I256:
Applied Natural Language Processing

Marti Hearst
Week 2

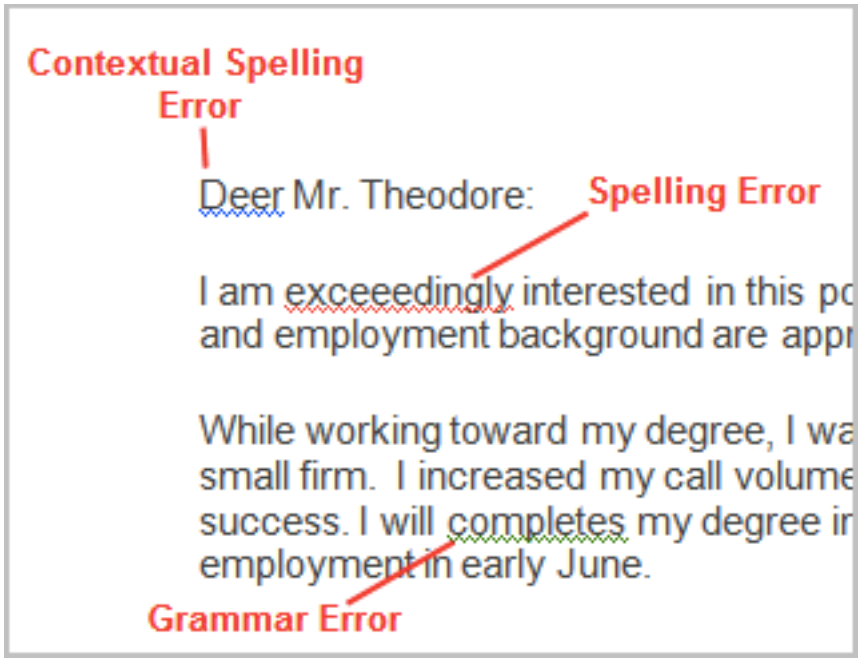
Today

- NLP Applicatinos
- Course Topics and Concept Map
- Segmentation, Syntax, and Semantics
- Coding Practice: NLTK FreqDist
- Wednesday's Assignment

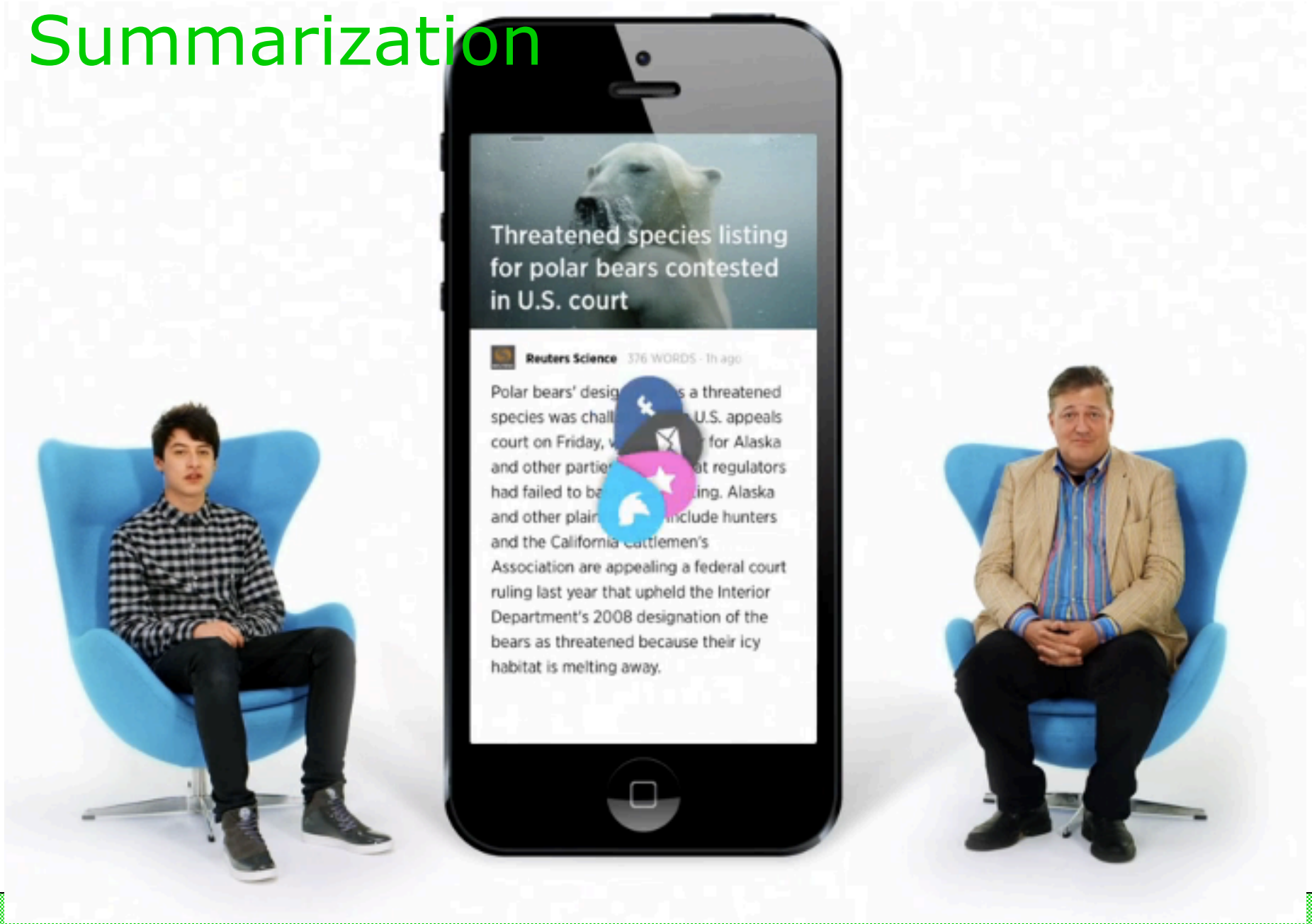
What are some

REAL WORLD APPLICATIONS OF NLP?

Grammar and Spelling Correction

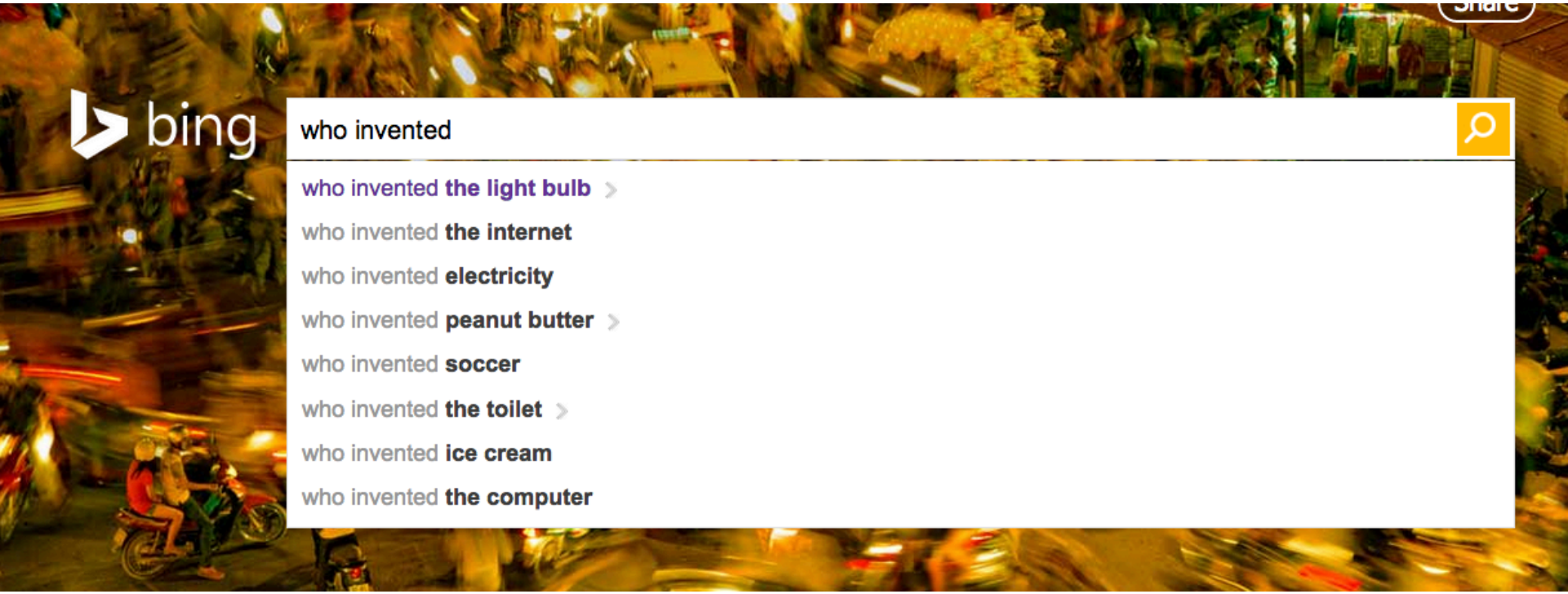


Summarization



http://summly.com/press-videos.html

Is auto-suggest NLP or something else?



