Taal- en informatietechnologie

Week 4

Informatie-extractie

23 september 2003

Instru teki ennätystuloksen

Sairaalateknologiayritys Instrumentarium teki liikevaihdolla ja -voitolla mitattuna satavuotisen historiansa parhaan tuloksen. Liikevaihto kasvoi kymmenen prosenttia. Valtaosan liikevaihdosta tuova anestesia ja tehohoito kasvoi peräti 14 prosenttia.

Nato-lähde: Turkki-umpikujaan tulossa uusi ehdotus

Naton neuvosto lykkäsi myöhemmäksi kokoustaan, jossa sen tulisi ratkaista umpikujaan ajautunut kiista Turkin avustuspyynnöstä. Nato-lähde vahvisti Helsingin Sanomille tiistaina päivällä, että pääsihteeri **George Robertson** valmistelee uutta esitystä asiassa.

Rond vier uur kwam Lieve Kate halen, even trap geschuurd, en dan naar de volleybal. Was zeer leuk, het ging niet helemaal zo
goed, maar toch gewonnen. Achteraf een beetje blijven plakken, altijd best leuk. En toen ik thuis was, nog een goed anderhalf uur zitten schurer/krabben aan de trap. Tof, stom, vermoeiend werk. Morgen nog doen.

VS woedend over veto in NAVO

De Verenigde Staten hebben woedend gereageerd op de blokkade voor NAVO-hulp aan Turkije die Frankrijk, Duitsland en België gisteren hebben opgeworpen. Volgens de Amerikaanse ambassadeur bij de NAVO, Nick Burns, kampt de alliantie nu met een 'geloofwaardigheidcrisis' vanwege het veto van de drie.

Tekst vs Informatie

- Tekst *is niet* informatie!
- Informatie zit in tekst
- Informatie is taalonafhankelijk
- Informatie uit tekst halen betekent de taal kennen
  - grammatica
  - conventies

Artikel uit 1946

Ettelijke zelfmoordpogingen werden verijdeld

Aanslag op Göring c.s. onthuld

NEURENBERG, 1 Oct. (Daily Tel.). — Dat Dönitz wel wist waar het op zou uitdraaien blijkt uit het feit, dat op de avond van 10 Mei jl., toen hij juist door den Britsen aanklager aan een zwaar kruisverhoor was blootgesteld geweest, door Amerikaanse bewakers in zijn cel met van schoenveters in elkaar gedraaid koord was gevonden. De veters waren langs slinkse weg in zijn bezit gekomen.

Hadden de Amerikanen zijn cel niet zo grondig geïnspecteerd, dan zou Dönitz zich hebben opgehangen.

Deze mededeling is afkomstig van den Amerikaansen kolonel Andruss, den gouverneur van de gevangenis van Neurenberg. De kolonel zei verder, dat de zelfmoorden van Robert Ley, den leider van het Duitse arbeidsfront, en van dr. Conti, den vroegeren Duitsen minister van volksgezondheid, „niet door mensenhanden te voorkomen waren geweest". Het pleit voor de waakzaamheid zijner manschappen, dat niet meer zelfmoorden zijn gepleegd.

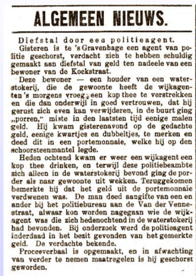
In een van de zomen van het ondergoed van Christine Schröder, Hitler's vroegere privé-secretaresse, een van de belangrijkste getuigen van het proces, werd onmiddellijk nadat zij in de gevangenis was gearriveerd, een capsule met cyanide gevonden, gelijk aan die, met behulp waarvan Himmler indertijd zelfmoord heeft gepleegd.

