



BlueKiwi

**"What's in the news?" - or: why
Angela Merkel is not significant**

Andrej Rosenheinrich, Dr. Bernd Eickmann
Forschung und Entwicklung, Unister GmbH, Leipzig

GELD.de
Dein persönlicher Finanzoptimierer

kredit.de

Reisende

auto.de
Deutschlands großes Autoportal

U-DEALS

Shopping.de
UND WAS GIBT'S MORGEN? ●●●●●

news.de

börsen**NEWS**.de 

 **Hotelreservierung!.de**
Dein Hotel zum besten Preis

U love Travel
**Urlaubs
TOURS** 

ab-in-den-urlaub!.de
Wir schicken Deutschland in den Urlaub

 **myimmo.de**

fluege.de 

PARTNERSUCHE  **de**
... lieben Sie los! 

EMPFEBLUNGEN

Das Geheimnis der Reisebranche!
Handverlesene Luxushotels. Mitglieder sparen bis zu 70%.

Mein Schiff Taufangebot für Sie
Profitieren Sie von den Taufwochen und sichern sich 100€ Bordguthaben

Punktlandung Traumurlaub!
Schnell noch Sommer buchen! Erlebe Deine Türkei.

IM NETZ
50 Ergebnisse anzeigen

Anzeigen zu: Golf Mexiko

Golf Spielen
www.fernmitgliedschaft.de/golf
2015 ab 79 € incl. 50€ Gutscheine. und Gratis Greenfee-Bonus Karte
Golfmitgliedschaft ab 79€ Greenfee-Karte kostenlos
Kundenstimmen Testsieger
★★★★★ Bewertung für fermmitgliedschaft.de

Golf House Online-Shop
www.golfhouse.de/
Alles zum Golfen bei Golf House mit Bestpreis-Garantie bestellen!
Alles für junge Talente - Golfausrüstung

Golf- & Seehotel Engstler
www.engstler.com/
Mitten in Velden, direkt am See. ideal für Familien + Golfer.
Golfen in Kärnten günstige Pauschalen
Familienurlaub Wellness am Wörthersee

Mexiko
www.tripadvisor.de/Hotelbewertungen
Mexiko Hotels. 4171 Hotels & 680000 Bewertungen.
3.921.997 Personen folgen TripAdvisor auf Google+
★★★★★ Bewertung für tripadvisor.de

Golf von Mexiko - Wikipedia
http://de.wikipedia.org/wiki/Golf_von_Mexiko
Golf von Mexiko (Reliefdarstellung) Gewässer Atlantischer Ozean Landmasse Nordamerika (USA , Mexiko , Kuba) Geographische Lage 25.3686 -90.3905 ...

Hotels in der Region Golf von Mexiko für Ihren Urlaub buchen
http://www.holidaycheck.de/destination-Reiseinformationen_Golf-von+Mexiko-did_1701.1
III Urlaub in der Region Golf von Mexiko mit Hotel vom TESTSIEGER HolidayCheck★★★★★
Günstige Hotels Golf von Mexiko Reise hier buchen...

Ölpest im Golf von Mexiko 2010 - Wikipedia
https://de.wikipedia.org/wiki/9%C3%96lpest_im_Golf_von_Mexiko_2010
Die Ölpest im Golf von Mexiko 2010 wurde durch die Explosion der Ölbohrplattform Deepwater Horizon am 20. April 2010 ausgelöst und ist eine der ...

Alle 50 Ergebnisse anzeigen

BESTE ERGEBNISSE

NACHRICHTEN

Ölplattform: "Deepwater Horizon" tötete massenhaft Delfine
welt.de vor 7 Tagen

Fünf Jahre nach Deepwater Horizon - alles wieder gut?
srf.ch vor 38 Tagen

Ölkatastrophe "Deep Water Horizon": BP muss Milliarden-Entschädigung zahlen
sueddeutsche.de vor 887 Tagen

Deepwater Horizon: Ausgleichszahlungen für Ölkatastrophe
handelsblatt.com vor 7 Tagen

Golf von Mexiko: Tote und Verletzte bei Unglück auf Ölplattform
spiegel.de vor 22 Tagen

Alle 9 Ergebnisse anzeigen

AUF EINEN BLICK

Golf von Mexiko
Der bis zu 5203 m tiefe Golf von Mexiko ist eine Meeresbucht und als Randmeer des Atlantiks sowie als nordwestlicher Teil des Amerikanischen Mittelmeers nahezu von Nordamerika eingeschlossen. Die östliche, nördliche und nordwestliche Küste liegt in den Vereinigten Staaten, die südwestliche und südliche in Mexiko; im Südosten grenzt der weiterlesen...