“Factiod” Question Answering

WEB IMAGES VIDEOS MAPS NEWS MORE 7 Sign in

bing who invented the light bulb

Also try: Who Invented the Telephone · Who Invented the Television · Who Invent…

1,270,000 RESULTS Any time

Incandescent light bulb inventors

Thomas Edison · Joseph Swan · Hiram Maxim

Data from Wikipedia

Light Bulb History - Invention of **the Light Bulb**
www.ideafinder.com/history/**invent**ions/**lightbulb**.htm
1878 **Thomas Edison** founded the **Edison Electric Light** Company 1878 **Hiram Maxim** ... **Thomas Edison**, **Joseph Swan**, **Hiram Maxim**, ... **Thomas Edison's light bulb** ...

Hiram Maxim - **Wikipedia**, the free encyclopedia
en.wikipedia.org/wiki/**Hiram**_Stevens_**Maxim**
This article is about **Hiram** Stevens **Maxim**. ... **Maxim invented** the ... One of these actions regarded the incandescent **bulb**, for which **Maxim** claimed that **Edison** was ...
Birth · Family · Emigration and knighthood · Profession · The **Maxim** machine gun

Who invented the light bulb? - Did you know?
didyouknow.org › inventions
Who invented the light bulb? No, it wasn't **Thomas Edison**. ... In the late 1870s, **Hiram** Stevens **Maxim** ... **Joseph Swan**'s 1878 invention ...

What connects **the light bulb** filament, and a machine gun ...
aarticles.net/technology-and-internet/16051-chto-svyazyvaet...
Hiram Maxim. Inventor. Inventor of the incandescent **bulbs** in the world is **Thomas Edison**, ... (**Hiram** Stevens **Maxim**, 1840-1916) **Thomas Edison** patented the …

Who REALLY **invented the Light Bulb** ? Because I believe it ...

Incandescent light bulb

An incandescent light bulb, incandescent lamp or incandescent light globe is an electric light which produces light with a wire filament heated to a high temperature by an electric current passing through it, until it glows. The hot filament is p… +

en.wikipedia.org

Patent holder: Thomas Edison

Inventors: Thomas Edison · Joseph Swan · Hiram Maxim

Related people


Thomas Edison
Patent holder


Joseph Swan
Inventor


Hiram Maxim
Inventor


Henry Woodward


Lewis Howard L…

People also search for

Google when was 2001 released marti.hearst@gmail.com

Web Shopping News Videos Images More Search tools

About 1,310,000,000 results (0.67 seconds)

January 1, 1968 (USA)

2001: A Space Odyssey, Release date

Feedback

2001: A Space Odyssey (film) - Wikipedia, the free ...
en.wikipedia.org/wiki/**2001**:_A_Space_Odyssey_(film) Wikipedia
Clarke concurrently wrote the novel **2001**: A Space Odyssey which was published soon after the film was **released**. The story deals with a series of encounters ...

2001: A Space Odyssey (no… 2001: A Space Odyssey is a 1968 science fiction novel by Arthur C ...

A Space Odyssey (soundtr… Lux Aeterna - The Blue Danube - Gayane - Atmosphères - ...

More results from wikipedia.org »

2001: A Space Odyssey (1968) - IMDb
www.imdb.com/title/tt0062622/ Internet Movie Database
★★★★☆ Rating: 8.3/10 - 313,399 votes
2001: A Space Odyssey Lost Footage Not Planned for **Release** ...
2001: A Space Odyssey -- The sci-fi masterpiece from acclaimed

2001: A Space Odyssey

G 2h 41m - Mystery/Fantasy

Watch trailer

★★★★☆ 8.3/10 - IMDb

Story follows the ascent of mankind into the near-future space age through minimalist performances and a strong visual style.

Release date: January 1, 1968 (USA)
Director: Stanley Kubrick
Sequel: 2010
Screenplay: Arthur C. Clarke, Stanley Kubrick
Music composed by: Aram Khachaturian, Richard Strauss, Johann Strauss II, György Ligeti

Cast

View 10+ more


Arthur C. Clarke

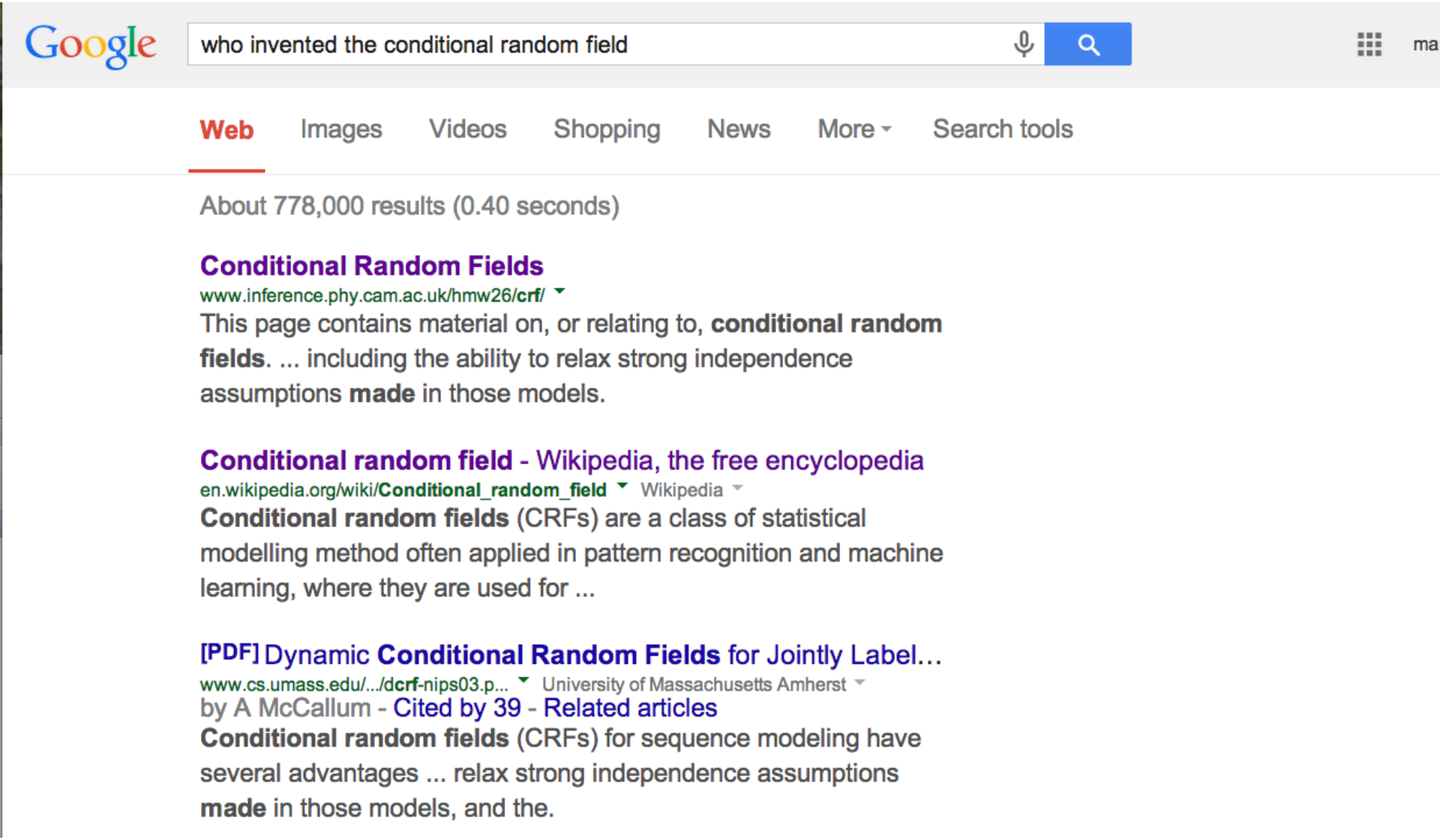

Keir Dullea
David Bowman


Gary Lockwood
Frank Poole


William Sylvester
Heywood R. Floyd

People also search for

View 15+ more


Google
who invented the conditional random field
ma
Web
Images
Videos
Shopping
News
More
Search tools
About 778,000 results (0.40 seconds)
Conditional Random Fields
www.inference.phy.cam.ac.uk/hmw26/crf/
This page contains material on, or relating to, conditional random fields. ... including the ability to relax strong independence assumptions made in those models.
Conditional random field - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Conditional_random_field Wikipedia
Conditional random fields (CRFs) are a class of statistical modelling method often applied in pattern recognition and machine learning, where they are used for ...
[PDF] Dynamic Conditional Random Fields for Jointly Label...
www.cs.umass.edu/.../dcrf-nips03.p... University of Massachusetts Amherst
by A McCallum - Cited by 39 - Related articles
Conditional random fields (CRFs) for sequence modeling have several advantages ... relax strong independence assumptions made in those models, and the.

