



Statistik und Text Mining — eine statistische Sicht

Antony Unwin

Lehrstuhl für RechnerOrientierte Statistik und Datenanalyse

Universität Augsburg

Statistik und Text Mining

Antony Unwin

2. November, 2011

Aufräumen einiger Irrglauben

- Statistik ist langweilig
 - man kann in jedem Gebiet Statistik einsetzen
- Statistik ist zahlenbehaftet
 - viel wichtiger ist die Deutung der Zahlen, und da muss man kreativ sein
- Statistik ist eigentlich versteckte Mathematik
 - Mathematik ist für die Theorie wichtig, nicht für die Praxis
- Für statistische analysen soll man möglichst viele Daten sammeln
 - für statistische Analysen soll man möglichst gute Daten sammeln

Statistik und Text Mining

Antony Unwin

2. November, 2011

Statistik?



Daten

title	year	length	budget	rating	votes	r1	r2	r3
\$	1971	121	6.4	348	4.5	4.5	4.5	14.5
\$1000 a Touchdown	1939	71	6	20	0	14.5	4.5	
\$21 a Day Once a Month	1941	7	8.2	5	0	0	0	
\$40,000	1996	70	8.2	6	14.5	0	0	0
\$50,000 Climax Show, The	1975	71				3.4	17	24.5
\$pent	2000	91	4.3	45	4.5	4.5	14.5	14.5
\$windle	2002	93	5.3	200	4.5	0	4.5	4.5
'15'	2002	25	6.7	24	4.5	4.5	4.5	14.5
'38	1987	97	6.6	18	4.5	4.5	0	0
'49-'17	1917	61	6	51	4.5	0	4.5	4.5
'68	1988	99	5.4	23	4.5	0	4.5	14.5
'94 du bi dao zhi qing	1994	96				5.9	53	4.5
'?' Motorist, The	1906	10	7	44	4.5	0	0	0
'A' 1965	10	6.7	11	14.5	0	0	0	14.5
'A' gai waak	1983	106	7.1	1259	4.5	4.5	4.5	4.5
'A' gai waak juk jaap	1987	101	7.2	614	4.5	4.5	4.5	4.5
'Breaker' Morant	1980	107	7.9	2718	4.5	4.5	4.5	4.5
'Bullitt': Steve McQueen's Commitment to Reality	1968							
'Croccodile' Dundee II	1988	110	5	7252	4.5	4.5	4.5	4.5
'E' 1981	7	8.6	15	0	0	0	4.5	4.5
'El Chicco' - der Verdacht	1995	90				3.9	10	24.5

Movies Datensatz

Statistik und Text Mining

Statistiken

Muster	Frequenz Kommen	Linkstatistik linkstat	L.I.R.	p
,PDS VVFIN PPER ADV	11	0	12.99	0.0001
,dann VVFIN PPER ADV	11	0	12.99	0.0001
,PPER VAFIN ADV ADV	14	1	10.79	0.0005
\$.ADV.,KOUS PPER	10	0	11.80	0.0003
ADV.,PPER VAFIN ADV	10	0	11.80	0.0003
,ADV VMFIN PPER ADV	10	0	11.80	0.0003
,PPER VVFIN ADV.	10	0	11.80	0.0003
,KOUS PPER PPR ADV	10	0	11.80	0.0003
ADV.,PDS VAFIN ADV	9	0	10.62	0.0005
,ADV VVFIN PPER ADV	19	4	7.64	0.0031
\$.PPER VVFIN ADV nicht	8	0	9.44	0.0010
,KON PDS VAFIN ADV	8	0	9.44	0.0010
ADV VMFIN PPER ADV ADV	8	0	9.44	0.0010
PDS VVFIN PPER ADV ADV	8	0	9.44	0.0010
,ich VVFIN ADV.	8	0	9.44	0.0010
KON PPER VAFIN ADV ADV	8	0	9.44	0.0010
,PPER VVFIN ADV nicht	8	0	9.44	0.0010
,KON PDS VAFIN ADV	8	0	9.44	0.0010
,KON KOUS PPER ADV	13	2	6.80	0.0051
,PDS VAFIN ADV ADV	13	2	6.80	0.0051