27°C

10%

CHRONIK

- Golf Mexiko** gerade eben
- ww Golf** gerade eben
- Golf** gerade eben
- champagner** vor 2 Minuten

Bearbeiten | Alles löschen | Deaktivieren

Empfehl uns weiter

Why news-search?

Alexa ranking :

- 11. bild.de
- 13. spiegel.de
- 17. chip.de
- 19. focus.de
- 21. wetter.com




Implementing news-search

- fetching news
- preprocessing
- store news in index
- implemented a search

→ **Whats next?**

- aggregate
- summarize
- find most important news



The screenshot shows a news feed interface with the title "NACHRICHTEN" and a green circular icon with a document symbol. The feed contains five news items, each with a title, source, and time relative to the current time, and a small thumbnail image.

- FIFA - Festnahmen erschüttern die FIFA**
fr-online.de vor 1 Tag
- Fifa-Skandal: Fifa-Präsident beruft Sondersitzung ein**
sueddeutsche.de vor 8 Std.
- Fußball FIFA: FIFA-Ethikkommission sperrt verhaftete Offizielle**
handelsblatt.com vor 1 Tag
- FIFA versinkt im Chaos**
dw.de vor 1 Tag
- Fifa: Immer wieder Blatter**
zeit.de vor 4 Std.

Elasticsearch significant terms aggregation

„The significant_terms aggregation finds uncommonly common terms in your data-set. What do we mean by uncommonly common? These are terms that are statistically unusual.“
(From „Elasticsearch: The Definitive Guide“)

- foreground set of documents vs. background set
- compare term frequency in both sets
- if frequency in foreground set is higher as expected term is significant

→ Angela Merkel is not significant because she is in the news every day.

Significant terms of headlines and entities

Use aggregation on headline terms:

- not really descriptive
- not considering grammar
- names will be split

Use aggregation on entities:

- clean, valid results
- not descriptive
- you will not find „new“ news
- you can not use it to aggregate news

→ We have to move from tokens to facts!

Use NLP for news summaries

- Implemented an nlp based approach to simplify headlines
 - 1) Preprocessing (pos tagging)
 - 2) Chunking (phrase detection)
 - 3) Extract facts using patterns
 - 4) Summary generation

Idea: Try to find the important terms that summarize a sentence based on its structure

Input: headline

- sentence splitting
- cleanup to remove special characters
- tokenizing
- part of speech tagging (Stanford PoS Tagger, STTS Tagset)

Example:

"Angela Merkel trifft am Mittwoch Vladimir Putin in Moskau."

Angela_NE Merkel_NE trifft_VVFIN am_APPRART Mittwoch_NN
Valdimir_NE Putin_NE in_APPR Moskau_NE ._\$.

