# NLP for the Web – WS 2016/2017 Lecture 1 Introduction


**Prof. Dr. Chris Biemann**
**Seid Muhie Yimam, MA**

**AG Sprachtechnologie / Language Technology**
**FB Informatik / Computer Science**
**Universität Hamburg / University of Hamburg**

***Slides partially adopted from Torsten Zesch, Györgi Szarvas, Richard Eckard de Castilho, Delphine Bernhard and others***

# Introduction: Lecturers

**Prof. Dr. Chris Biemann**
biemann@informatik.uni-hamburg.de

Informatikum, F-429
Tel. 040 / 42883 - 2386
office hours by appointment

**Seid Muhie Yimam, MA**
yimam@informatik.uni-hamburg.de

Informatikum, F-415
Tel. 040 / 42883 - 2418
office hours by appointment

# AG Sprachtechnologie / Language Technology

- LT@UHH since WS 2016 - http://www.lt.tu-darmstadt.de
- Projects:
  - ABSA-DB: Aspect-based Sentiment Analysis for DB Products and Services (Deutsche Bahn)
  - DIVID-DJ: Data Extraction and Interactive Visualization of Unexplored Textual Datasets for Investigative Data-Driven Journalism (Der Spiegel)
  - JOIN-T: Joining Ontologies and Semantics Induced from Text (DFG)
  - SEMSCH: Semantic Writing Aid (DFG)
- Industry Partners
  - IBM Research USA
  - SIEMENS AG, München
  - Deutsche Bahn, Frankfurt
  - Spiegel Verlag

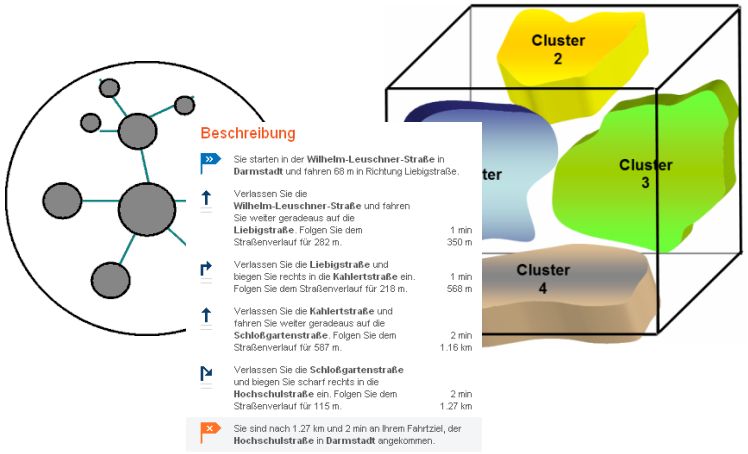
# Selection of recently finished theses @ LT

- Peter Klöckner: **Learning to Identify Antonymy**. MA Thesis, 08/2016
- Maria Pelevina**: Contextualization based on Word Sense Embeddings**. MA Thesis, 08/2016
- Jonas Wacker, **Ambient Search - just-in-time document recommendations for conversations**, BA Thesis, 03/2016
- Dominik Sobania**: A Focused Web Crawler Driven by Self-Optimizing Classifiers**, MA Thesis, 09/2015
- Johannes Simon: **Word Sense Induction Using Distributional Semantics**. MA Thesis, 05/2015
- Tim Feuerbach**, Including distributional semantics in a coreference system**, BA Thesis TU Darmstadt 09/2014
- Gerold Hintz, **Lojban as a middle ground between natural language and semantic ontology**, MA thesis, TU Darmstadt 08/2014
- Uli Fahrer**, Analysis of the German Parliament Elections 2013 with Twitter**, BA Thesis, TU Darmstadt 05/2014
- Benjamin Milde**: Unsupervised acquisition of acoustic models for speech-to-text alignment**, MA Thesis, in collaboration with Dirk Schnelle-Walka, TK group, TU Darmstadt 04/2014