Geheimzinnig mes

Een onopgelost mysterie was de aanval op Göring, Keitel of Jodl. Wien van de drie hij gold is niet na te gaan. Toen dit drietal indertijd door een der gangen van de gevangenis liep om te worden verhoord, werd plotseling een vlijmscherp, lang jachtmes geworpen, dat zich slechts een paar centimeter van de mannen af in de muur boorde. Het kwam uit de richting van het plafond. Op dat ogenblik werkten 800 S.S.-mannen in het gebouw, die reparaties verrichtten en er waren niet genoeg Amerikaanse bewakers aanwezig om alle gevangenen een afzonderlijk escorte mede te geven.

Hess heeft zich reeds toen hij nog in Britse handen was trachten op te hangen. In Neurenberg werd een tijdlang ernstig gevreesd, dat Ribbentrop zou pogen zelfmoord te plegen. Zijn mede-gevangene Hans Frank bood toen aan zijn cel met Ribbentrop te delen, om een oogje op den ex-minister van buitenlandse zaken te kunnen houden.

1915



ALGEMEEN NIEUWS.

Diefstal door een politieagent.

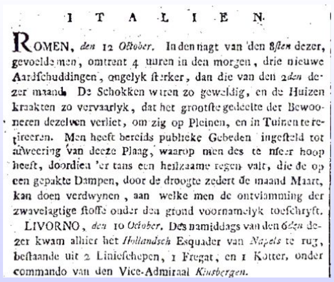
Gisteren is te 's Gravenhage een agent van politie geschorst, verdacht zich te hebben schuldig gemaakt aan diefstal van geld ten nadeele van een bewoner van de Kockstraat.

Deze bewoner — een houder van een waterstokerij, die de gewoonte heeft de wijkagenten 's morgens vroeg een kop thee te verstrekken en die dan onderwijl in goed vertrouwen, dat hij gerust zich even kan verwijderen, in de buurt ging „porren," miste in den laatsten tijd eenige malen geld. Hij kwam gisterenavond op de gedachte geld, eenige kwartjes en dubbeltjes, te merken en deed dit in een portemonnaie, welke hij op den schoorsteenmantel legde.

Heden ochtend kwam er weer een wijkagent een kop thee drinken, en terwijl deze politiebeambte zich alleen in de waterstokerij bevond ging de porder als naar gewoonte uit wekken. Teruggekomen bemerkte hij dat het geld uit de portemonnaie verdwenen was. De man deed aangifte van een en ander bij het politiebureau aan de Van der Vennestraat, alwaar kon worden nagegaan wie de wijkagent was die zich hedenochtend in de waterstokerij had bevonden. Bij onderzoek werd de politieagent inderdaad in het bezit gevonden van het gemerkte geld. De verdachte bekende.

Procesverbaal is opgemaakt, en in afwachting van verder te nemen maatregelen is hij geschorst geworden.

1785



I T A L I E N.

ROMEN, *den* 12 *October.* In den nagt van den 8*sten* dezer, gevoelde men, omtrent 4 uuren in den morgen, drie nieuwe Aardschuddingen, ongelyk sterker, dan die van den 2*den* dezer maand. De Schokken waren zo geweldig, en de Huizen kraakten zo vervaarlyk, dat het grootste gedeelte der Bewooneren dezelven verliet, om zig op Pleinen, en in Tuinen te retireeren. Men heeft bereids publieke Gebeden ingesteld tot afweering van deeze Plaag, waarop men des te meer hoop heeft, doordien 'er tans een heilzaame regen valt, die de opeen gepakte Dampen, door de droogte zedert de maand Maart, kan doen verdwynen, aan welke men de ontvlamming der zwavelagtige stoffe onder den grond voornamelyk toeschryft.