Input: tagged sentence

- use a regular grammar on the sentence

Nominal phrase: NP → NE+

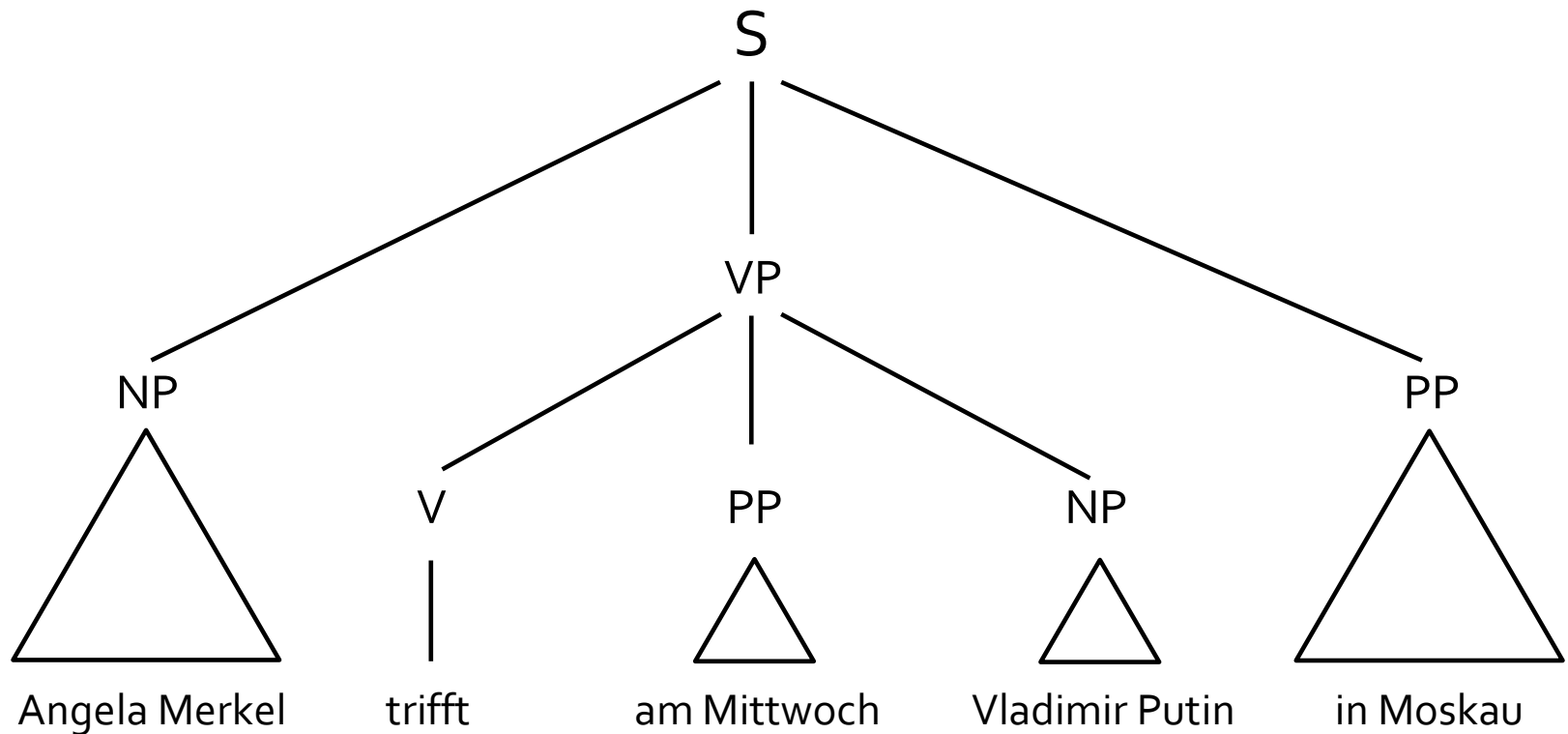
Prepositional phrase: PP → APPART NN

Verbal phrase: VP → V PP NP

Angela_NE Merkel_NE trifft_VVFIN am_APPRART Mittwoch_NN
Valdimir_NE Putin_NE in_APPR Moskau_NE ._\$.

[NP Angela Merkel] [VP trifft [PP am Mittwoch] [NP Vladimir Putin]]
[PP in Moskau]

NLP: chunking step – (partial) syntax tree representation



NLP: extract facts

- Input: partial syntax tree
- Self implemented extraction language
- Goal: extract important keywords using a set of rules

NP<Subj> VP → (VFIN<Pred> NP<Obj>)
Angela Merkel trifft Vladimir Putin

- Apply all rules on the syntax tree → set of matches
- Filter out the best result
- Example: filter by length

Merkel trifft
Merkel trifft Putin
Merkel trifft Putin in Moskau

NLP: summary generation

We want to

- generate normalized term
- generate syntactically well-formed summary

For the normalized term

- Normalize keywords from processing step
- Concatenate them and add to index document

„Merkel treffen Putin“

For the well formed summary

- Remove unwanted parts from syntax tree

NP<Subj> VP → (VFIN<Pred> NP<Obj>)

[NP Angela Merkel] [VP trifft [PP am Mittwoch] [NP Vladimir Putin]]
[PP in Moskau] → Angela Merkel trifft Vladimir Putin

- Preprocess news when indexing
- Use significant terms aggregation on normalized keyword term

"key": "GDL ankündigen Streik",

"key": "Neun Tote zwischen Rockergruppen",

"key": "Kampf gegen Schleuser",

"key": "Schiiten-Milizen zurückerobersollen Sunniten-Stadt Ramadi",

"key": "Lokführer streiken Dienstag",

"key": "Ermittlungen gegen Bundespolizisten",

"key": "Merkel als Gastrednerin"

- Look up well-formed summary for significant terms found

We

- aggregated news
- summarized news
- found significant news



See it live on bluekiwi.de!

Thank you!

