#nerdstuff voor communicatiewetenschappers

Hoe je computer je kan helpen met je inhoudsanalyse

Damian Trilling

d.c.trilling@uva.nl
@damian0604

Afdeling Communicatiewetenschap
Universiteit van Amsterdam

Versie 0.1
Maart 2013

Waarom dit document?

Waarom dit document?

Om...

- ...je te laten zien wat allemaal kan.
- ...je inspiratie te bieden.
- ...je werk te besparen.
- ...je te laten zien waar je verder kan zoeken.

Waarom dit document?

Om...

- ...je te laten zien wat allemaal kan.
- ...je inspiratie te bieden.
- ...je werk te besparen.
- ...je te laten zien waar je verder kan zoeken.

Let op!

- Het is "work in progress".
- Het kan fouten bevatten.
- Het is zeker niet volledig.

Even voor de duidelijkheid

Dit document...

- ...is geen stap-voor-stap-handleiding.
- ...bevat niet per se de beste, mooiste, en misschien niet eens de makkelijkste oplossing.
- ...gaat ervan uit dat je meedenkt en -googelt.

Beschouw het als een blokkendoos – het een en ander kun je gebruiken, sommige delen niet, en misschien moet je nog wat blokjes ergens anders vandaan halen.

Data verzamelen

Data verzamelen

Principes

- Zo min mogelijk handmatige stappen
- Reproduceerbaarheid
- Output moet geschikt zijn voor (automatische/handmatige) codering en analyse
- Automatische opslag van de data

RSS-Feeds als databron

Voordelen

- bijna alle nieuwssites en blogs beschikken over RSS-Feeds
- Je hoeft de artikelen niet handmatig te downloaden
- Je kan geen artikelen missen.

RSS-Feeds als databron

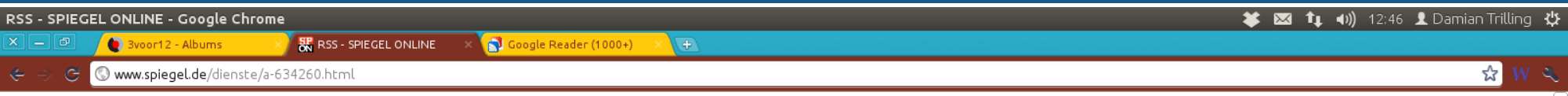
Voordelen

- bijna alle nieuwssites en blogs beschikken over RSS-Feeds
- Je hoeft de artikelen niet handmatig te downloaden
- Je kan geen artikelen missen.

Wat krijg je?

- URL
- Kop
- Teaser (eerste alinea) of hele blogpost
- Datum
- Link naar volledig artikel

RSS-Feeds



Wo immer Sie im Netz das RSS-Zeichen oder den Schriftzug RSS sehen, gelangen Sie zu einem oder mehreren Feeds. Klicken Sie einfach darauf:

Nach einem Klick auf das RSS-Symbol zeigt Ihnen der Browser in der Regel direkt den Feed, damit Sie ihn abonnieren konnen, und eine Hilfe dazu. Wenn nicht, installieren Sie einen neueren Browser - denn alle aktuellen verstehen RSS.

RSS wird unterstützt von Internet Explorer ab Version 7, Firefox ab Version 2, Safari ab Version 2 und Opera ab Version 5.7. Google Chrome unterstützt RSS-Feeds mit einer Erweiterung (hier klicken...).

DIE WICHTIGSTEN RSS-FEEDS IM ÜBERBLICK

alle Topmeldungen von SPIEGEL ONLINE
http://www.spiegel.de/schlagzeilen/tops/index.rss

EILMELDUNGEN - die Breaking News von SPIEGEL ONLINE
http://www.spiegel.de/schlagzeilen/eilmeldungen/index.rss