Real-World Applications of NLP

- Spelling Suggestions
- Grammar Checking
- Synonym Generation
- Information Extraction
- Text Categorization
- Summarization
- Essay Scoring
- Automated Customer Service
- Speech Recognition
- Speech Generation
- Machine Translation
- Question Answering
- Improving Web Search Engine results
- Automated Metadata Assignment
- Online Dialogs

TEXT ANALYSIS FOR SOCIAL SCIENCES

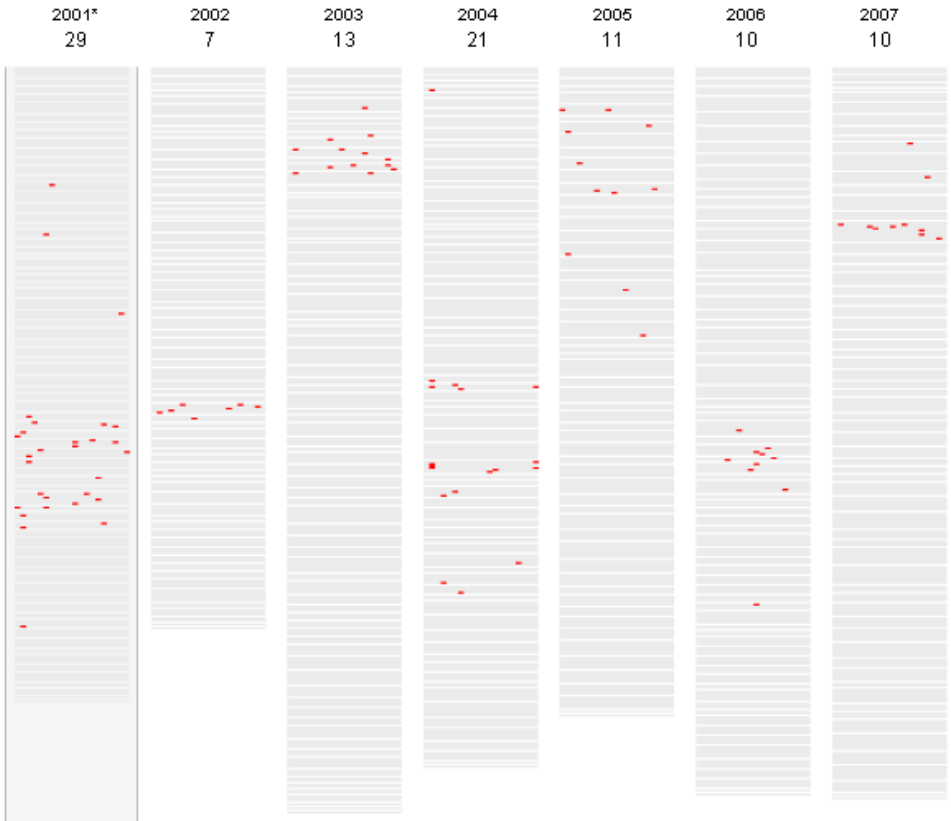
Example: State of the Union Addresses

The 2007 State of the Union Address

Over the years, President Bush's State of the Union address has averaged almost 5,000 words each, meaning the the President has delivered over 34,000 words. Some words appear frequently while others appear only sporadically. Use the tools below to analyze what Mr. Bush has said.

Search | or choose a word here.

Use of the phrase "Tax" in past State of the Union Addresses



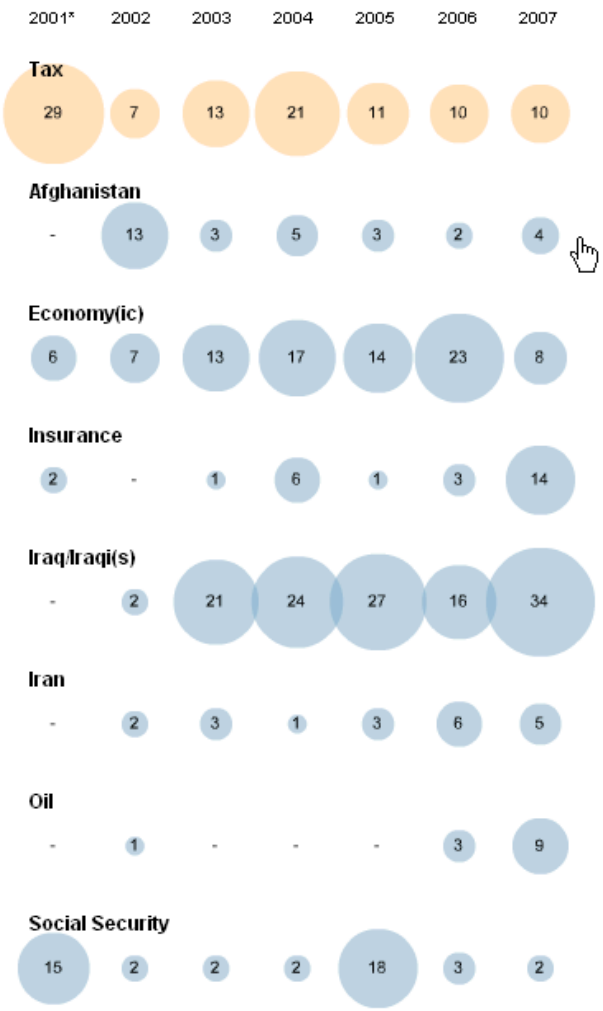
Next Instance of 'Tax'

The word in context

I believe in local control of schools. We should not, and we will not, run public schools from Washington, D.C. Yet when the federal government spends **TAX** dollars, we must insist on results. Children should be tested on basic reading and math skills every year between grades three and eight. Measuring is the only way to know whether all our children are learning. And I want to know, because I refuse to leave any child behind in America.

-- 2001 (Paragraph 14 of 73)

Compared with other words

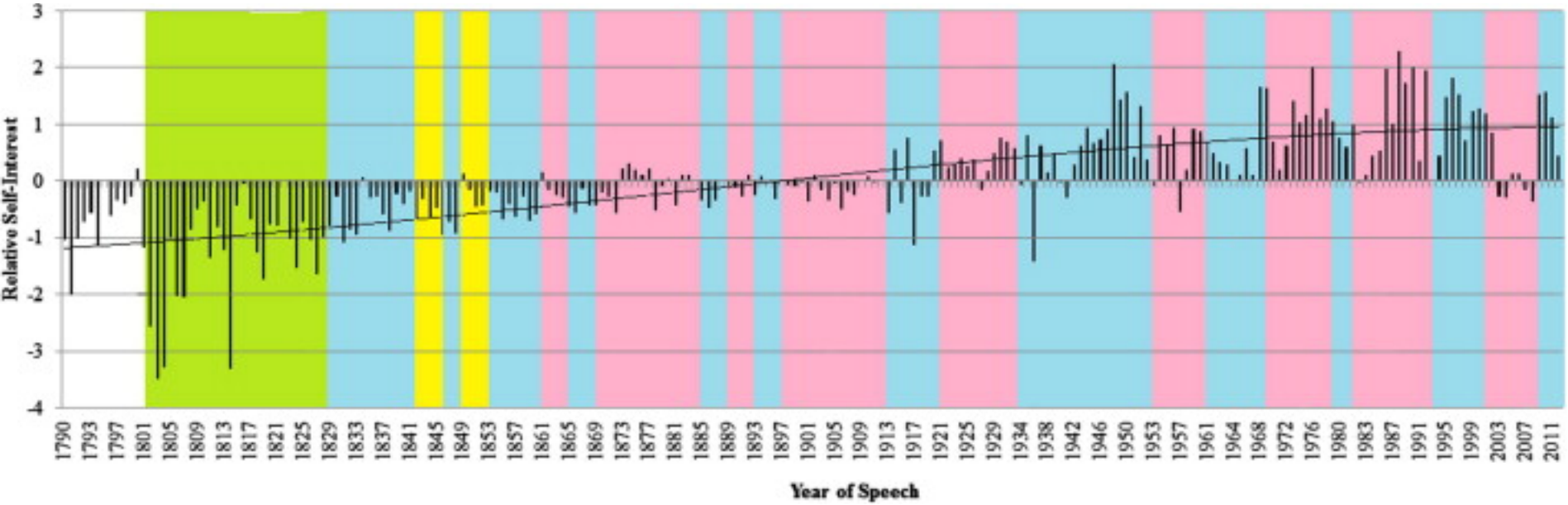


* As a newly elected president, Mr. Bush did not deliver a formal State of the Union address in 2001. His Feb. 27 speech to a joint session of Congress was analogous to the State of the Union, but without the title.

