News mining for conflict forecasting

Thomas Chadefaux — Uppsala

21 September 2017

Sources

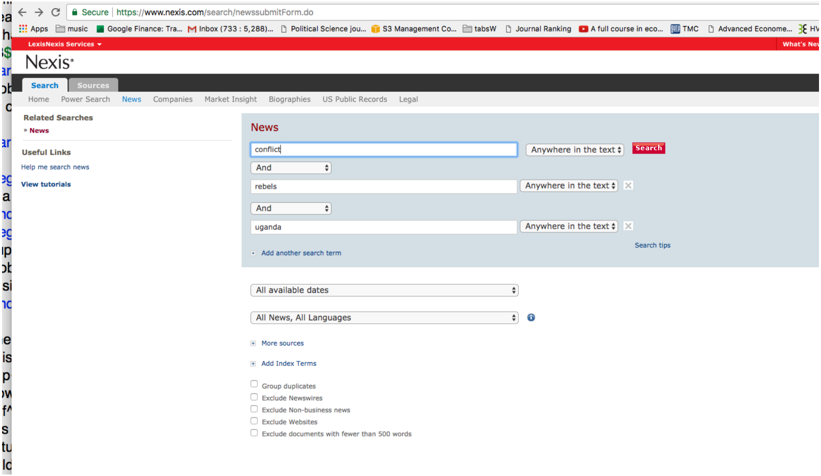
Factiva searches

Can use booleans:

- Simple searches:
  - X and Y
  - X or Y
  - X not Y
  - More complex booleans:
    - X and (Y or Z)
- More complex:
  - “atleast2 deaths” → at least two mentions of “deaths”
  - “x same y” → looks for mentions of x in the same paragraph as y
  - “President w/3 Trump” → “president” within three words of Trump. E.g., President Donald J. Trump
  - “x/F100” → looks for x within first 100 words of the article
  - “fight*" → returns fighting, fighter, fights, etc.
  - Only look for keywords in articles of a certain size -E.g., “rebels and wc > 1000”

Alternatives to Factiva

LexisNexis Newsdesk



Secure | https://www.nexis.com/search/newssubmitForm.do

News

conflict | Anywhere in the text

And

rebels | Anywhere in the text

And

uganda | Anywhere in the text

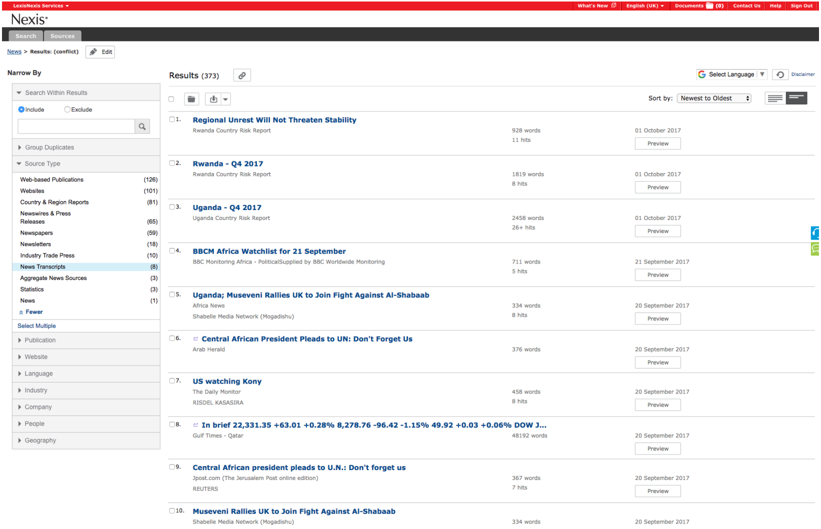
All available dates

All News, All Languages

- □ Group duplicates
- □ Exclude Newswires
- □ Exclude Non-business news
- □ Exclude Websites
- □ Exclude documents with fewer than 500 words

Figure 1:

What the results look like



News > Results: (conflict) Edit

Narrow By

▾ Search Within Results

● Include ○ Exclude

▸ Group Duplicates

▾ Source Type

Web-based Publications	(126)
Websites	(101)
Country & Region Reports	(81)
Newswires & Press Releases	(65)
Newspapers	(59)
Newsletters	(18)
Industry Trade Press	(10)
News Transcripts	(8)
Aggregate News Sources	(3)
Statistics	(3)
News	(1)

Fewer

Select Multiple

▸ Publication

▸ Website

▸ Language

▸ Industry

▸ Company

▸ People

▸ Geography

Results (373)

Select Language ▼ Disclaimer

Sort by: Newest to Oldest

□1. **Regional Unrest Will Not Threaten Stability**
Rwanda Country Risk Report — 928 words, 11 hits — 01 October 2017 — Preview

□2. **Rwanda - Q4 2017**
Rwanda Country Risk Report — 1819 words, 8 hits — 01 October 2017 — Preview

□3. **Uganda - Q4 2017**
Uganda Country Risk Report — 2458 words, 26+ hits — 01 October 2017 — Preview