Tabelle 3: Typische sprachliche Muster im Teilkorpus *Fiktionalisierte Subtextentwicklungsmuster des GenMov-Vorhandenheitskorpus*

Scharloth u. Bubenhofer

2. November, 2011

z.B. Literary Digest Poll von 1936



- 10 Millionen Auto-, Telefonbesitzer angeschrieben, 2.376 Million Teilnehmer
 - 57% für Alf Landon, 43% für FDR
- Gallup befragte 50.000, sagte FDR gewinnt
- Resultat: **FDR 62% Alf Landon 38%**
- Warum?
 - keine zufällige Stichprobe
 - freiwillige Rückmeldung



Statistik und Text Mining

Antony Unwin

2. November, 2011

Statistik in Ihrem Leben (1) SMS



- Ein Wort mit sechs Buchstaben
 - pqrs S
 - abc Sa
 - ghi Sah
 - tuv Sagt
 - jkl Schul
 - def Schule
- Ein Wort mit sechs Buchstaben
 - pqrs S
 - abc Sa
 - tuv Rat
 - jkl Paul
 - def Säule
 - mno Säulen
 - Pauken

Google 29.10.2011:
Säulen c 1.430m
Pauken c 1.740m

Statistik in Ihrem Leben (2) Google



- Google Page Rank Algorithmus (entscheidet die Reihenfolge der Seiten nach einer Suchanfrage)
- Sprachalternativen
 - „am Mittwoch, den 2. November“ oder „am Mittwoch, dem 2. November“
 - color / colour
 - Semmel / Brötchen
 - ...

Zwei Vorgehensweisen (Tukey)



- Detektiv
 - Informationen aus Daten herauslesen
 - Ideen/Hypothesen generieren, Explorative Datenanalyse
 - beschreiben, untersuchen, entdecken
 - Kontext verstehen und einbringen, subjektiver Ansatz
- Richter
 - überprüfen ob Schlüsse gerechtfertigt sind
 - Hypothesen testen
 - hinterfragen, kritisieren, eingrenzen
 - technische Modelle anwenden, objektiv bleiben

Filme — Beispiel für EDA



- Informationen über 87248 Filme von der IMDb Webseite 2007
- Für jeden Film
 - Titel
 - Jahr
 - Länge (Minuten)
 - Budget
 - durchschnittliche Bewertung der Zuschauer auf einer Skala von 1 bis 10
 - Anzahl Stimmen
 - Zertifikat
 - Typ (sieben binäre Variablen)

Filme — Eigenschaften



- Fehlende Budget und Zertifikat Daten
- Länge
 - Ausreißer
 - Rundungen
 - „Kurzfilme“
- Anzahl pro Jahr
- Bewertungen und Stimmen
 - hohe Bewertungen
 - viele Bewertungen

Filme — Ideen testen?



- Idealerweise braucht man neue Daten, um Ideen zu prüfen.
(Die neuen Daten müssen mit den ursprünglichen direkt vergleichbar sein.)
- Annahmen müssen gemacht werden, um Modelle zu entwerfen, die die beobachtete Variabilität der Daten erzeugen können.
- Auf Basis der Modelle überprüft man, ob die Ergebnisse hätten zufällig entstehen können.
- Wenn die Chancen unter dem Modell gering sind, dass so ein Ergebnis oder ein extremeres Ergebnis gefunden sein könnte, wird das Ergebnis „signifikant“ genannt.

ABER: Vieles kann man nicht testen.

Statistische Konzepte (1)



- Variabilität und Zufall, Verteilungen von Zufallsvariablen
 - Wie lang ist ein Satz?
 - Wie häufig kommt eine Kollokation vor?
 - Wie oft wird eine Wortform verwendet?
- Unabhängigkeit
 - Sind Wörter unabhängig voneinander?
 - Sind Kollokationen unabhängig voneinander?
 - Sind Texte unabhängig voneinander?