# Vision

information


knowledge



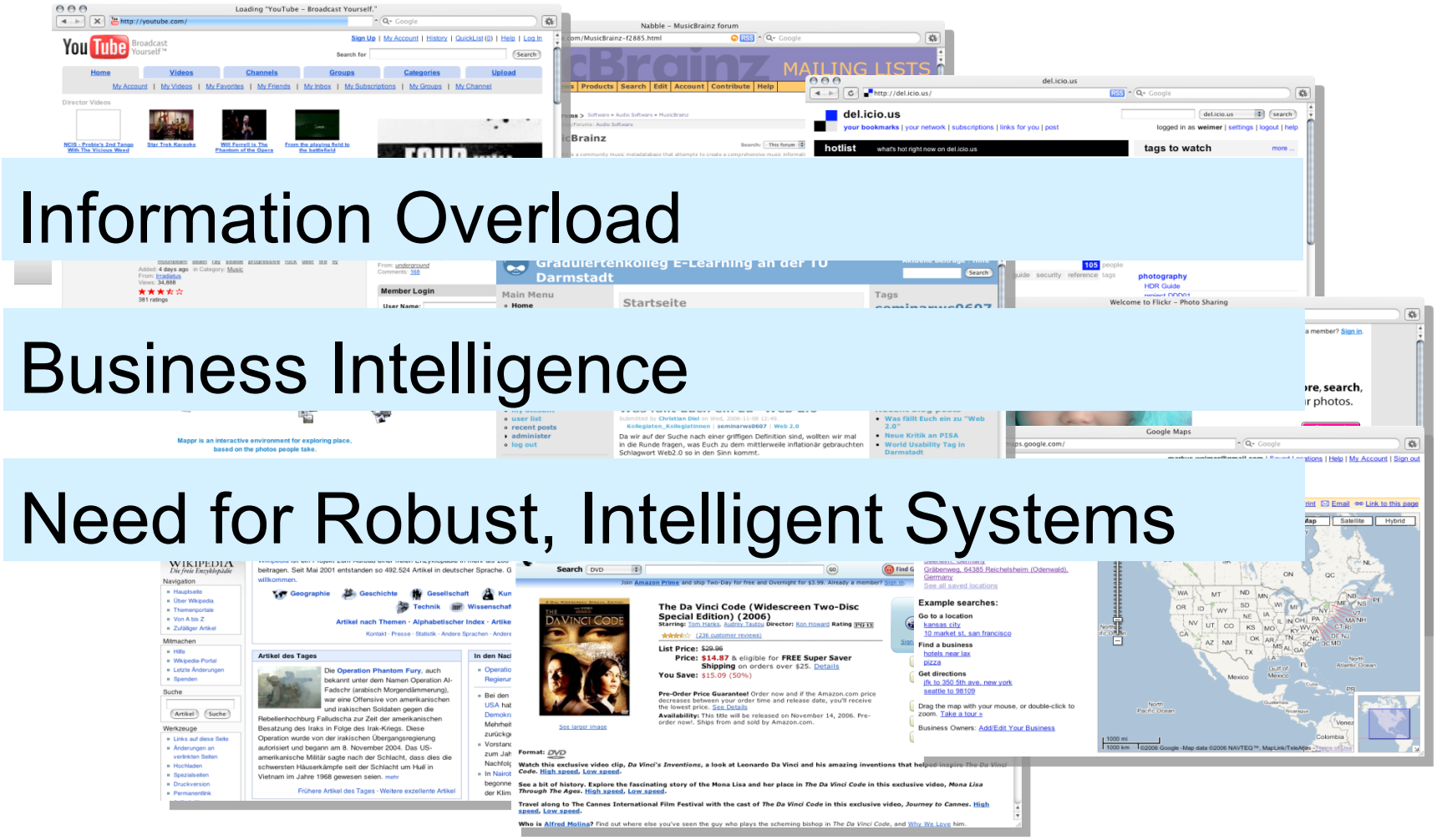

Focus on text and speech in NLP

NLP – Natural Language Processing

# Course Goals

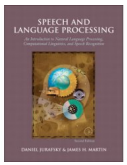
- Learn the basic principles underlying **NLP systems** and the **Unstructured Information Management Architecture (UIMA)**
- Learn **techniques and tools** used to:
  - explore the Web as a corpus
  - perform Web Mining
  - provide NLP based UIM applications for the Web
- Gain insight into **open research problems** in natural language processing

# Why Care?

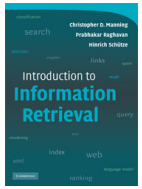

Information Overload
Business Intelligence
Need for Robust, Intelligent Systems

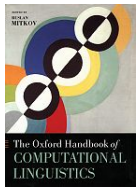
# Textbooks

- Computerlinguistik und Sprachtechnologie. Eine Einführung. Kai-Uwe Carstensen, Christian Ebert, Cornelia Endriss, Susanne Jekat, Ralf Klabunde. Heidelberg: Spektrum-Verlag, (egal welche Auflage).
  - http://www.linguistics.rub.de/CLBuch/buch.html
  - http://dx.doi.org/10.1007/978-3-8274-2224-8 (online version)
- Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Second Edition. Daniel Jurafsky and James H. Martin. Prentice-Hall, 2nd edition, 2008.
  - http://www.cs.colorado.edu/~martin/slp.html
- Chris Manning and Hinrich Schütze, Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA: May 1999.
  - http://nlp.stanford.edu/fsnlp/
- Chris Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.
  - http://nlp.stanford.edu/IR-book/information-retrieval-book.html
- Ruslan Mitkov, The Oxford handbook of computational linguistics, OxfordUniversity Press. 2003.
  - http://ukcatalogue.oup.com/product/9780198238829.do





# General Information

- There will be a final exam for the NLP4Web lecture
- In the practice classes you will work on a project
  - These will give you some practical experience in NLP
  - The assignments will be graded
  - They are mandatory for passing the Module!
- The practice classes are supervised by
  - Seid Muhie Yiman yimam@informatik.uni-hamburg,.de
- The lecture slides, handouts, readings etc. can all be found on the Moodle platform:
  - course name: Natural Language Processing and the Web (NLP'n'WEB)
  - enrolment key: NLP4WEB2017

# General Information (II)

- **At the beginning of each lecture → literature list for this lecture**
  - Literature can be found on the Web or in the Moodle platform
- **We will assume knowledge of what's in the mandatory parts**
  - Could be asked in the exam, even if it was not explicitly covered on the lecture slides
- **Reading also the optional parts is strongly encouraged**
- **Most lectures have a references section**
  - Interesting reads if you want to know more
- If you are interested and might want specific literature hints → feel free to ask

# Final exam

- **When and where:**
  TBA , ~beginning of March, 2017
- **Relevant registration times as usual**
- **Content:** lecture, readings, practice class

You will be able to pass the final exam based on the lecture only, but knowledge from the practice class will help you attain a good grade

# Practice class

- Practice class: 12:15 – 13:45 in G-102 on Tuesdays **-** First practice class starts right today
- Successfully passing the class means you reach 50% of points in the exercises.
- You should bring your own laptop
- If you need additional help regarding the practice class, send an email and ask for an appointment

The assignments in the practice class will require a significant amount of time, so do not start the day before.

# Practice class (II)

- **Milestone 1 (20%)** – Creating UIMA Pipelines
  - Submission: Homework solution + Pipeline code
- **Milestone 2 (30%)** – Creating processing component using Machine Learning
  - Submission: Documentation + Code + Short presentation
- **Milestone 3 (50%)** – NLP Project (in groups)
  - Submission: Documentation + Code + Final presentation

# Programming framework: UIMA

- „Unstructured Information Management Architecture“
- Open source incubator project at Apache
- Framework for NLP applications, provides
  - Component types with special interfaces,
  - Data types, and
  - Control flow between components (*pipeline*-style)
- Used in research, but also in e.g. IBMs Business Intelligence Apps
- Emerging standard (research and industry) → Important to know!
- **Using UIMA will be mandatory for some tasks**

Questions, suggestions, etc.