□4. **BBCM Africa Watchlist for 21 September**
BBC Monitoring Africa - PoliticalSupplied by BBC Worldwide Monitoring — 711 words, 5 hits — 21 September 2017 — Preview

□5. **Uganda; Museveni Rallies UK to Join Fight Against Al-Shabaab**
Africa News
Shabelle Media Network (Mogadishu) — 334 words, 8 hits — 20 September 2017 — Preview

□6. **Central African President Pleads to UN: Don't Forget Us**
Arab Herald — 376 words — 20 September 2017 — Preview

□7. **US watching Kony**
The Daily Monitor
RISDEL KASASIRA — 458 words, 8 hits — 20 September 2017 — Preview

□8. **In brief 22,331.35 +63.01 +0.28% 8,278.76 -96.42 -1.15% 49.92 +0.03 +0.06% DOW J...**
Gulf Times - Qatar — 48192 words — 20 September 2017 — Preview

□9. **Central African president pleads to U.N.: Don't forget us**
Jpost.com (The Jerusalem Post online edition)
REUTERS — 367 words, 7 hits — 20 September 2017 — Preview

□10. **Museveni Rallies UK to Join Fight Against Al-Shabaab**
Shabelle Media Network (Mogadishu) — 334 words — 20 September 2017

Figure 2:

Processing the results

R package for factiva/lexisNexis

```
# Install relevant package (a similar one exists for Facti
#install.packages('tm.plugin.lexisnexis')

library(tm) # text mining package
library(tm.plugin.lexisnexis)

# Import corpus
source <- LexisNexisSource("All_News,_All_Languages2017-09-
corpus <- Corpus(source, readerControl = list(language = NA
```

R package for factiva/lexisNexis

```r
# See how many articles were imported
corpus
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed)
## Content:  documents: 55
```

```r
# Get an overview of the contents of the first article and
print(corpus[1]$content)
```

```
## [[1]]
## <<PlainTextDocument>>
## Metadata:  17
## Content:  chars: 2720
```

R package for factiva/lexisNexis

```r
# See the metadata for the first article
meta(corpus[[1]])
```

```
##   author       : character(0)
##   datetimestamp: 2017-09-20
##   description  : character(0)
##   heading      : US watching Kony
##   id           : TheDaily201709201
##   language     : NA
##   origin       : The Daily Monitor
##   intro        : character(0)
##   section      : character(0)
##   subject      : character(0)
##   coverage     : character(0)
##   company      : character(0)
##   stocksymbol  : character(0)
##   industry     : character(0)
##   type         : character(0)
```

R package for factiva/lexisNexis

```r
# Get the text of the first article
corpus[[1]]$content
```

```
##  [1] "The US is working with African Union Mission in So
##  [2] "STUTTGART- The American troops have withdrawn from
##  [3] " leader Joseph Kony but they are \"keenly watching
##  [4] " are operating. The commander of the US Africa Com
NA
##  [6] "The American government recently announced they we
##  [7] " outfit that has been weakened. Money spent Gen Wa
##  [8] " who had abducted and killed thousands in the regi
##  [9] " in 2006, fled to DR Congo and later to CAR in 200
## [10] " are now weak. \"This past April, it got to a poin
```

R package for factiva/lexisNexis

```r
#remove stopwords using the standard list in tm
dtm <- DocumentTermMatrix(corpus)
inspect(dtm)
```

```
## <<DocumentTermMatrix (documents: 55, terms: 8046)>>
## Non-/sparse entries: 21138/421392
## Sparsity           : 95%
## Maximal term length: 55
## Weighting          : term frequency (tf)
## Sample             :
##                      Terms
## Docs                   and for from have said south that
##   DailyMoni201709109   56  11   12   13    2    17   14
##   DailyMoni201709125   24  10    6   12    0     0   12
##   TheDaily201709108    58  11   12   13    2    17   14
##   TheGuardi2017090219 100  34   24   14    2     0   51
##   TheIndepe201709163   34  12    7   14    7     1    8
##   TheNewYo2017082353   40   8    5    3    6     1   18
```

Collect What?

What to collect?

- Number of articles mentioning any keyword
- Number of hits for each keyword
  - More flexible
  - Multiple TS
- Store all possible articles about country x. Ideal for future uses

Keywords

	Static	Dynamic
Human-defined	Chadefaux 2014	
Automatically inferred		APSR (forth.)

Underlying question: build a data also useful for others, or maximize prediction power?

More advanced processing?

- Tone
- Collocations/n-grams
- TABARI, etc.

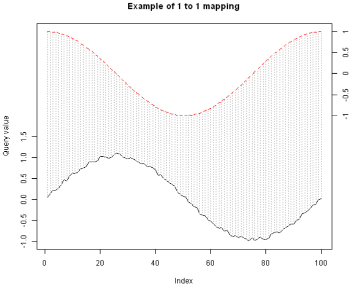
Underlying questions: looking for perceptions, or event coding?

Things to keep in mind

- Article inflation
- Articles correlation

Vague paper idea

- Dynamic time warping for news count. Works well with irregular data.
  - Find /match patterns in TS data


Example of 1 to 1 mapping
Query value
Index