseriös für ein Ranking abzulehnen oder zurückzustufen. Wer kann sich ausmalen, um wie viel weiter das Modell mittlerweile von Google entwickelt wurde, wie groß das Dokumentkorpus ist und die Geschwindigkeit, das alles zu verarbeiten?

Einsatzmöglichkeiten für word2vec

Über Klassifizierungen lassen sich (auch neue) Begriffe bereits bestehenden zuordnen. Damit ließen sich ggf. automatisch Texte generieren, wie Fischerländer zeigte. Nach der Zuordnung eines Automodells X über das Vektorenmodell zu 75 % für „Klein- und Kompaktwagen“ und zu 25 % zu „Sportwagen“ könnten Sätze wie „X ist ein Klein- und Kompaktwagen mit sportlicher Note“ generiert werden.

Aber auch für die Optimierung der internen Verlinkung kann man wertvolle Erkenntnisse ableiten. Ebenso ließe sich die interne Suchfunktion damit verbessern.

In der Szene ist bereits länger bekannt, dass sich die Rankingsignale immer mehr in Richtung Content verschieben. Stefan Fischerländer zeigte eindrucksvoll in der Praxis, wie so etwas funktionieren kann, und öffnete somit sicher dem einen oder der anderen die Augen, dass die Vision einer intelligenten Maschine, die Webseiten aufgrund inhaltlicher Qualität anstatt von Titles, Überschriften oder Backlinks beurteilt, gar nicht mehr so visionär ist, sondern zunehmend Realität. Wer selbst so etwas nachvollziehen oder nachbauen möchte, dem empfiehlt Fischerländer, Python zu lernen.

```
>>> model.most_similar(positive=['Koenig', 'Frau'], negative=['Mann'])

[(u'Gemahlin', 0.7290278673171997), (u'Koenigs', 0.7272396087646484),
(u'Regentin', 0.711432158946991), (u'Koenigin', 0.6957824230194092),
(u'Titularkoenigs', 0.6942150592803955), (u'Thronerbin',
0.6831889748573303), (u'Thronbesteigung', 0.653880774974823),
(u'Eduards', 0.6514211893081665), (u'Regentschaft', 0.6499212980270386),
(u'Maetresse', 0.6497464776039124)]
```

Abb. 14: Was ist König minus Mann plus Frau? *(Quelle: Stefan Fischerländer)*

```
>>> model.most_similar(positive=['Deutschland'])

[(u'Schweiz', 0.7215635776519775), (u'Oesterreich', 0.6851878762245178),
(u'Grossbritannien', 0.6840066313743591), (u'Bundesrepublik',
0.6752889752388), (u'Niederlanden', 0.6633405089378357), (u'USA',
0.645158588886261), (u'Benelux-Laendern', 0.6436079144477844),
(u'Westdeutschland', 0.6359853148460388), (u'Benelux-Staaten',
0.6338233947753906), (u'Belgien', 0.6330624222755432)]
```

Abb. 15: Was ist ähnlich dem Wort „Deutschland“? *(Quelle: Stefan Fischerländer)*

```
>>> model.most_similar(positive=['Canberra', 'Kanada'],
negative=['Australien'])

[(u'Ottawa', 0.7005172371864319), (u'Ontario', 0.6972213983535767),
(u'Montreal', 0.6887485384941101), (u'Toronto', 0.6846897006034851),
(u'Winnipeg', 0.6674027442932129), (u'Mississauga', 0.6657336950302124),
(u'Saskatoon', 0.6404088735580444), (u'Halifax', 0.6335928440093994),
(u'Alberta', 0.6323922276496887), (u'Saskatchewan', 0.626349687576294)]
```

Abb. 16: Canberra minus Australien plus Kanada? *(Quelle: Stefan Fischerländer)*

```
>>> model.doesnt_match("Fruehstueck Mittagessen Restaurant
Abendessen".split())
'Restaurant'

>>> model.doesnt_match("Rom Florenz Gardasee Venedig".split())
'Gardasee'

>>> model.doesnt_match("Entwickler Developer SEO Programmierer".split())
'SEO'
```

Abb. 17: Auch möglich: Welches Wort passt nicht? *(Quelle: Stefan Fischerländer)*

Fazit

Wie immer – ein Besuch der SEO Campixx lohnt sich eigentlich grundsätzlich für alle, die sich ernsthaft mit SEO beschäftigen. Es ist für jeden und für jedes Vorwissen mehrfach etwas dabei und die Möglichkeit des Networkings sollte man auch nicht unterschätzen. Im nächsten Jahr wird die zentrale Campixx-Week aufgelöst. Sie wird laut dem Veranstalter in einzelnen Großstädten in Deutschland stattfinden. Das wird vielen die Entscheidung für die gezielte Teilnahme erleichtern und auch das Budget für Reisekosten spürbar schonen.

Save the date: Die nächste SEO Campixx wird am 1. und 2. März 2018 sein. Wieder in Berlin, allerdings diesmal familienfreundlich unter der Woche (Do./Fr.). Weitere Infos finden Sie unter *www.campixx.de*.¶