Ben Werschkul/The New York Times

The *New York Times*

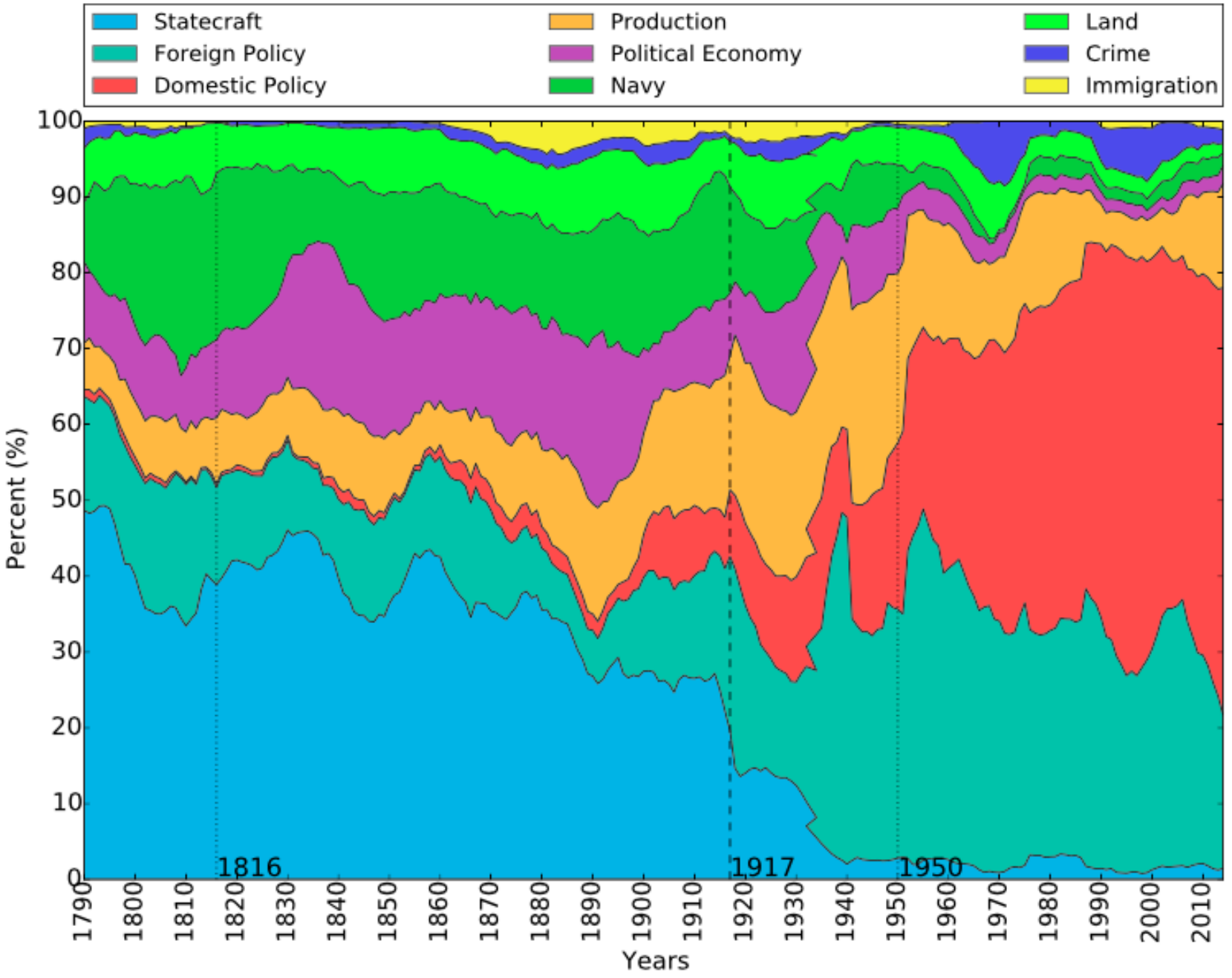
W.J. Chopik et al. / Personality and Individual Differences 66 (2014) 128–133



Republicans (Red), Democrats (Blue), Whigs (Yellow), Democratic-Republicans (Green).

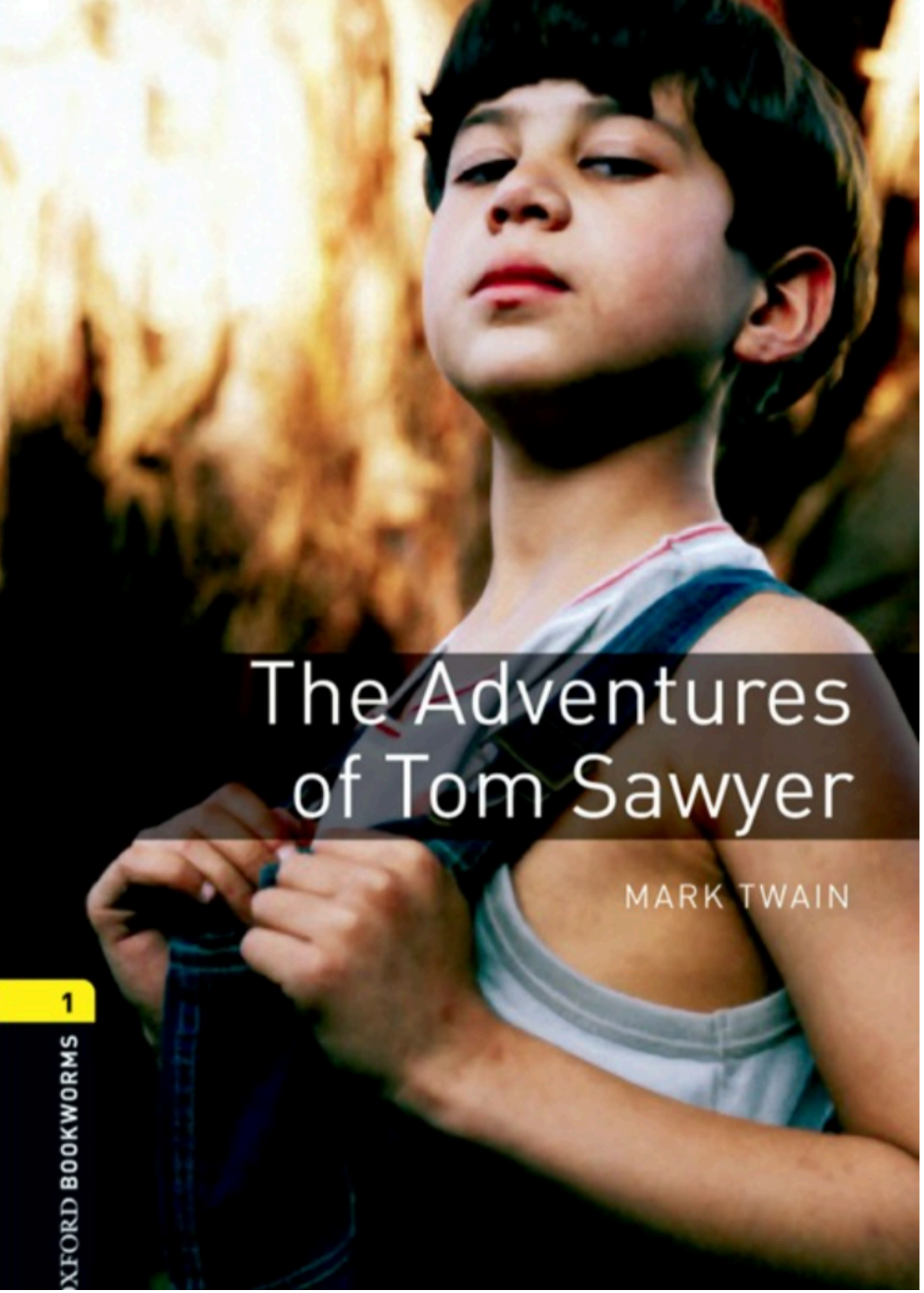
Pure self-interest (first-person singular pronouns),
Moderate self-interest (family references, first-person plural pronouns)
Other-interest (other-focused pronouns, friends references)
were simultaneously regressed onto year

The analysis revealed a decrease in words related to other-interest (e.g., "his/her," "neighbor") and an increase in words related to self-interest (e.g., "I, me, mine," "mother").



Rule, Cointet, and Bearman. "Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014." PNAS 2015

WHAT DO WE DO WHEN WE DO NLP?



"TOM!"

No answer.