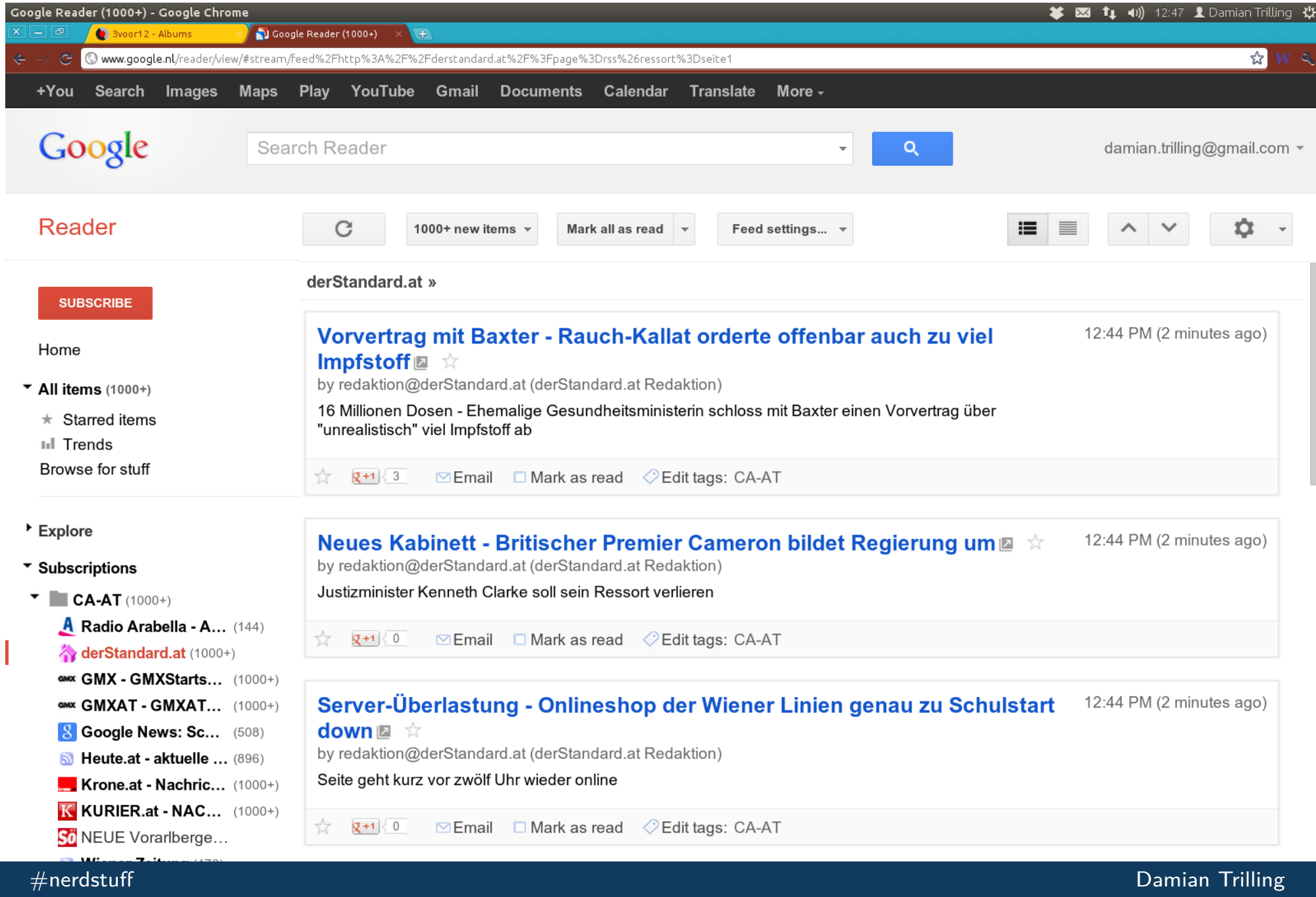
Hoe ziet het eruit?

```
<item>
<title>Umbau beim VfL Wolfsburg: Magaths Spaetlese</
    title>
<link>http://www.spiegel.de/sport/fussball
    /0,1518,824556,00.html#ref=rss</link>
<description>Ploetzlich traeumen sie in Wolfsburg
    wieder von der Europa League. Nach drei Siegen in
    Folge scheint die Mannschaft von Felix Magath
    endlich so etwas wie eine Struktur zu haben. Hat
    sich der Dauerumbau der vergangenen Monate also
    doch gelohnt?</description>
<pubDate>Sat, 31 Mar 2012 14:12:18 +0200</pubDate>
</item>
```

Hoe werkt het?