LIVORNO, *den* 10 *October.* Des namiddags van den 6*den* dezer kwam alhier het *Hollandsch* Esquader van *Napels* te rug, bestaande uit 2 Linieschepen, 1 Fregat, en 1 Kotter, onder commando van den Vice-Admiraal *Kinsbergen.*

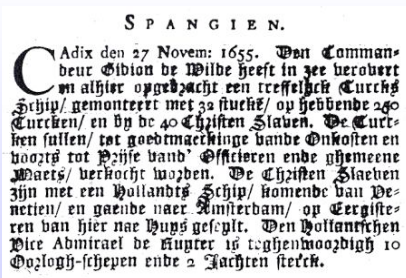
1732

I T A L I E N.

ROmen *den* 19 *July.* Gistere-morgen is de zaek van den Heer Sardini, door de Vergadering *de Nonnullis* afgehandeld; en men verzeekerd, dat hy van zyn Prelaetschap gedegradeerd is, en 10 jaren in 't kasteel van St. Angelo zitten zal.

Genua den 19 *July.* Zaturdag-avond ging een van onze gewapende barken wederom zeyl na Corsica; van waer zondag nog 5 transportscheepen met 200 muylezels en eenige Keyzerlyke Officieren, overquamen. In 't begin van de week arriveerde ook van Bonifacio onze Commissaris-Generael de Graef Gentile, en gistere-morgen de Prins van Wurtenberg, die met alle teekenen van eer by 't aen land stappen, door twee Heeren Gedeputeerden ontfangen, en in 't Carmeliter-klooster geleyd wierd, alwaer een logement voor zyn Hoogheyt gereed gemaekt was.

1655



SPANGIEN.

CAdix den 27 Novem: 1655. Den Commandeur Gidion de Wilde heeft in zee verovert en alhier opgebracht een treffelijck Turcks Schip/ gemonteert met 32 stucken/ op hebbende 250 Turcken/ en by de 40 Christen Slaven. De Turcken sullen/ tot goedtmaeckinge vande Onkosten en voorts tot Prijse vand' Officieren ende ghemeene Maets/ verkocht worden. De Christen Slaeven zijn met een Hollandts Schip/ komende van Venetien/ en gaende naer Amsterdam/ op Eergisteren van hier nae Huys geseylt. Den Hollantschen Vice Admirael de Ruyter is teghenwoordigh 10 Oorlogh-schepen ende 2 Jachten sterck.

Information Extraction

- (Jackson & Moulinier, 3.1)
- Information retrieval = vind documenten relevant voor query
- Information extraction = extraheer specifieke informatie-"snippets" uit vrije tekst
- Gebruikt oppervlakkige NLP

Achterliggend principe: analogie

- Conventions worden uitgedrukt door
  - Functiewoorden en frasen, in conventionele patronen
  - Inhoud op conventionele plaatsen in die patronen
- Analogieprincipe (De Saussure, 1916)
  - A staat tot B als A' tot B'
  - Wanneer A en B een vergelijkbare structuur bevatten, dan bevatten A en B hetzelfde type informatie

Vereenvoudigingen

- Specifieke informatie?
  - Informatie binnen een bepaald domein
    - Terrorisme
    - Financiele transacties
    - Personele wijzigingen
    - Vacatures
- Oppervlakkige NLP?
  - "oppervlakkig ontleden", zoals op middelbare school
  - Herkennen van
    - eigennamen (personen, locaties, organisaties)
    - patronen

Conventies in domeinen

- Patronen in berichtgeving

The Oceanport said its fourth-quarter profit was $1.7 million
Gateway said its fourth-quarter profit was 12 US cents
Kaiser said its first-quarter loss was exacerbated by $7 million
Dan River Inc. said its second-quarter loss was $6.1 million
IBM said its first quarter profit was $1.2 billion
Compaq said its first-quarter profit was 12 cents per share
Cable operator Comcast said its second quarter profit was four cents a share
Chinadotcom said its third-quarter loss was $20.5 million

- [Company] said its [nth] quarter [loss/profit] was [$ (per/a share)]