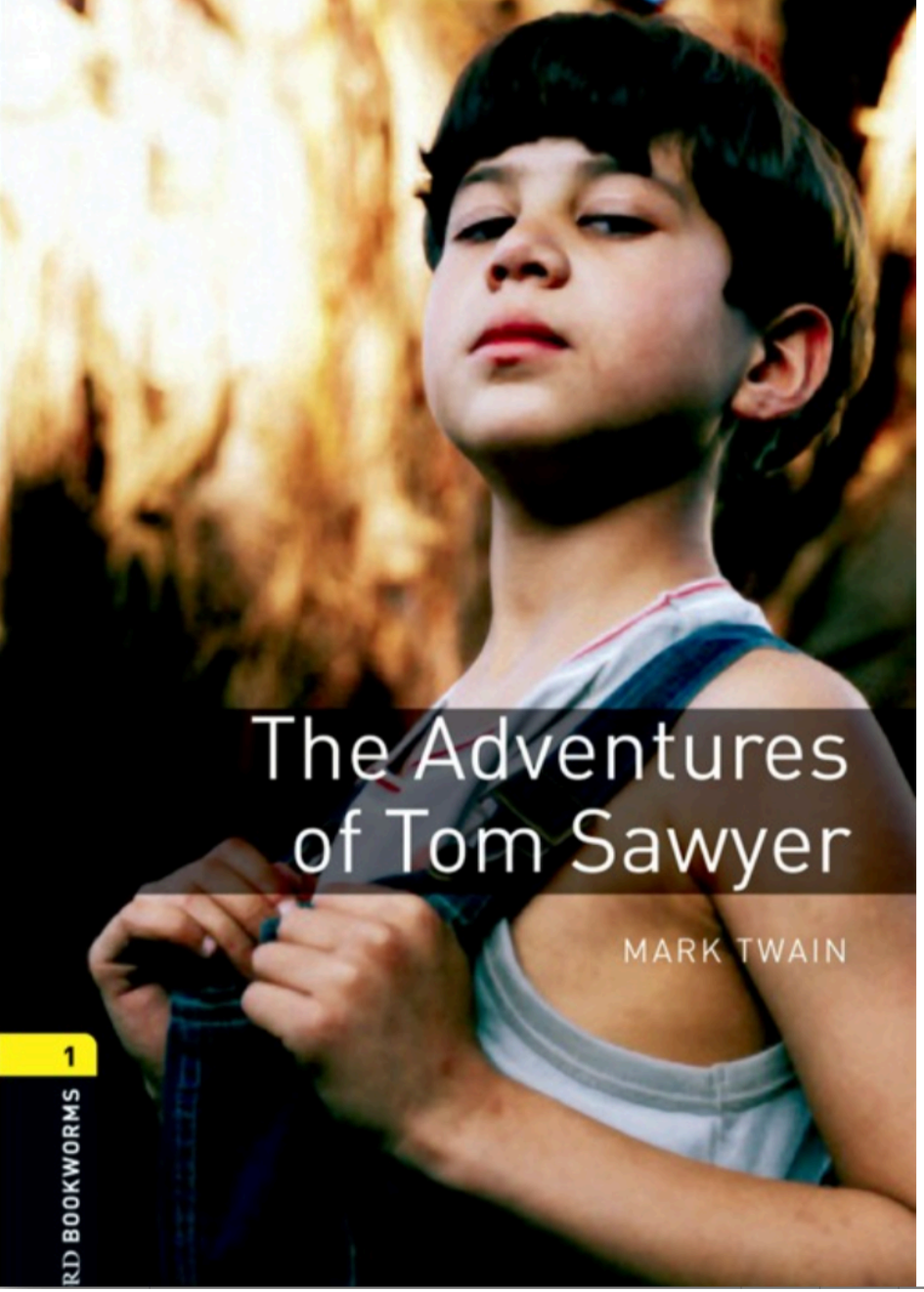
"TOM!"

No answer.

"What's gone with that boy, I wonder? You TOM!"

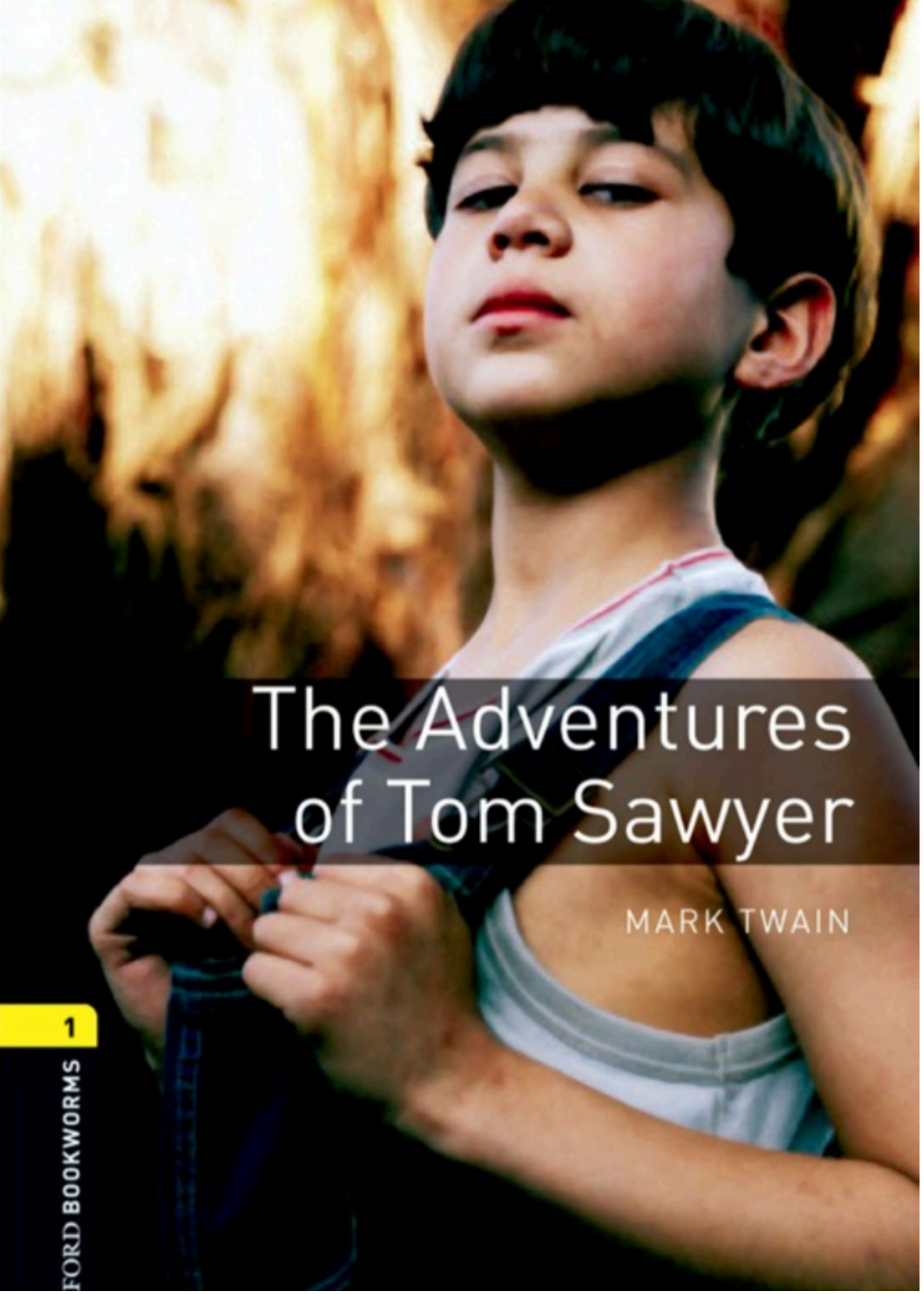
No answer.

The old lady pulled her spectacles down and looked over them about the room.



Bag of Words

tom no answer tom no answer what's gone with that boy , I wonder ? you tom ! no answer the old lady pulled her spectacles down and looked over them about the room .



nouns

"TOM!"

No answer.