1. Je abonnert je met GoogleReader op de Feeds die je wilt onderzoeken
2. Je laat GoogleReader de data opslaan en exporteert ze naar een XML-bestand
3. Je zet het XML-bestand om naar een Excel-bestand (of SPSS, STATA, R)
4. Je slaat de artikelen op met wget.

RSS-Feeds


Google Reader (1000+) - Google Chrome
12:47 Damian Trilling
3voor12 - Albums
Google Reader (1000+)
www.google.nl/reader/view/#stream/feed%2Fhttp%3A%2F%2Fderstandard.at%2F%3Fpage%3Drss%26ressort%3Dseite1
+You Search Images Maps Play YouTube Gmail Documents Calendar Translate More
Google
Search Reader
damian.trilling@gmail.com
Reader
1000+ new items
Mark all as read
Feed settings...
SUBSCRIBE
Home
All items (1000+)
Starred items
Trends
Browse for stuff
Explore
Subscriptions
CA-AT (1000+)
Radio Arabella - A... (144)
derStandard.at (1000+)
GMX - GMXStarts... (1000+)
GMXAT - GMXAT... (1000+)
Google News: Sc... (508)
Heute.at - aktuelle ... (896)
Krone.at - Nachric... (1000+)
KURIER.at - NAC... (1000+)
NEUE Vorarlberge...
derStandard.at »
Vorvertrag mit Baxter - Rauch-Kallat orderte offenbar auch zu viel Impfstoff
12:44 PM (2 minutes ago)
by redaktion@derStandard.at (derStandard.at Redaktion)
16 Millionen Dosen - Ehemalige Gesundheitsministerin schloss mit Baxter einen Vorvertrag über "unrealistisch" viel Impfstoff ab
+1 3 Email Mark as read Edit tags: CA-AT
Neues Kabinett - Britischer Premier Cameron bildet Regierung um
12:44 PM (2 minutes ago)
by redaktion@derStandard.at (derStandard.at Redaktion)
Justizminister Kenneth Clarke soll sein Ressort verlieren
+1 0 Email Mark as read Edit tags: CA-AT
Server-Überlastung - Onlineshop der Wiener Linien genau zu Schulstart down
12:44 PM (2 minutes ago)
by redaktion@derStandard.at (derStandard.at Redaktion)
Seite geht kurz vor zwölf Uhr wieder online
+1 0 Email Mark as read Edit tags: CA-AT

Van GoogleReader naar een XML-bestand

URL

http://www.google.com/reader/atom/feed/http://rss.orf.at/news.xml?n=1000

. . . URL van de RSS-feed?n=Aantal artikelen

Van GoogleReader naar een XML-bestand

URL

http://www.google.com/reader/atom/feed/http://rss.orf.at/news.xml?n=1000

. . . URL van de RSS-feed?n=Aantal artikelen

- Een *.xml-bestand per feed
- Denk eraan URLs met speciale letters eerst te coderen (http://meyerweb.com/eric/tools/dencoder/)