E-Mail to:

biemann@informatik.uni-hamburg.de

(lecture)

yimam@informatik.uni-hamburg.de

(practice class)

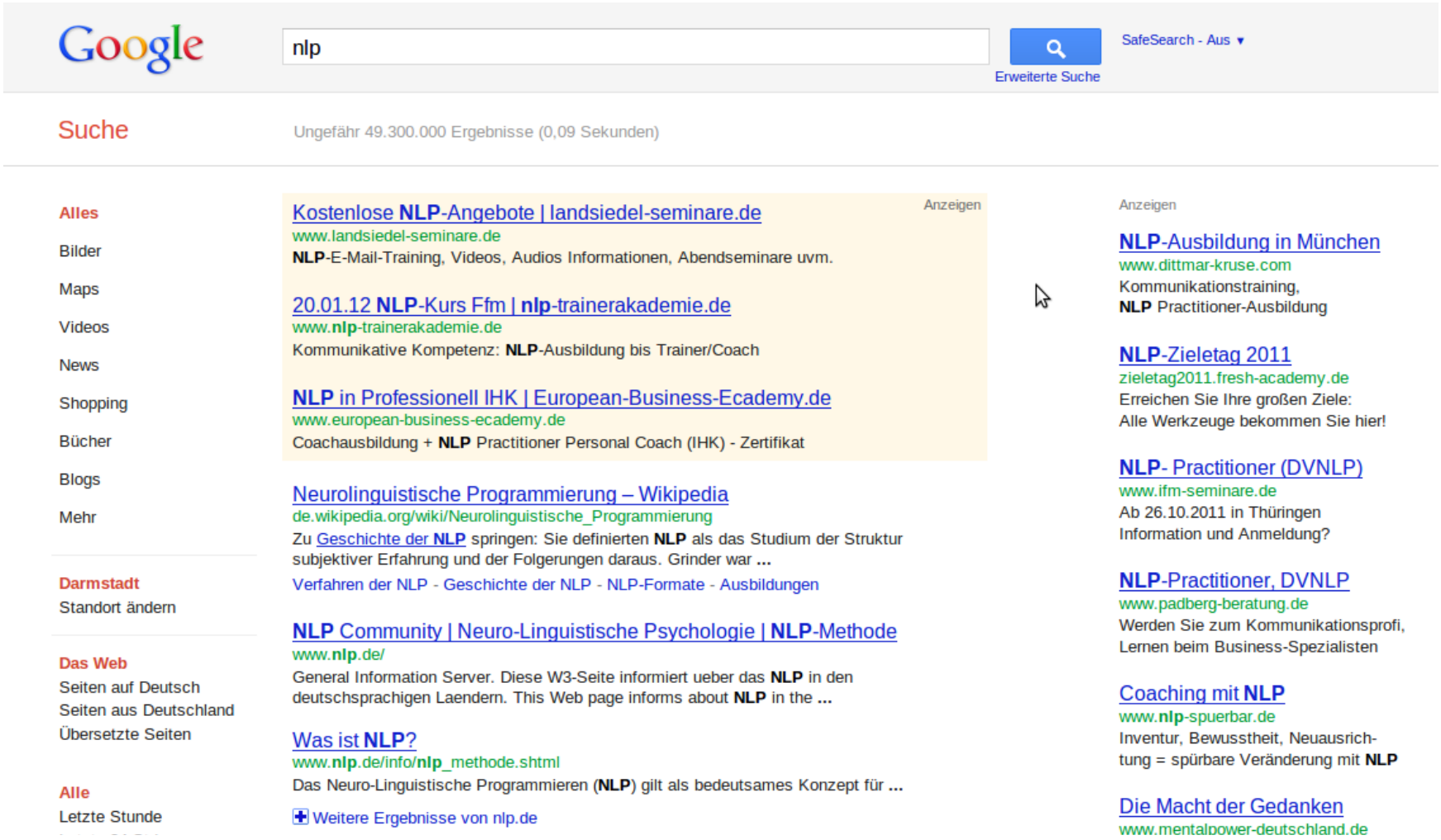
# Syllabus

1. NLP, Web 2.0
2. Levels of Linguistic Analysis
3. UIMA: Introduction
4. UIMA: Architecture and Applications
5. Machine Learning for Genre and POS tagging
6. Information Retrieval I
7. Information Retrieval II
8. Summarization
9. Lexical Semantic Knowledge from Wiki{pedia, onary}
10. Question Answering
11. Subjectivity and Sentiment I
12. Subjectivity and Sentiment II
13. Web as Corpus I
14. Web as Corpus II

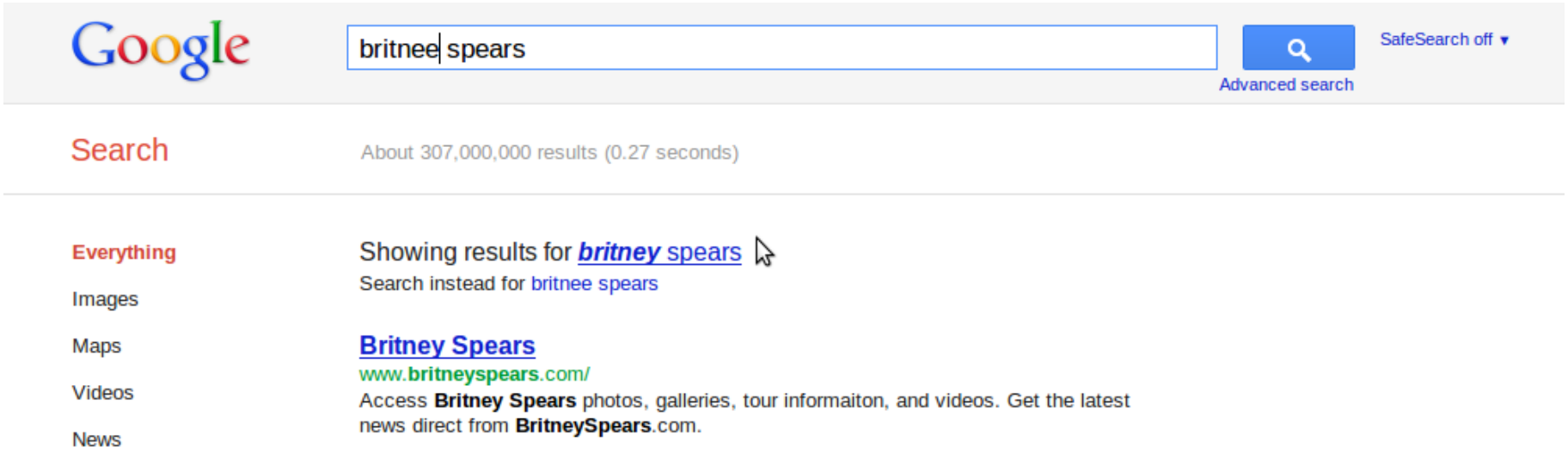
# Warm up

- Hands up
  - Computer Science?
  - Other disciplines?
  - Bachelor?
  - Master?
- Other NLP Lectures?
  - at UHH?
  - elsewhere?