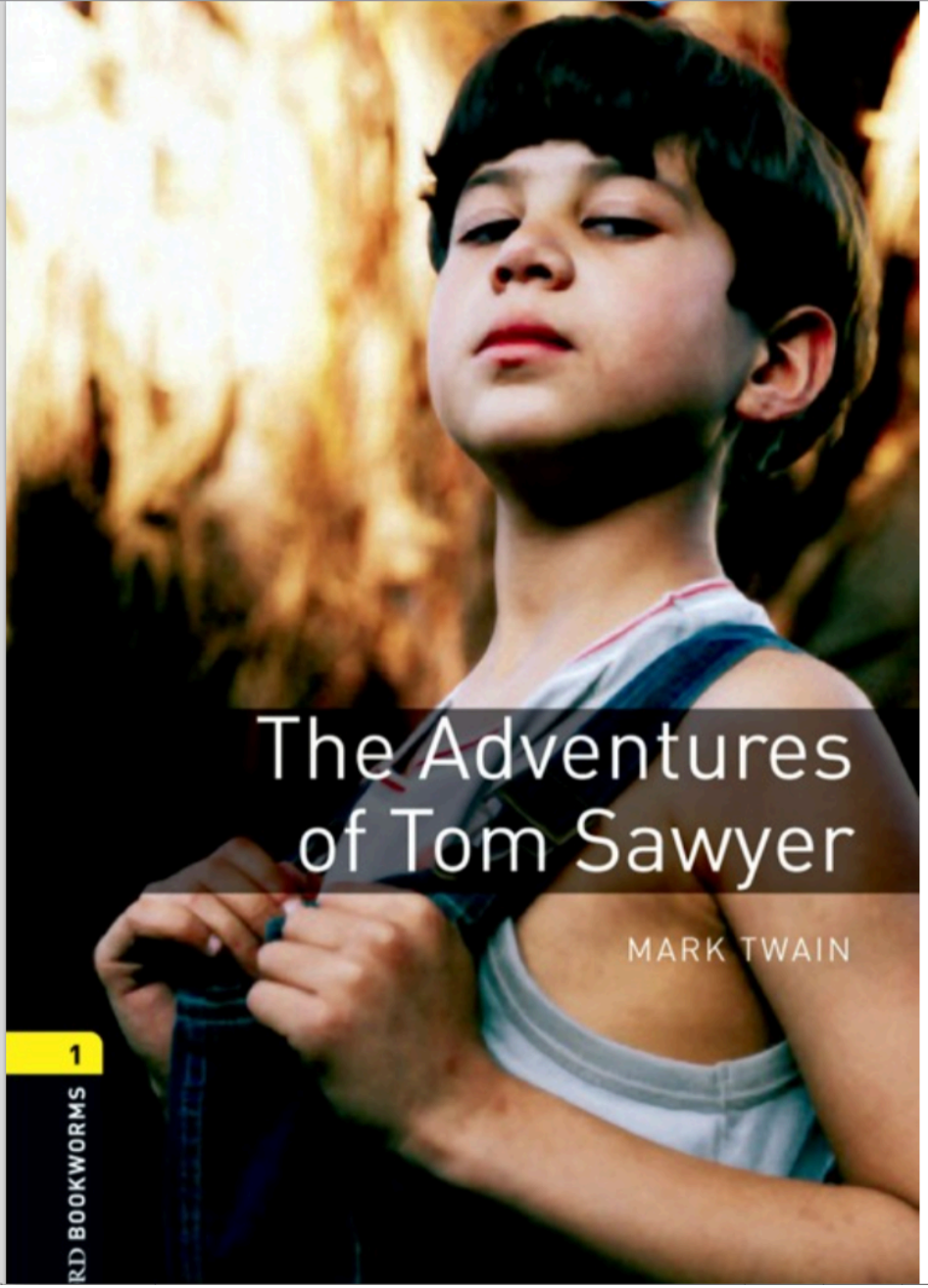
"TOM!"

No answer.

"What's gone with that boy, I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.



People (Proper Nouns)

"TOM!"

No answer.

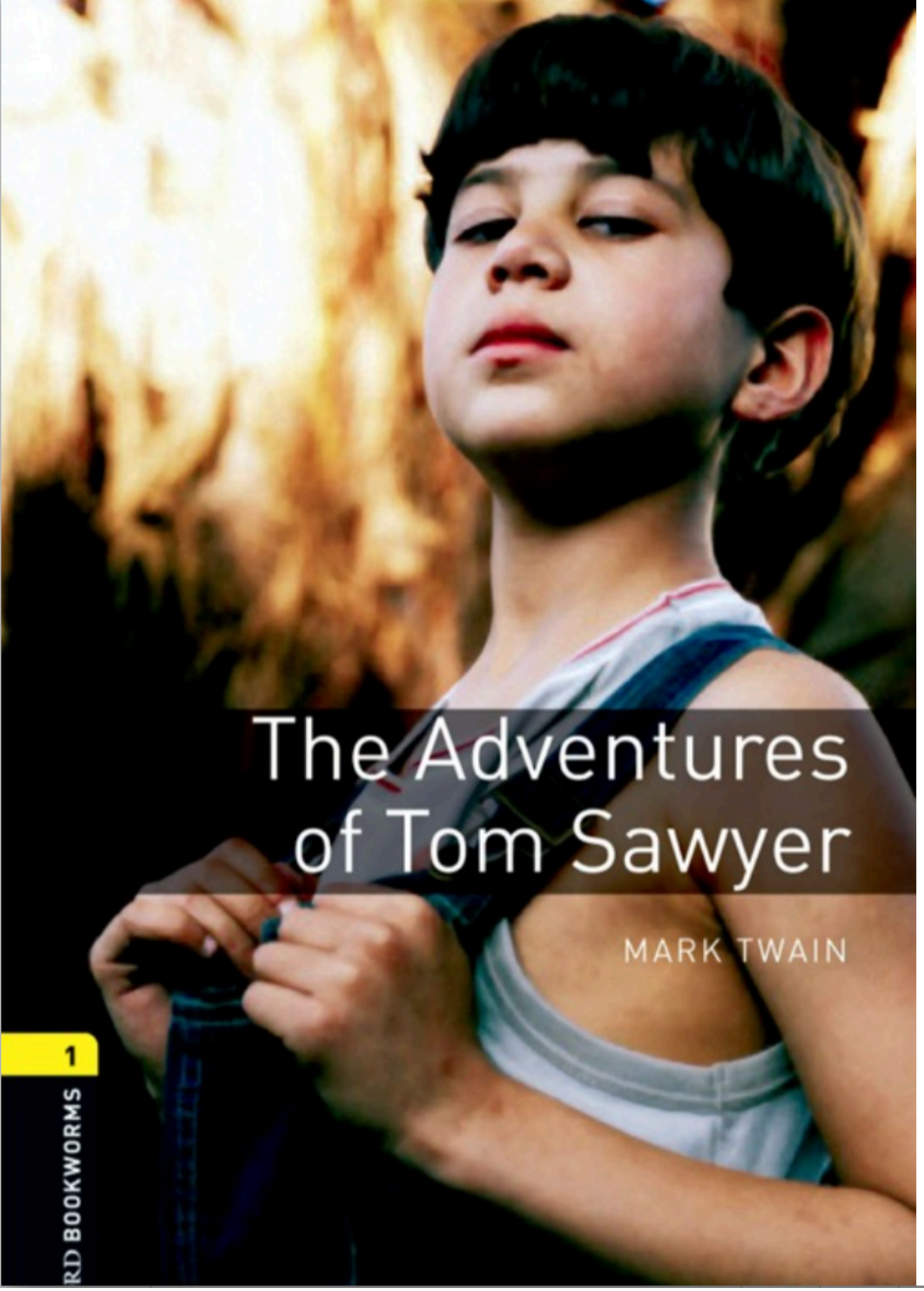
"TOM!"

No answer.

"What's gone with that boy, I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.



Relations (syntax)

"TOM!"

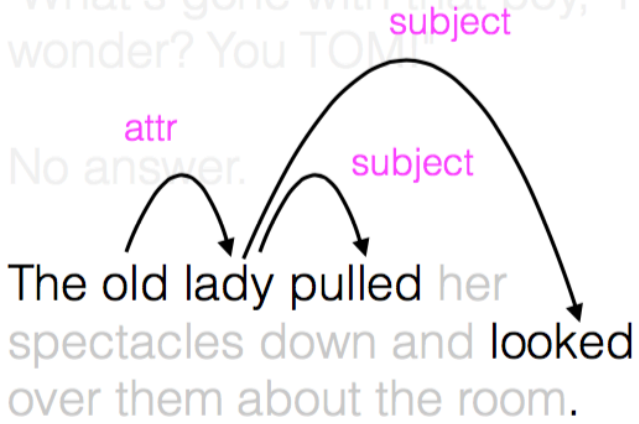
No answer.

"TOM!"

No answer.

"What's gone with that boy, I wonder? You TOM!"

No answer.



Speaker identification

"TOM!"

No answer.

"TOM!"

No answer.

"What's gone with that boy, I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.

Coreference

"TOM!"

No answer.

"TOM!"

No answer.

"What's gone with that boy, I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.

COURSE TOPICS

COURSE CONCEPT MAP

Segment-ation

Syntax

Semant-ics

Word-Level	Phrase-Level	Sentence-Level	Discourse-Level
Tokenization	Chunking	Sentence Boundary Detection	TextTiling
Morphology / Stemming / Part of Speech Tagging	Chunking / Information Extraction	Parsing	Rhetorical Structure Theory Parsing
Thesauri / Word Similarity	Information Extraction	Sentiment Classification / QA / Word Sense Disam.	Summarization/ Categorization / Discourse Analysis

Statistics

Note: Topics may be added or dropped.

COURSE CONCEPT MAP

Word-Level	Phrase-Level	Sentence-Level	Discourse-Level

single

the **single**

Ruthless released **the single** "Panic Zone" in 1987 with Macola Records, which was later included on the compilation album N.W.A. and the Posse.