Van GoogleReader naar einen XML-bestand

URL

http://www.google.com/reader/atom/feed/http://rss.orf.at/news.xml?n=1000

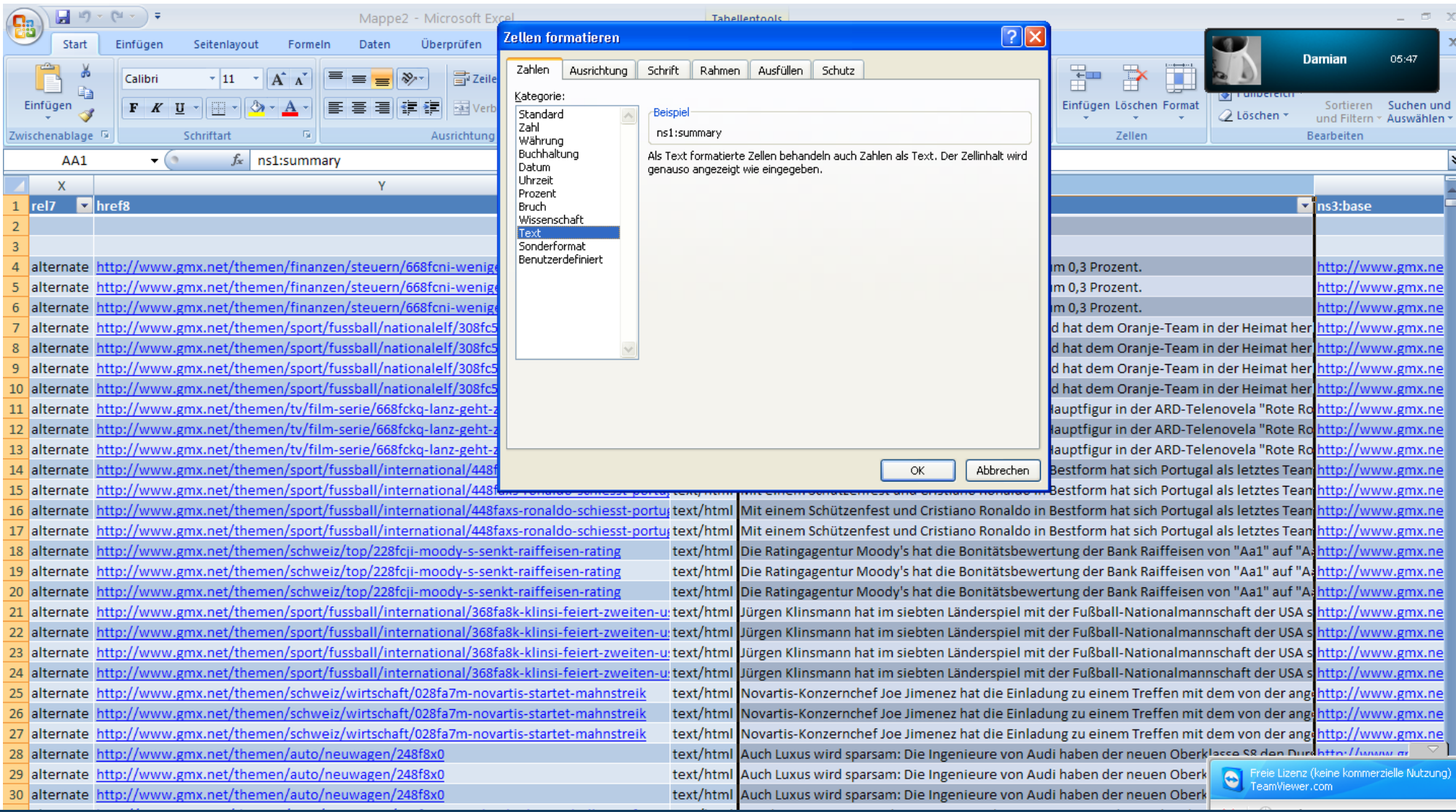
. . . URL van de RSS-feed?n=Aantal artikelen

- Een *.xml-bestand per feed
- Denk eraan URLs met speciale letters eerst te coderen (http://meyerweb.com/eric/tools/dencoder/)
- Voor n>1000 bestanden in meerdere stappen opslaan

  <gr:continuation>CI672pPfi58C</gr:continuation> in erster XML-Datei –> &c=CI672pPfi58C

  toevoegen aan URL om de volgende 1000 artikelen te downloaden

Excel-bestand aanmaken (werkt niet op de Mac!)



Artikelen opslaan met wget

url.txt

http://www.gmx.at/themen/wissen/mensch/108g5xi-baeuerlich-schiefe-zaehne

http://www.gmx.at/themen/unterhaltung/klatsch-tratsch/408g740-fuermann-bittet-um-verzeihung

http://www.gmx.at/themen/nachrichten/aufruhr-arabien/268g70u-regierung-will-zuruecktreten

http://www.gmx.at/themen/nachrichten/panorama/828g54y-neues-zur-klage-gegen-republik

http://www.gmx.at/themen/nachrichten/panorama/968g72s-millionstrafe-wegen-oelpest

http://www.gmx.at/themen/unterhaltung/klatsch-tratsch/368g6yc-kein-babybauch-nur-fast-food

wget-commando

```
wget -i urls.txt
```

(In de volgende versie komt hier een stuk over het verzamelen van Tweets)

(Automatisch) coderen

Woorden en regular expressions (regexp) coderen

De situatie

- Je hebt een dataset (Excel, SPSS, STATA) met Tweets
- Je wilt een variabele aanmaken die altijd 1 is als een bepaald woord genoemd wordt (en anders 0)
- Dat kan natuurlijk ook als je alle Retweets, alle @-mentions of alle Links (http://) wilt coderen.
- Sterker nog: Je kunt ook zeggen dat het woord (of de hele tweet) aan een bepaald patroon moet voldoen: bijvoorbeeld met een cijfer beginnen, gevolgd door een hoofdletter, . . .
  (de zogenaamde regular expressions (regexp) - kijk maar op de STATA-slide)

Hoe doe je dat?