Statistische Konzepte (2)



- Stichproben (aus einer Grundgesamtheit)
 - Repräsentativität, Verallgemeinerung
 - Kann man Zeitungsartikel als eine Stichprobe betrachten? Wovon?
 - Kann man Kopora als zufällige Stichproben betrachten?
- Vergleichen
 - Womit sollen Statistiken verglichen werden?
 - Häufigkeiten bei anderen Texten des Autors?
 - Häufigkeiten bei anderen Autoren?

Statistische Konzepte (3)



- Schätzen (von Modellparametern), Konfidenz Intervalle
 - Wie oft wird eine Kollokation verwendet?
 - Wie gut sind diese Schätzer?
- Testen: Signifikanz, p-Werte, Fehler I und II Art, multiples Testen
 - Ist ein Ergebnis sehr unwahrscheinlich oder nicht?
 - Wenn wir alle Wörter testen, müssen 5%, bzw. 1% signifikant sein.
- Assoziation und Kausalität
 - Sind zwei Formen assoziiert oder folgt eine der anderen?

Software für EDA und Statistik



- Mondrian für Interaktive Graphik und EDA
 - intuitiv, flexibel, begrenzt
 - rosuda.org/software/Mondrian/
- R für Statistik im herkömmlichen Sinne
 - komplex, mächtig, umfangreich
 - “This is R. There is no if. Only how.”
 - cran.r-project.org/

R als Sprache (1)



- Alles ist ein Objekt und die Objekte können, nach gewissen Regeln, mit einander kombiniert werden. z.B. `table(table(Lemma))`
- Objekte können verschiedenartig angesprochen werden (wie Figuren in russischen Romanen).
- Variablen
 - `m2007$length`, `m2007[,“length”]`, `m2007[,3]`, `with(m2007, length)`
- Daten
 - `m2007[110,3]`, `m2007[,3]`, `m2007[100:250,1:3]`, `m2007[c(4,5:20),seq(3,11,2)]`
 - `m2007[-(1:3),-c((1:6),9)]`
 - `m2007[m2007$length>240,1]`

R als Sprache (2)



- Wichtige Befehle sind polymorph: *summary*, *plot*
- Es gibt manchmal mehrere Befehle, die dieselbe Aufgabe erledigen
 - z.B. *hist*, *truehist*, *ihist*, *histogram*, ...
- Oft kann man Ziele mit verschiedenen Methoden erreichen
 - komplex, einfach, knapp, langatmig, ...
- Inkonsistenzen: Strukturen, die identisch sein sollten, sind es nicht.
- R kann mit Paketen erweitert werden. Wie mit Umgangssprachen werden die Pakete nicht offiziell überprüft.
- Pakete und „Masking“ von Befehlen. In anderen Umgebungen kann es sein, dass Befehle andere Bedeutungen haben.

Statistik und Text im Web



- Namevoyager: www.babynamewizard.com/voyager#
 - Häufigkeiten von Vornamen in den Vereinigten Staaten seit 1880
 - attraktiv, informativ, nett
 - Namen von Jungs über die Zeit
www.babynamewizard.com/archives/2007/7/where-all-boys-end-up-nowadays
- n-grams: ngrams.googlelabs.com oder www.culturomics.org
 - Zeitreihen für Worthäufigkeiten in Korpora, die „4% aller Bücher“ enthalten
 - interessant, provokant, ausbaufähig
- wordcount.org/main.php

Zusammenfassung



- Statistik ist für wissenschaftliche Untersuchungen unerlässlich.
- Die Rollen
 - Detektiv (Explorative Datenanalyse) und
 - Richter (Hypothesen testen)sind beide für statistische Analysen wichtig.
- Es sind nicht die mathematischen Theorien, die für das Verständnis von Statistik notwendig sind, sondern die grundlegenden Konzepte.