Ruthless released the single "Panic Zone" in 1987 with Macola Records, which was later included on the compilation album N.W.A. and the Posse. N.W.A was still in its developing stages, and is only credited on three of the eleven tracks, notably the uncharacteristic record "Panic Zone," "8-Ball," "Dopeman," which marked the first collaboration of Arabian Prince, DJ Yella, Dr. Dre, and Ice Cube. …

Linguistic Characterizations

Linguistic Characterizations

Ruthless released the single "Panic Zone" in 1987 with Macola Records, which was later included on the compilation album N.W.A. and the Posse. N.W.A was still in its developing stages, and is only credited on three of the eleven tracks, notably the uncharacteristic record "Panic Zone," "8-Ball," "Dopeman," which marked the first collaboration of Arabian Prince, DJ Yella, Dr. Dre, and Ice Cube. Mexican rapper Krazy-Dee …

Linguistic Characterizations

Segment-ation

Syntax

Semant-ics

Ruthless released the single "Panic Zone" in 1987 with Macola Records, which was later included on the compilation album N.W.A. and the Posse. N.W.A was still in its developing stages, and is only credited on three of the eleven tracks, notably the uncharacteristic record "Panic Zone," "8-Ball," "Dopeman," which marked the first collaboration of Arabian Prince, DJ Yella, Dr. Dre, and Ice Cube. Mexican rapper Krazy-Dee …

SEGMENTATION

The process of dividing text into meaningful units, such as words, sentences, or topics. (English Wikipedia)

Segmentation: Sentence Level

Segment-ation

Ruthless released the single "Panic Zone" in 1987 with Macola Records, which was later included on the compilation album N.W.A. and the Posse. N.W.A was still in its developing stages, and is only credited on three of the eleven tracks, notably the uncharacteristic record "Panic Zone," "8-Ball," "Dopeman," which marked the first collaboration of Arabian Prince, DJ Yella, Dr. Dre, and Ice Cube. Mexican rapper Krazy-Dee …

Segmentation: Sentence Level

Segment-ation

Ruthless released the single "Panic Zone" in 1987 with Macola Records, which was later included on the compilation album N.W.A. and the Posse.

N.W.A was still in its developing stages, and is only credited on three of the eleven tracks, notably the uncharacteristic record "Panic Zone," "8-Ball," "Dopeman," which marked the first collaboration of Arabian Prince, DJ Yella, Dr. Dre, and Ice Cube.

Mexican rapper Krazy-Dee …

SYNTAX

The way in which words are put together to form phrases, clauses, or sentences; the part of grammar dealing with this.

(Merriam-Webster free dictionary).

Parse

Syntax

```
(ROOT
  (S
    (NP (NNS Ruthless))
    (VP (VBD released)
      (NP
        (NP (DT the) (JJ single) (`` ``) (NN Panic) (NN Zone) ('' ''))
        (PP (IN in)
          (NP (CD 1987))))
      (PP (IN with)
        (NP
          (NP (NNP Macola) (NNPS Records))
          (, ,)
          (SBAR
            (WHNP (WDT which))
            (S
              (VP (VBD was)
                (ADVP (RB later))
                (VP (VBN included)
                  (PP (IN on)
                    (NP
                      (NP (DT the) (NN compilation) (NN album) (NN N.W.A.))
                      (CC and)
                      (NP (DT the) (NN Posse)))))))))))
    (. .)))
```

```
nsubj(released-2, Ruthless-1)
root(ROOT-0, released-2)
det(Zone-7, the-3)
amod(Zone-7, single-4)
nn(Zone-7, Panic-6)
dobj(released-2, Zone-7)
prep(Zone-7, in-9)
pobj(in-9, 1987-10)
prep(released-2, with-11)
nn(Records-13, Macola-12)
pobj(with-11, Records-13)
nsubjpass(included-18, which-15)
auxpass(included-18, was-16)
advmod(included-18, later-17)
rcmod(Records-13, included-18)
prep(included-18, on-19)
det(N.W.A.-23, the-20)
nn(N.W.A.-23, compilation-21)
nn(N.W.A.-23, album-22)
pobj(on-19, N.W.A.-23)
cc(N.W.A.-23, and-24)
det(Posse-26, the-25)
conj(N.W.A.-23, Posse-26)
```

Syntax

http://nlp.stanford.edu:8080/parser/index.jsp

SEMANTICS

The study of the meanings of words and phrases in language.
(Merriam-Webster Free Dictionary)

What Does This Mean?

Semant-ics

Ruthless released the single "Panic Zone" in 1987 with Macola Records, which was later included on the compilation album N.W.A. and the Posse.

Semantic Analysis
(settling for information extraction)

Semant-ics

Ruthless released the single "Panic Zone" in 1987 with Macola Records, which was later included on the compilation album N.W.A. and the Posse.

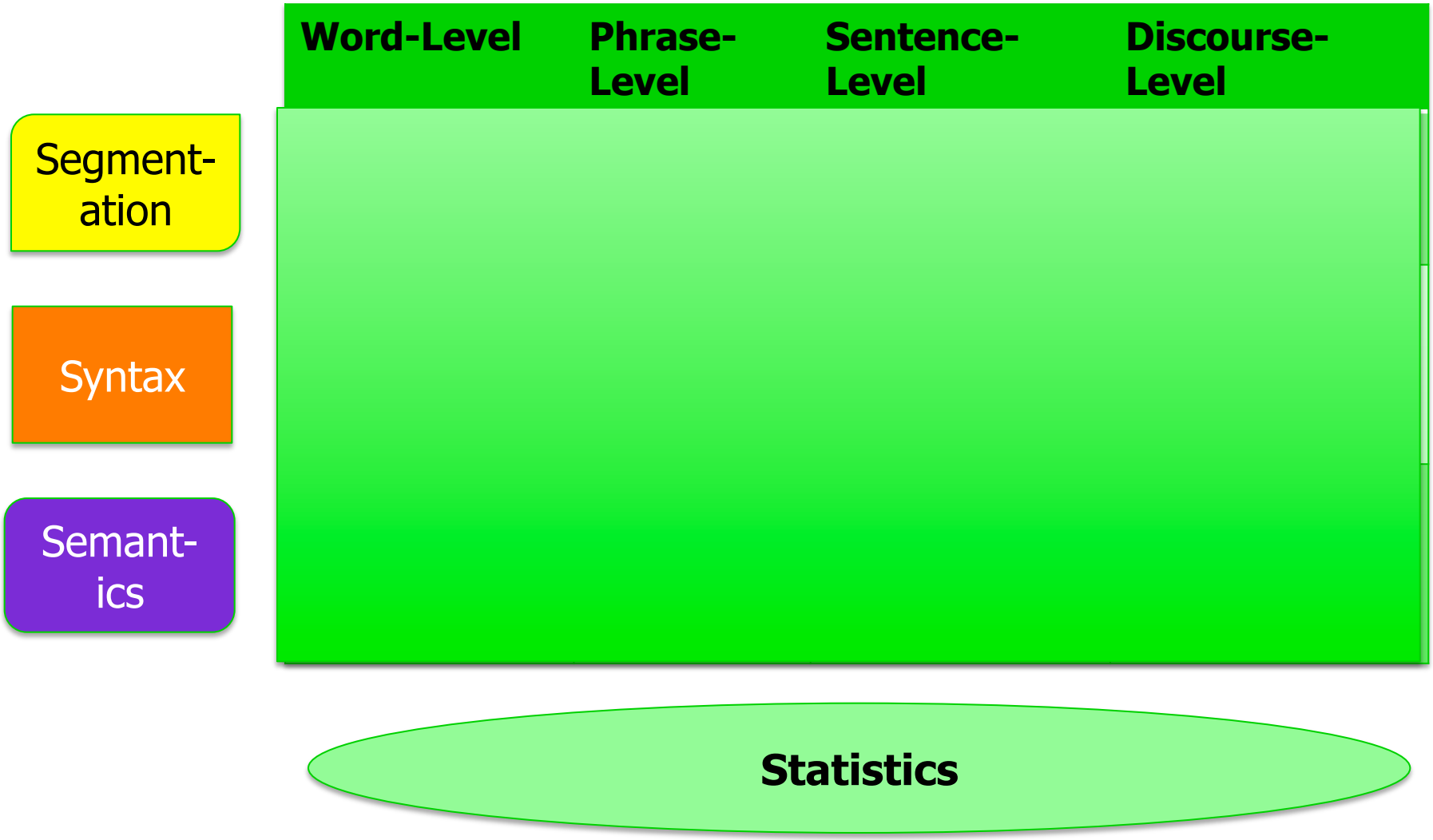
Submit Clear

Ruthless released the single "Panic Zone" in 1987 with Macola Records, which was later included on the compilation album N.W.A. and the Posse.

Potential tags:
- LOCATION
- TIME
- PERSON
- ORGANIZATION
- MONEY
- PERCENT
- DATE

http://nlp.stanford.edu:8080/ner/process

COURSE CONCEPT MAP



Note: Topics may be added or dropped.

KEY TECHNIQUES

- Looking at Data / Error Analysis
- Evaluating with Training / Development / Test Sets
- Predicting with Sequences (ngrams / Language models)
- Machine Learning with Feature Creation and Selection
- Word Windows and Context Vectors

The Importance of World and Language Knowledge

- He arrived at the lecture.
- He chuckled at the lecture.

- He arrived drunk.
- He chuckled drunk.

- He chuckled his way through the lecture.
- ✘ He arrived his way through the lecture.