Met SPSS

```
* Create dummy variable for all tweets containing CDA.
COMPUTE cda=INDEX(UPCASE(status),"CDA")>0.
EXECUTE.
* How often is CDA mentioned?.
FREQUENCIES cda.
```

(als de variabele die de tweets bevat “status” heet)
UPCASE zet de hele tweet eerst om naar hoofdletters, vandaar dat ook de zoekterm “CDA” in hoofdletters moet staan.

Hoe doe je dat?

Met SPSS – alle tweets, die “white” bevatten, maar niet “white house”

```
COMPUTE white=INDEX(UPCASE(status),"WHITE")>0.
COMPUTE whitehouse=INDEX(UPCASE(status),"WHITE
HOUSE")>0.
EXECUTE.
IF (white=1) AND (whitehouse=0) whitemaarnietwhitehouse=1.
RECODE whitemaarnietwhitehouse (SYSMIS=0).
```

Hoe doe je dat?

Met SPSS – alle tweets, die “white” bevatten, maar niet “white house”

```
COMPUTE white=INDEX(UPCASE(status),"WHITE")>0.
COMPUTE whitehouse=INDEX(UPCASE(status),"WHITE
HOUSE")>0.
EXECUTE.
IF (white=1) AND (whitehouse=0) whitemaarnietwhitehouse=1.
RECODE whitemaarnietwhitehouse (SYSMIS=0).
```

Zo’n SPSS-analyse van LexisNexis-artikelen?

Lees de handleiding van Rens Vliegenthart op `http://www.polcomm.org/amsterdam-content-analysis-lab/manuals/`

Hoe doe je dat?

Met Excel

```
=IF(ISNUMBER(SEARCH("*cda*",E2)),1,"")
=IF(ISNUMBER(SEARCH("*cda*",E3)),1,"")
=IF(ISNUMBER(SEARCH("*cda*",E4)),1,"") ...
```

(ervan uitgaande dat de tweets in kolom E staan).

Hoe doe je dat?

Met STATA

```
gen romney=1 if regexm(status, "[Rr][Oo][Mm][Nn][Ee][Yy]")
gen obama=1 if regexm(status, "[Oo][Bb][Aa][Mm][Aa]")
recode romney obama (.=0)
```

Google eens naar regexp of kijk hier voor een overzicht:
`http://www.stata.com/support/faqs/data-management/regular-expressions/`

(Ja, je kan dit ook mooier oplossen door ook weer eerst de tweets om te zetten naar hoofdletters: gen status_hoofdletter=upper(status))

Woordenlijsten

De situatie

- Je hebt niet twintig of dertig, maar honderden of duizenden woorden
- Je kan deze woorden ook niet met een regexp-uitdrukking beschrijven – het zijn dus echt honderden verschillende woorden
- Je wilt niet weten welk woord gebruikt wordt, maar óf er een van de woorden gebruikt wordt.
- Bijvoorbeeld: Je hebt een lijst met 1000 positieve en 1000 negatieve woorden en je wilt weten of je tweets negatieve en/of positieve uitdrukkingen bevatten (dit noem je een *sentimentanalyse*)

Hoe doe je dat?

Met Python: Wat heb je nodig?

- een platte-tekst-bestand met de tweets (één tweet per regel)
- een platte-tekst-bestand met positieve woorden (één woord per regel)
- een platte-tekst-bestand met negatieve woorden (één woord per regel)
- het python-script van Neal Caren dat je vervolgens aan jouw behoeftes kan aanpassen (`http://www.unc.edu/~ncaren/haphazard/sentiment.py`)

Een uitgebreide beschrijving met downloadlinks voor het python-script, een positieve en een negatieve woordenlijst staat hier: `http://nealcaren.web.unc.edu/an-introduction-to-text-analysis-with-python-part-1/`

Hoe doe je dat?