# NLP in the Web – Search Engines



# NLP in the Web – Spelling Correction

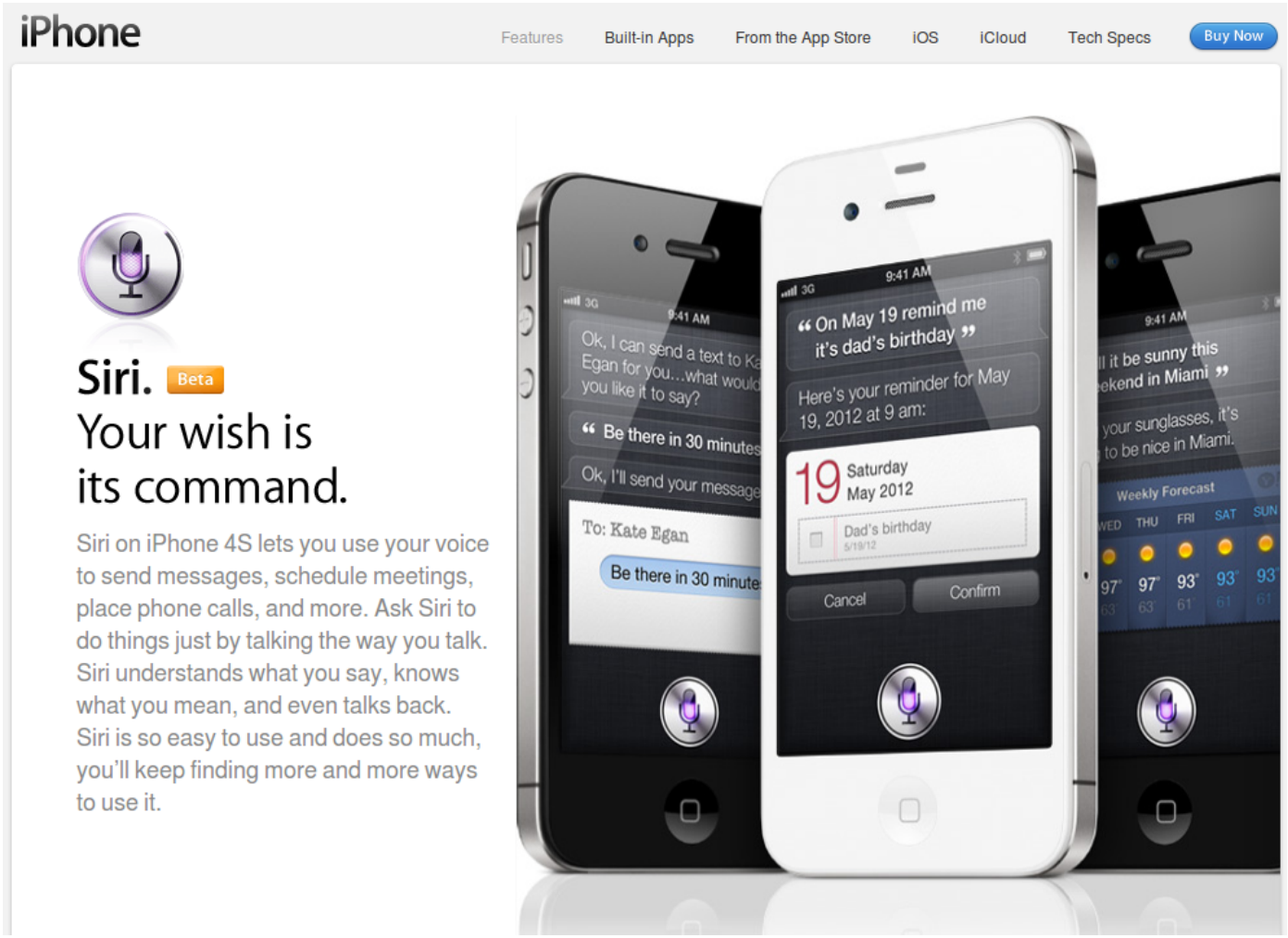


# Question Answering



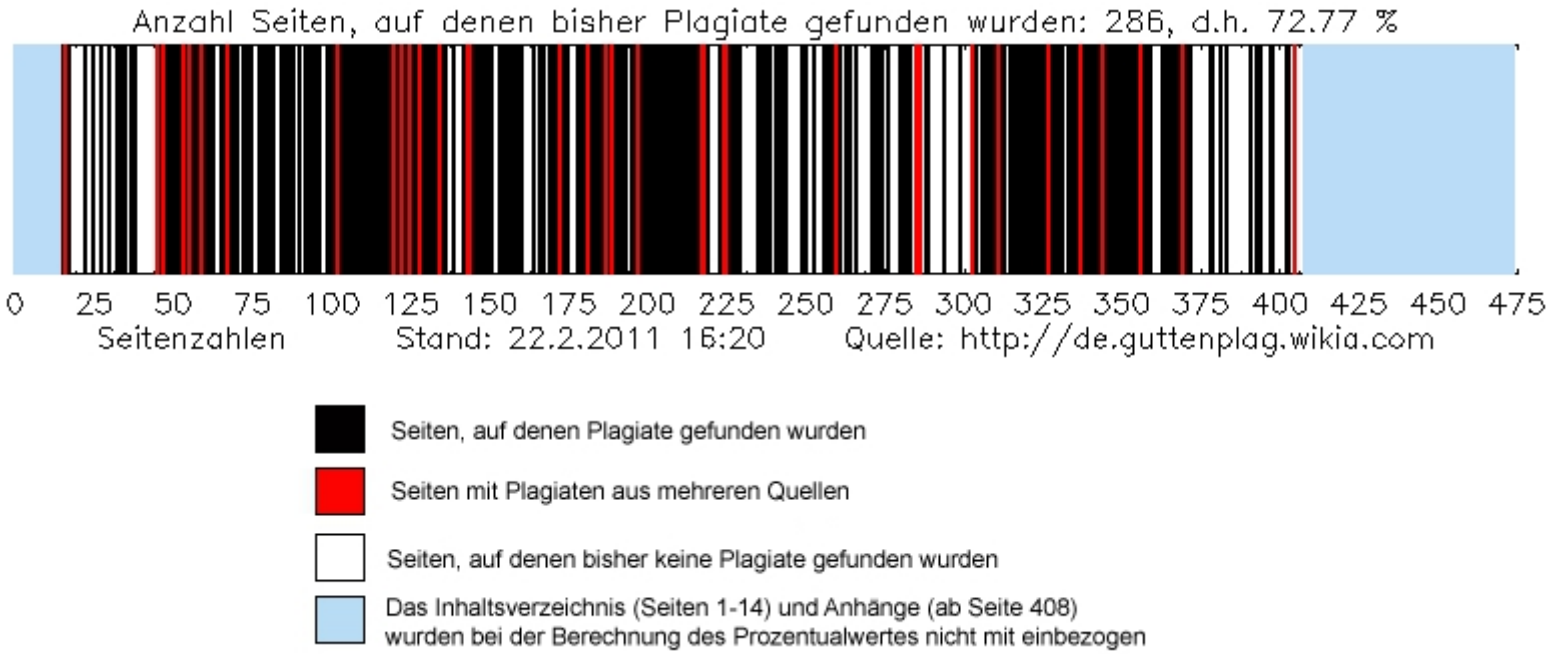
# NLP in the Web – Machine Translation

# NLP in the Web – Speech Recognition

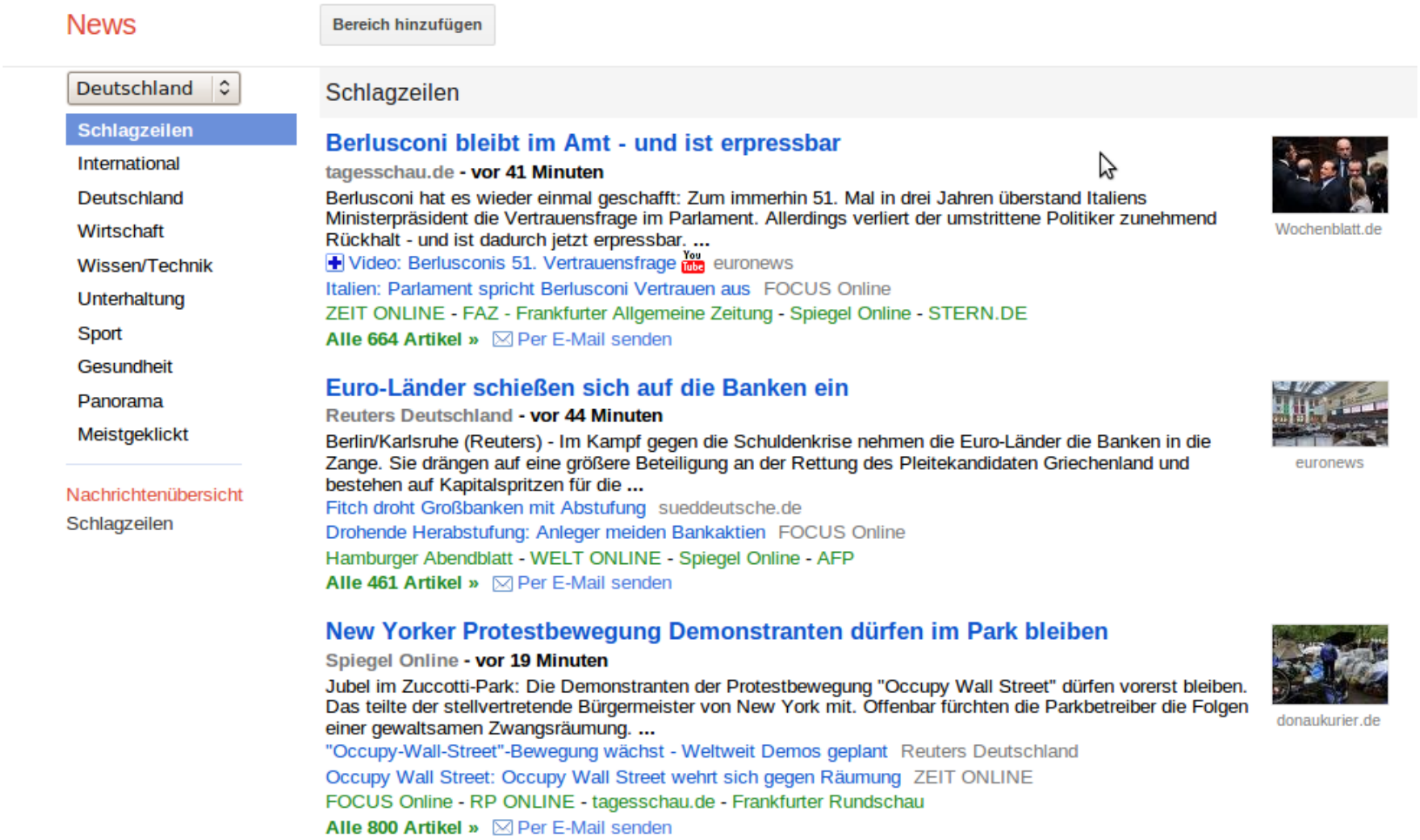


# NLP in the Web – Plagiarism Detection

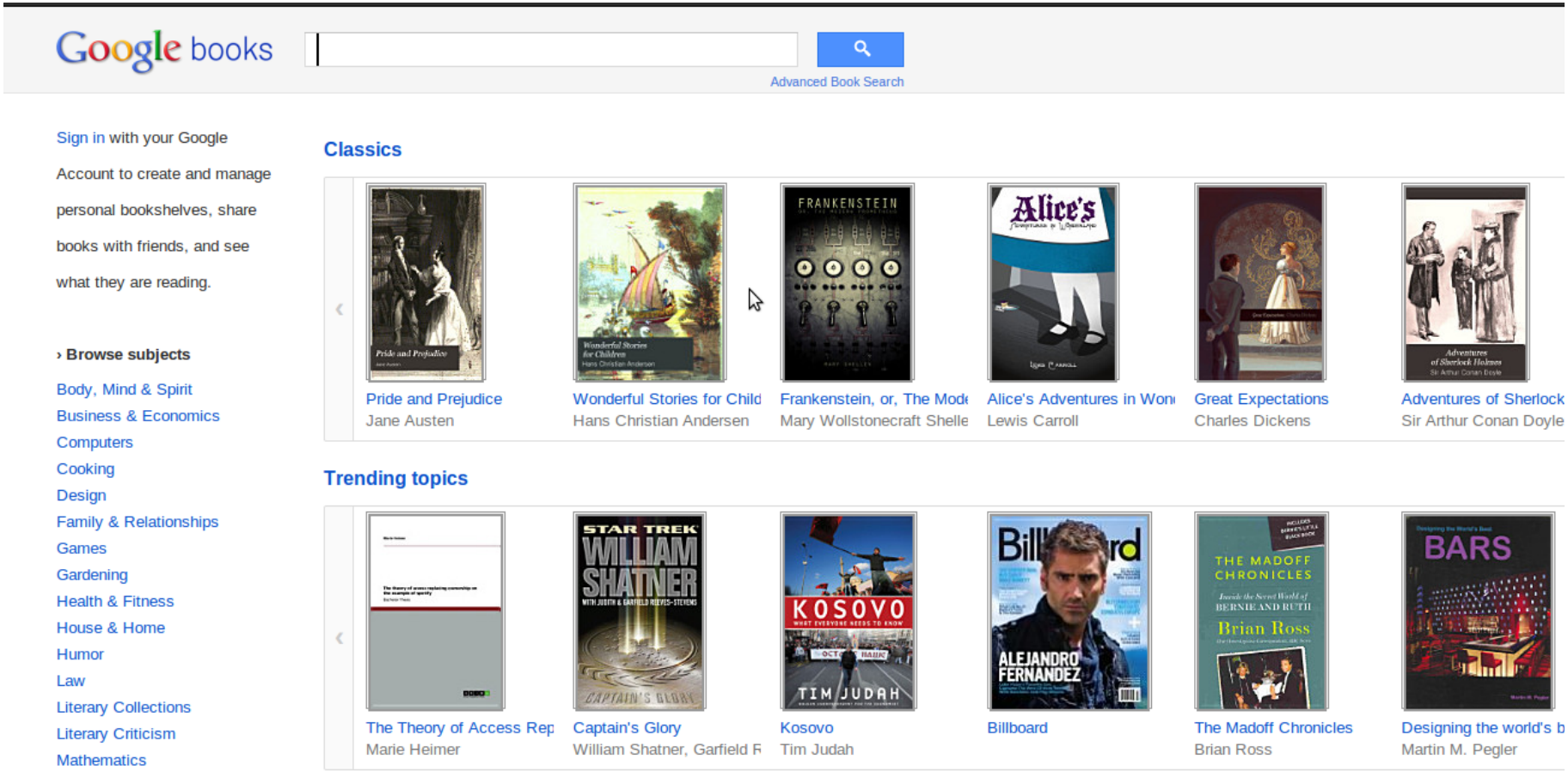
http://de.guttenplag.wikia.com/



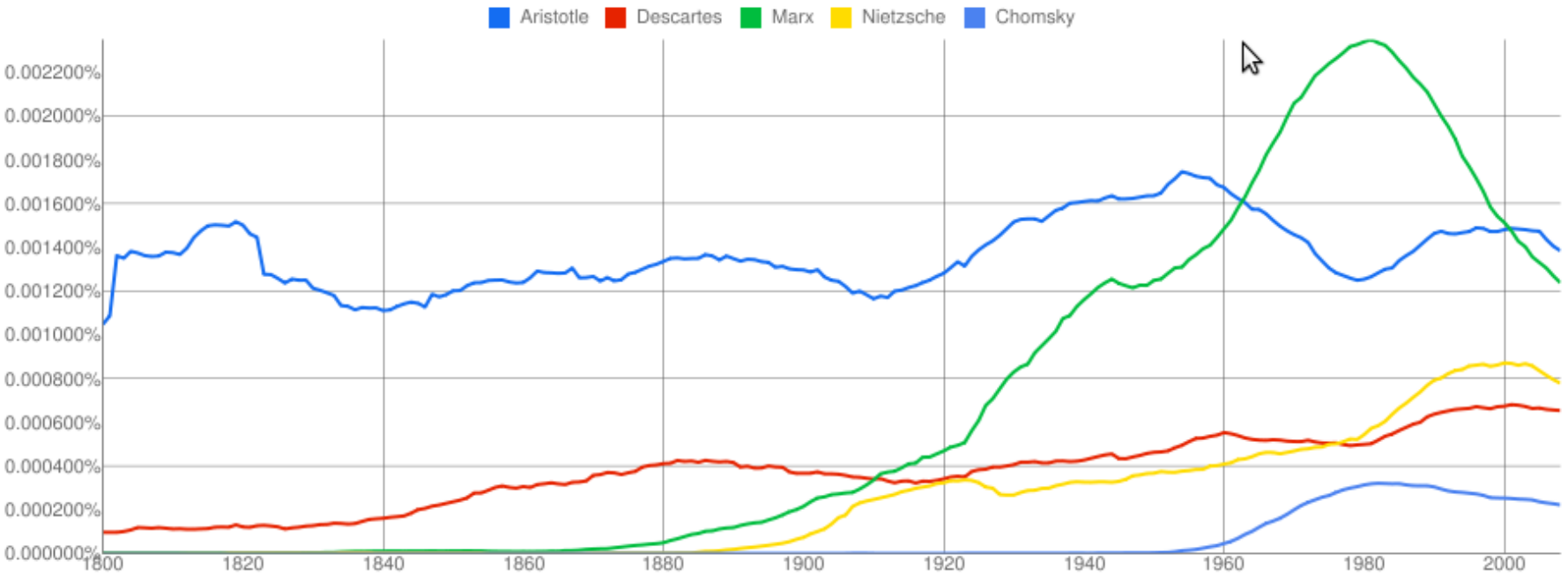
# NLP in the Web – Summarization



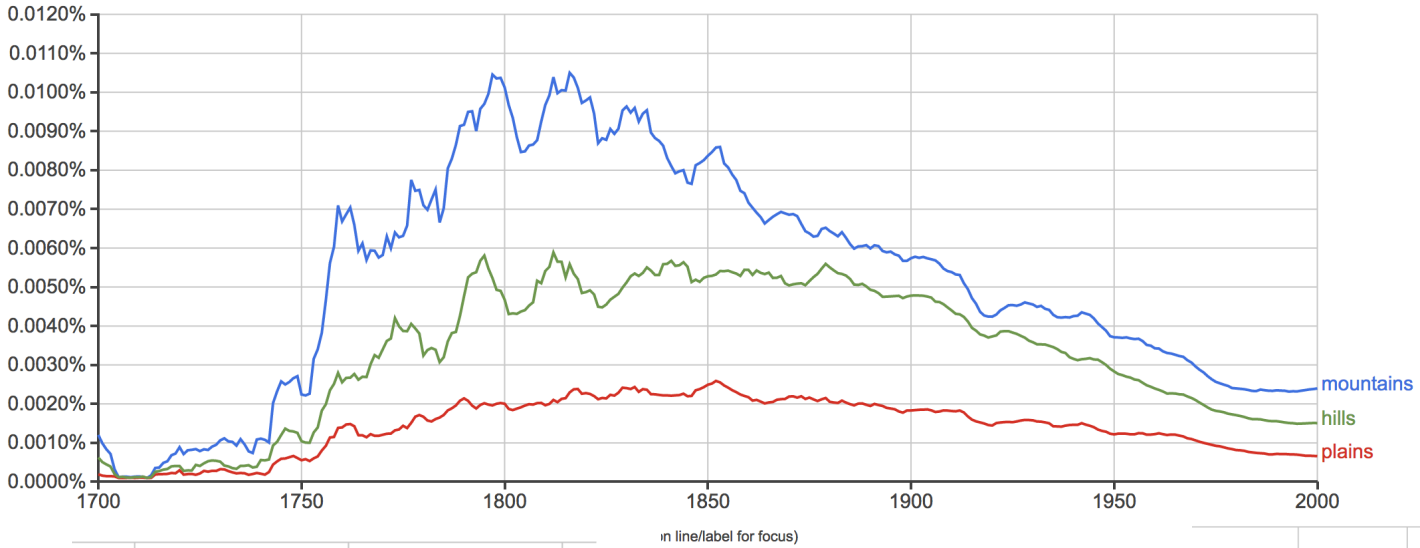
# NLP in the Web – OCR

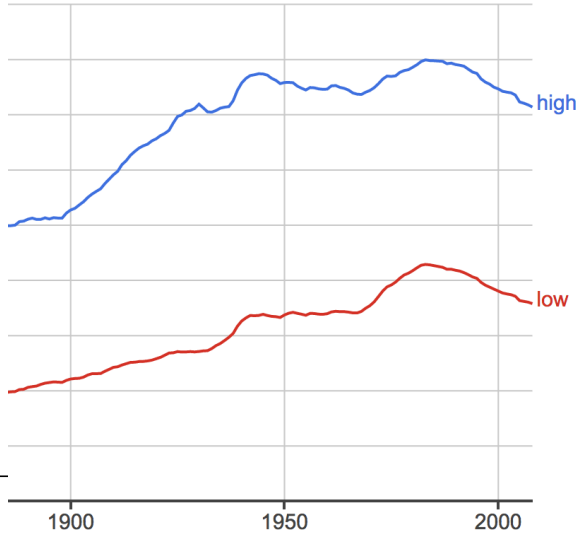


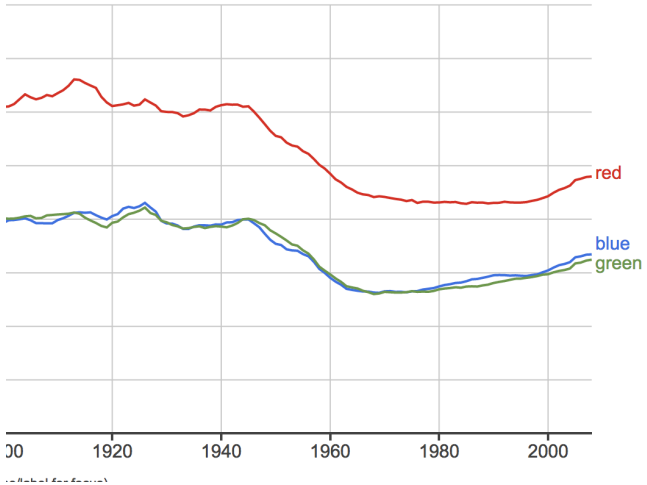
# NLP in the Web – Diachronic Analysis



# Fun With Google n-grams



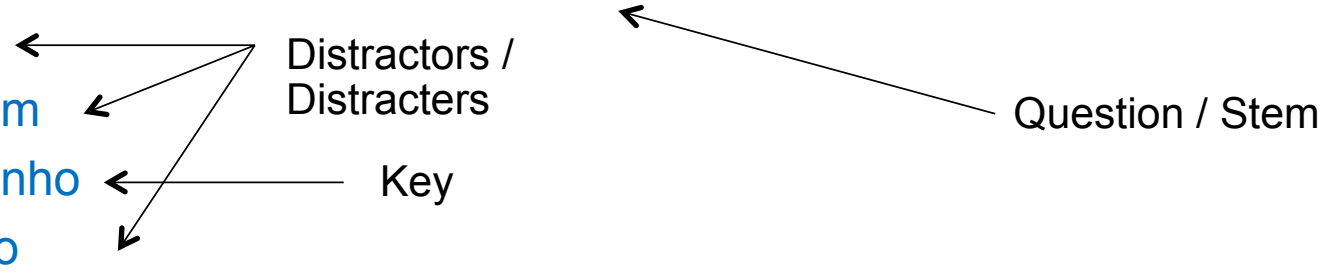



# NLP in the Web – Exercise Generation

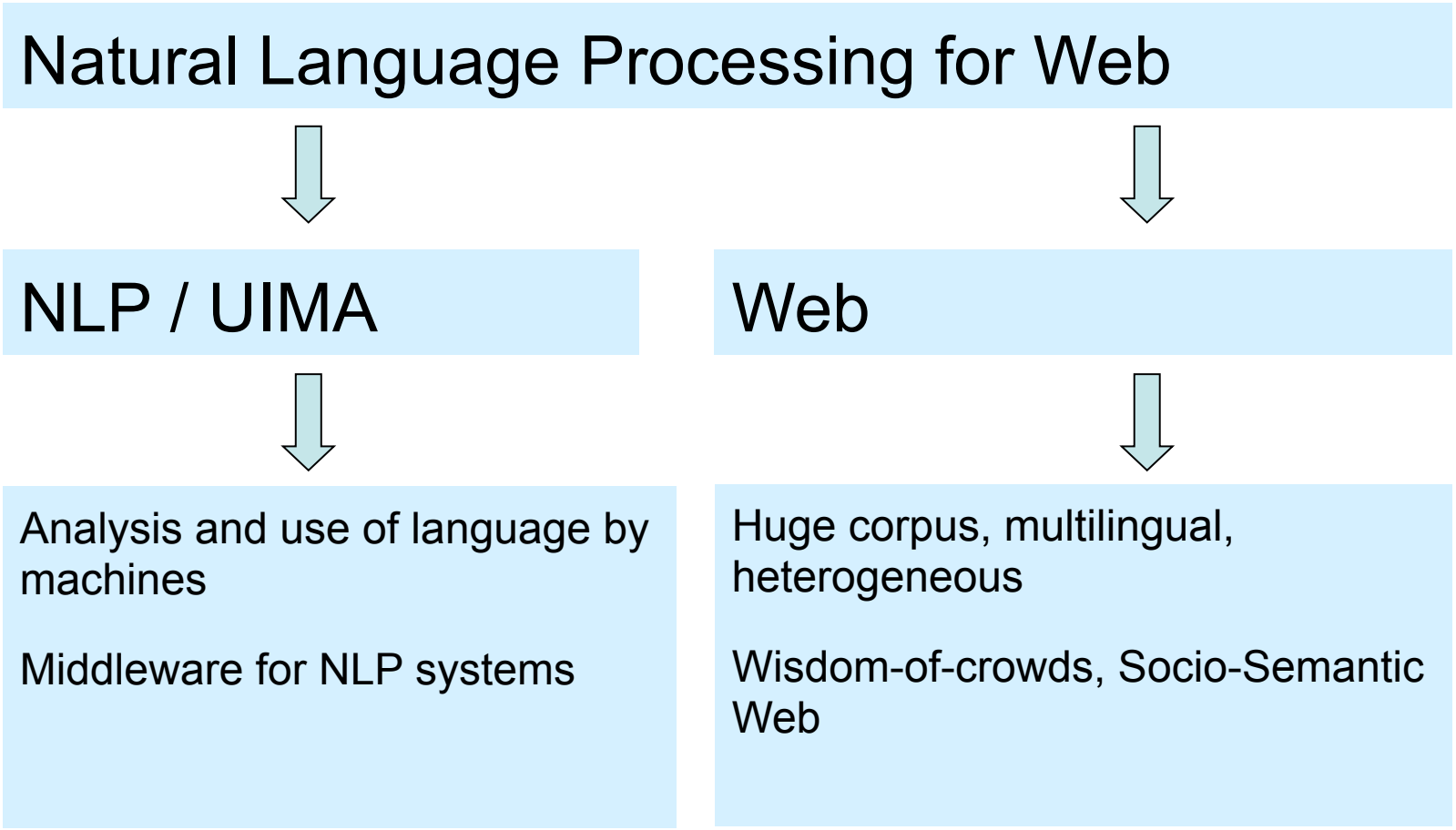