Exercise:
Indiv: (1m) Pair (2m)

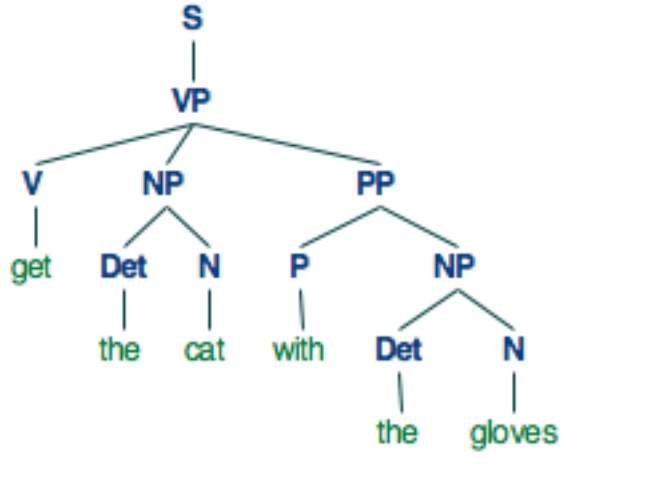
Describe in linguistic and semantic terms the two different meanings of this sentence as depicted by the two images:

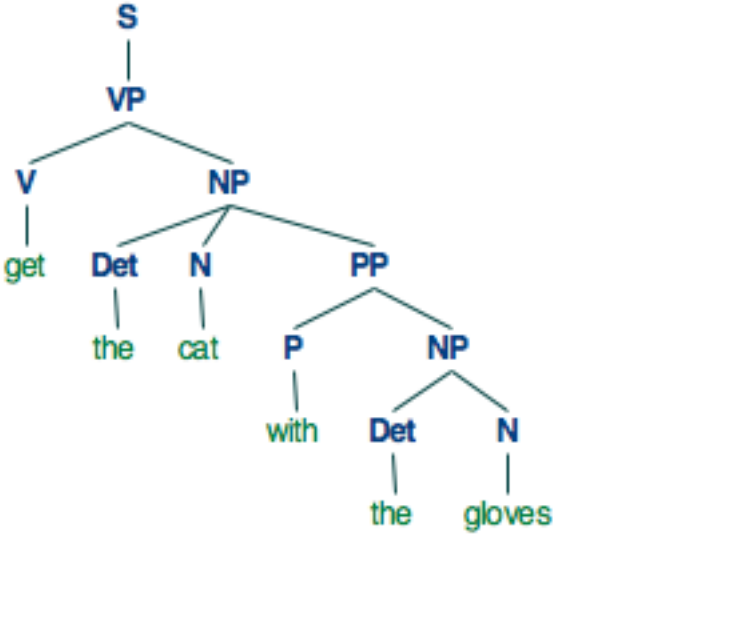
"Get the cat with the gloves."

(a)

(b)

Two Syntactic Structures





Two Syntactic Structures; So Easy with NLTK

In [3]:
```
import nltk
from nltk.draw.tree import draw_trees
```

In [7]:
```
grammar1 = nltk.CFG.fromstring("""
S -> NP VP | VP
VP -> V NP | V NP PP
PP -> P NP
V -> "saw" | "ate" | "walked" | "get" | "got"
NP -> "John" | "Mary" | "Bob" | Det N | Det N PP
Det -> "a" | "an" | "the" | "my"
N -> "man" | "dog" | "cat" | "telescope" |"park" | "gloves"
P -> "in" | "on" | "by" | "with"
""")
```

In [*]:
```
def quick_parse(sentence):
    cat_sent1 = sentence.split()
    ch_parser = nltk.ChartParser(grammar1)
    trees = ch_parser.parse(cat_sent1)
    for tree in trees:
        draw_trees(tree)
quick_parse("get the cat with the gloves")
```

For Wednesday: Chapter 1: Two Classes/Data Structures

Text

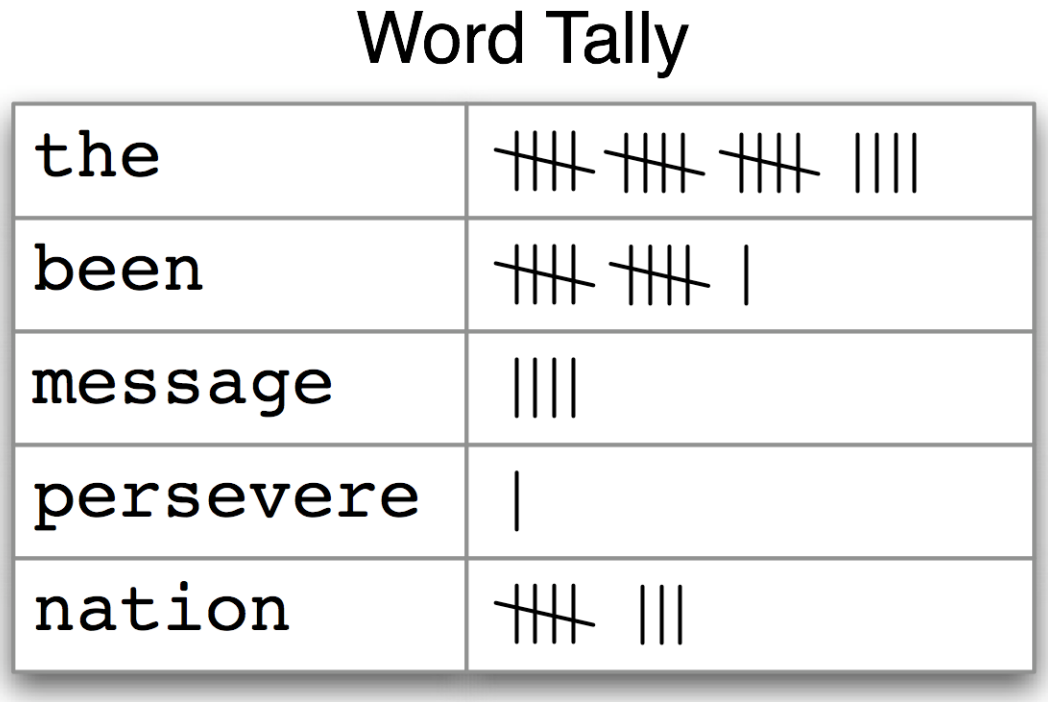
- Processed texts
  - One long list of words
  - No sentence markers
- Analysis tools
  - Word counting
  - Print concordance
  - Similarity between words
  - Plot dispersion
  - Regex search words
  - Collocations, bigrams

FreqDist

- Quick way to count all the items in a list. Also:
  - Frequency
  - Tabulate
  - Plot
  - Max
  - Iterate through
  - Compare size of distributions

NLTK Frequency Distribution

A convenient way to compute a tally.



NLTK Frequency Distribution

A convenient way to compute a tally.

NLTK's Frequency Distribution Object

This data structure makes it easy to tally up frequencies across words and other items, and incorporate them into list comprehensions (and later we'll see the conditional frequency distribution as well).

These are the functions supported by FreqDis:

```
fdist = FreqDist(samples) create a frequency distribution containing the given samples
fdist[sample] += 1    increment the count for this sample
fdist['monstrous']    count of the number of times a given sample occurred
fdist.freq('monstrous')    frequency of a given sample
fdist.N()    total number of samples
fdist.most_common(n)    the n most common samples and their frequencies
for sample in fdist:    iterate over the samples
fdist.max()    sample with the greatest count
fdist.tabulate()    tabulate the frequency distribution
fdist.plot()    graphical plot of the frequency distribution
fdist.plot(cumulative=True)    cumulative plot of the frequency distribution
fdist1 |= fdist2    update fdist1 with counts from fdist2
fdist1 < fdist2    test if samples in fdist1 occur less frequently than in fdist2
```

How to quickly find the bad or weird characters in a collection?

CODING PRACTICE

Here is how to quickly see the bad characters in a collection.

First, read in the State of the Union Corpus (it is part of NLTK).

In [14]:
```
import nltk
from nltk.corpus import state_union
```

Here is how to quickly see the bad characters in a collection.

First, read in the State of the Union Corpus (it is part of NLTK).

In [14]:
```
import nltk
from nltk.corpus import state_union
```

Compute a FreqDist over the Characters; Show how many Chars total

In [15]:
```
sotufd = nltk.FreqDist(state_union.raw())
sotufd.N()
```

Out[15]:
```
2073698
```

Here is how to quickly see the bad characters in a collection.

First, read in the State of the Union Corpus (it is part of NLTK).

In [14]:
```
import nltk
from nltk.corpus import state_union
```

Compute a FreqDist over the Characters; Show how many Chars total

In [15]:
```
sotufd = nltk.FreqDist(state_union.raw())
sotufd.N()
```

Out[15]:
```
2073698
```

Show all the character counts; Look at the end of the array to see the weird characters

In [16]:
```
l = sotufd.most_common(200)
l[-10:]
```

Out[16]:
```
[('>', 16),
 ('Z', 13),
 ('§', 12),
 ('¨', 12),
 ('@', 11),
 ('\\', 6),
 ('*', 2),
 ('~', 1),
 ('`', 1),
 ('%', 1)]
```

For Wednesday

- Adopt a text collection: what does this mean?
- Treating text as data; we'll get some practice

Exercise

- Pair up: do one of these:

1. Work on getting ipython notebook problems worked out
2. Start working on the assignment due on Wed.