Met Python: Wat moet je aanpassen in sentiment.py?

- Vooral de bestandsnamen in regels zoals tweets = open("obama_tweets.txt").read() veranderen
- Er worden nu telkens de woordenlijsten en de obama-dataset gedownload ⇒ deze regels verwijderen of met een # deactiveren
- Misschien wil je ook de woordenlijsten zelf aanpassen

Hoe doe je dat?

Met Python: Wat krijg je?

- Een bestand tweet_sentiment.csv met de originele tweets, gevolgd door een komma, het percentage positieve woorden, nog een komma, het percentage negatieve woorden.
- Dit soort CSV (comma-seperated values)-bestanden kun je met Excel (of SPSS, of STATA, of R) openen!
- Als je niet het percentage wilt weten maar het absolute aantal, vervang dan positive_counts.append(positive_counter/word_count) door positive_counts.append(positive_counter)

Machine learning

Een andere benadering, waarop ik hier niet verder inga, is het zogenoemde "machine learning": Met behulp van een (kleine) training-dataset, die je handmatig codeert, probeer je de computer aan te leren hoe hij moet coderen. Vervolgens laat je de computer de (grote) dataset coderen. In grote onderzoeksprojecten wordt dit al toegepast, en ook wordt er onderzoek naar gedaan op de UvA: `http://ascor.uva.nl/research/phd-research-projects/political-communication--journalism/overview/overview/content/folder/information-retrieval-for-information-services.html`. Het zou dus zomaar kunnen gebeuren dat we over een aantal jaren tools tot onze beschikking hebben die het mogelijk maken, frames te laten coderen zonder afhankelijk te zijn van woordenlijsten!

Tekst exploratief analyseren

n-grams

De situatie

- Je hebt een of meerdere platte-tekst-bestanden
- Je wilt weten welke woorden (of woordcombinaties) het meest voorkomen
- Dit kan bijvoorbeeld handig zijn om onderwerpen te identificeren

n-grams

De zin “Ik hou van inhoudsanalyses” bestaat uit

- vier woorden (logisch)
- drie bigrams: ik_hou, hou_van, van_inhoudsanalyses
- twee trigrams: ik_hou_van, hou_van_inhoudsanalyses

n-grams

De zin “Ik hou van inhoudsanalyses” bestaat uit

- vier woorden (logisch)
- drie bigrams: ik_hou, hou_van, van_inhoudsanalyses
- twee trigrams: ik_hou_van, hou_van_inhoudsanalyses

Waarom ngrams in plaats van losse woorden?