- Choose the correct answer among a set of possible answers

- Example (Mitkov et al., 2006)
- Who was voted the best international footballer for 2004?
  - (a) Henry
  - (b) Beckham
  - (c) Ronaldinho
  - (d) Ronaldo




- Usually 3 to 5 alternative answers

# NLP for the Web



# Natural Language Processing and the Web

- The web is an **application area** for NLP, e.g.:
  - Information retrieval:
    - Search engines
    - Question answering
    - News aggregation
  - Sentiment analysis
- Web is a **resource** to improve the quality of NLP, e.g.:
  - Web as a corpus
  - Analyzing web-based knowledge repositories
    - Wikipedia
    - Wiktionary
  - Recognizing synonyms, paraphrases and the like

# Social Semantic Web

- Social interactions on the Web lead to the creation of **explicit** and **semantically rich** knowledge representations
- Web of collective knowledge systems able to provide useful information based on **human contributions** which get **better** as more people participate
- Combines technologies, strategies and methodologies from the **Semantic Web**, **Social Software** and the **Web 2.0**

# Social Software

- **Definition:**
  - A range of web-based software programs which allow users to interact and share data with other users

- **Examples:**
  - MySpace, Facebook, Flickr, YouTube
  - Amazon.com and eBay (commercial)

- **Properties:**
  - open APIs, service-oriented design, and the ability to upload data and media

# Characteristics of Web 2.0

+ The Wisdom of Crowds
+ Collective intelligence
+ Huge amount of data
+ Fast growing

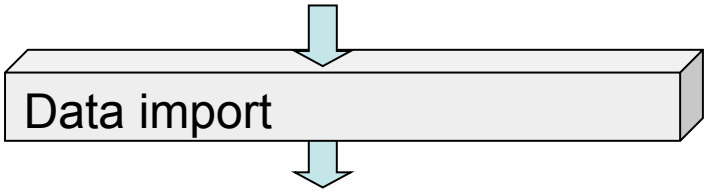
- Noise