Named entity recognition

- Wie, wat, waar?
- Tilburg corpus: 120 M woorden uit nieuwsartikelen
- 83% namen correct herkend, 88% persoonsnamen
- Oplossing: patronen

burgemeester van [GEMEENTE]
, zei [PERSOON].
de aandelen van [BEDRIJF]

Named entity recognition (2)

- Beginpunt:
  - 53k namen van het web geplukt (41k pers., 5k loc., 7k orgs.)
  - >3M namen in 120M woorden krantenartikelen
- Cyclus:
  1. Zoek patronen bij niet-ambigue namen
  2. Zoek nieuwe namen in meest zekere patronen
  3. Voeg namen toe, ga terug naar 1

(Buchholz & Van den Bosch, 1999)

Commerciele IE

- Succesverhaal: http://www.flipdog.com
- IE-fabriek:
  - Webcrawling zoekt naar pagina's met vacatures (text classification)
  - Permanent 500,000 in stock
  - Informatie-extractie: type baan, bedrijf, locatie, opleidingsniveau, salaris, …
  - Opslag in en retrieval uit web-toegankelijke database
  - Wekelijks verversen: nieuwe erin, verlopen vacatures eruit

Informatie-extractie uit vacatures

- Wederom conventies:

```
Job type at top, in bold
    "System analyst"
Candidate profile:
    "must possess a [BS] degree in [chemistry]."
Salary offer
    "Salary: [$, $-$, competitive]"
Contact info
    "Contact: [person, email, phone]"
```

- Conventies kunnen automatisch geleerd worden op basis van voldoende voorbeelden

Bonusopdracht 1

- "Precisie van search engines"
- Google, Altavista, Alltheweb
  - http://www.google.com
  - http://www.altavista.com
  - http://www.alltheweb.com

Opdracht (vervolg)

- Google:
  - Grote coverage (meer dan 3 miljard webpagina's)
  - Page ranking, hubs/authorities
  - Standaard AND, maar zie *Advanced search*:
    - AND (all of the words ...)
    - OR (at least one of the words ...)
    - NOT (without the words ...)
    - Exact phrase

Opdracht (vervolg)

- Altavista:
  - Minder grote coverage
  - Meer opties onder *More precision*
  - AND, OR, NOT, Exact phrase

Opdracht (vervolg)

- Alltheweb:
  - Dekt bijna evenveel als Google
  - AND, OR, NOT, Exact phrase
  - Boolean queries
  - "RANK": "pac rank man" levert pagina's met "pac" op, en zo mogelijk ook met "man"

Opdracht (vervolg)

- Doel: door het doen van queries een zo hoog mogelijke *precision* halen op tekstclassificatie.
- Twee classificaties:
  - Nederlandse pagina met CV (wel/niet)
  - Nederlandse pagina met cursusinformatie (wel/niet)

Opdracht (vervolg)

- Voorbeeldqueries:
  - (gehuwd OR ongehuwd) AND opleiding AND werkervaring
  - docent AND (literatuur OR syllabus)
- Verzin zelf minimaal drie queries per taak

Opdracht (vervolg)

- Per query/taak/search engine: kijk naar de eerste 20 links
- Bereken *precision*: van de 20 gevonden pagina's, hoeveel zijn er daadwerkelijk CVs / cursuspagina's?
- Als er meer dan alleen CV / cursusinformatie op staat, dan telt dat als een FOUT.

Opdracht (vervolg)

- Rapporteer dus:
  - Voor Google, Altavista en Alltheweb
  - Precisie met minimaal 3 queries (dezelfde voor de 3 search engines)
  - Bediscussieer verschillen tussen
    - Search engines
    - Queries
  - Let op:
    - Gebuik drie hele verschillende queries

Opdracht (vervolg)

- Rapportage naar keuze in papieren of electronische vorm
- Inleveren bij Antal van den Bosch / antalb@uvt.nl
- Met 1 of 2 personen
- Deadline maandag 13 oktober
- Bespreking op 14 oktober, college Text classification