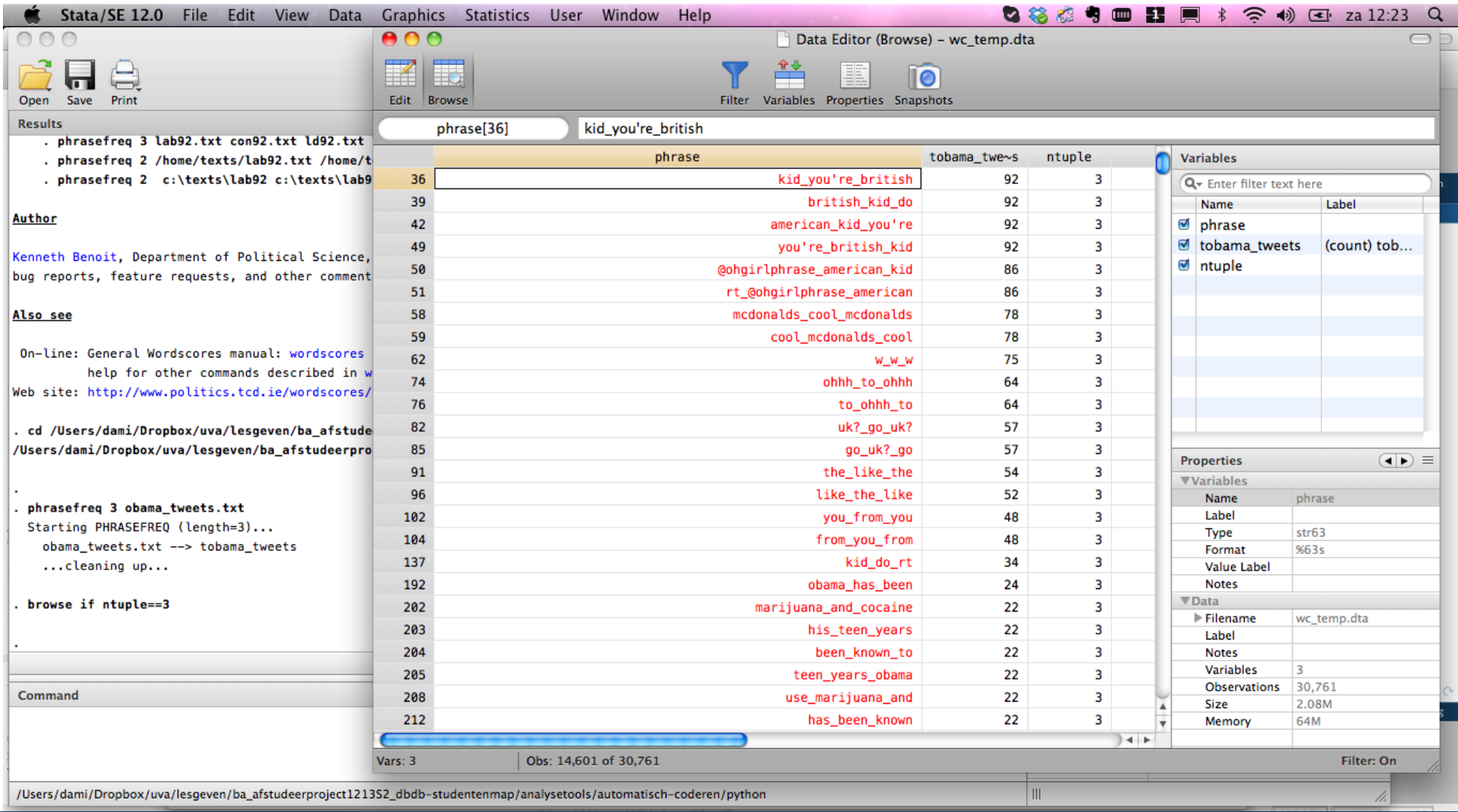
- Om een onderscheid te kunnen maken tussen white_house en "Mr White was in his house" zonder al deze valkuilen handmatig uit te sluiten
- Om vaak voorkomende uitdrukkingen te identificeren
- Omdat de betekenis van woorden nogal kan verschillen als er andere woorden voor of achter staan

Hoe doe je dat?

Met STATA

- Installeer het package wordscore (net install http://www.tcd.ie/Political_Science/wordscores/wordscores)
- voor wordcounts: wordfreq /home/dami/texts/lab92.txt /home/dami/texts/lab97.txt
- voor ngrams (trigrams in dit geval): phrasefreq 3 lab92.txt lab97.txt

Een voorbeeld: trigrams in Obama-Tweets



Hoe doe je dat?

Met Text-NSP, het Ngram Statistics Package (een verzameling perl-scripts)

- Soms heb je meer opties nodig: je wilt vaak voorkomende woorden (stopwords) uitsluiten (de, het, en, of, . . .) of je wilt dat sommige elementen genegeerd worden (html-tags of leestekens bijvoorbeeld)
- perl count.pl - -stop stopfile.txt - -nontoken nontoken.regex - -ngram 3 output.txt input.txt
- Met - -window 5 sta je toe dat maximaal twee woorden die niet bij je trigram horen ertussen staan
- Meer weten? Vraag het aan mij of lees de README: `http://search.cpan.org/dist/Text-NSP/`

Tot slot

Automatisering is mooi. . .

. . . maar vertrouw niet blind op de techniek! Check of je met andere middelen hetzelfde resultaat had gekregen.
En, misschien het belangrijkste: Hou een logboek bij! Wat heb je allemaal gedaan? Hoe ben je er gekomen? Welke keuzes heb je gemaakt – en waarom?

Welk onderwerp zou voor jouw project nuttig kunnen zijn?

Vragen of opmerkingen?

Damian Trilling

d.c.trilling@uva.nl
@damian0604