- Duplicates
- Content of different quality

# Challenges for NLP

- How to remove noise, e.g. duplicates?
- How to assess the quality of content?
- How to integrate the content of heterogeneous and scattered nature?
- How to deal with errors, e.g. spelling or grammar errors?
- How to „clean“ the data?

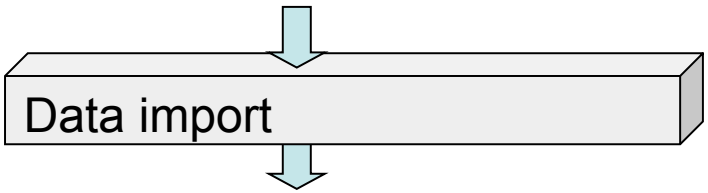
# Scattered and Incoherent Information

- Required information is scattered in heterogeneous semi-structured data sources
  - Forums, mailing lists, blogs, personal home pages, wikis




Forum reader, Blog reader,
Wiki reader, PDF reader, ...

- Data from different languages should be transparently integrated

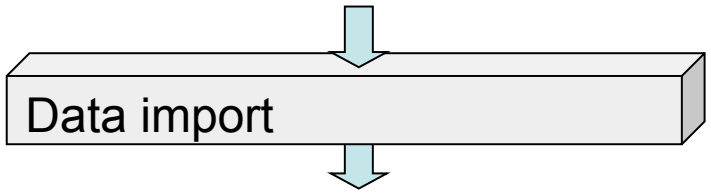



Language Identification

# Data Cleansing is Necessary

- User-generated content contains errors, smileys, abbreviations, etc.

```
Hi
        Micheal,
                              have u seen my
posting,last week u said that u
will look in to my problem thsi week.can i ask u
now?
```




Data cleansing

# Take-Home-Messages

- The web is an application area for NLP, e.g.:
  - Search engines, question answering, community mining
- Web is a resource to improve the quality of NLP, e.g.:
  - Web as a corpus, Wikipedia, Wiktionary
- Challenges for NLP
  - Noise, duplicates, quality of content, heterogeneous sources, …

# Readings for Next Lecture

## Mandatory

- Jurafsky & Martin
  - p.1-5, General introduction
  - p.45-47, Morphology
  - p.68-71, Segmentation
  - p.123-124, PoS-Tagging
  - p.385-392, Syntax & Parsing
  - p.611-619, Lexical Semantics, Word Senses, WordNet

- Mitkov
  - Ch. 10

## Next Lecture

# Linguistic Analysis Levels