

# Lecture 9

## Text Mining

COMP 474/6741, Winter 2021

### Introduction

- Text Mining in Science
- Text Mining Applications
- Language Technology (LT)
- Development Frameworks
- Example GATE Pipeline

### NLP

- Language Models
- Tokenization
- Sentence Splitting
- Morphology
- Part-of-Speech (POS) Tagging
- Chunking and Parsing
- Named Entity Recognition
- Entity Linking
- Pipelines

### Text Mining

- Example: Scientific Literature Mining
- Summary

### Notes and Further Reading

René Witte  
Department of Computer Science  
and Software Engineering  
Concordia University

## 1 Introduction

### Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks  
Example GATE Pipeline

## 2 Natural Language Processing (NLP)

### NLP

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS) Tagging  
Chunking and Parsing  
Named Entity Recognition  
Entity Linking  
Pipelines

## 3 Text Mining

### Text Mining

Example: Scientific Literature Mining  
Summary

## 4 Notes and Further Reading

### Notes and Further Reading

## Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks  
Example GATE Pipeline

## NLP

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS) Tagging  
Chunking and Parsing  
Named Entity Recognition  
Entity Linking  
Pipelines

## Text Mining

Example: Scientific Literature Mining Summary

## Notes and Further Reading

## Slides Credit

- Includes slides from Hoifung Poon, Chris Quirk & Scott Wen-Tau Yih, *Machine Reading for Precision Medicine*, [https://www.microsoft.com/en-us/research/uploads/prod/2018/01/1802\\_aaai-tutorial\\_precision-med.pdf](https://www.microsoft.com/en-us/research/uploads/prod/2018/01/1802_aaai-tutorial_precision-med.pdf)
- Includes slides from Matthew Honnibal & Ines Montani, *An introduction to spaCy*, [https://github.com/explosion/talks/blob/master/2017-08-28\\_spacy-101.pdf](https://github.com/explosion/talks/blob/master/2017-08-28_spacy-101.pdf)

## Too much (textual) information

- We now have electronic books, documents, web pages, emails, blogs, news, chats, memos, research papers, ...
- ... all of it immediately accessible, thanks to databases and Information Retrieval (IR)
- An estimated 80–85% of all data stored in databases are natural language texts
- But humans did not scale so well...

Results in the common perception of **Information Overload** (or even *information rage*)

The screenshot shows a Google search results page. The search bar at the top contains the query "Text Mining". Below the search bar are several navigation links: "All" (highlighted in blue), "Images", "News", "Videos", "Books", "More", "Settings", and "Tools". The main search results area displays a summary message: "About 774,000,000 results (0.44 seconds)". Below this summary are four small informational cards:

- Text Mining**: A diagram showing three overlapping circles labeled "Preprocess", "Index", and "Mine".
- THE TEXT MINING PROCESS**: A flowchart illustrating the steps: "Identify and select the text sources", "Preprocess the text (e.g., remove stop words, stemming)", "Index the text (e.g., create an inverted index)", and "Mine the text (e.g., find patterns, extract knowledge)".
- THE TEXT MINING PROCESS**: Similar to the previous card, it shows the process: "Identify and select the text sources", "Preprocess the text (e.g., remove stop words, stemming)", "Index the text (e.g., create an inverted index)", and "Mine the text (e.g., find patterns, extract knowledge)".
- Text Analytics**: A diagram showing a flow from "structured Data" and "unstructured Data" through "Social Media" and "Persons" to "Text Analytics". It includes a legend for "Content", "Organization", and "Location".

### Introduction

- Text Mining in Science
- Text Mining Applications
- Language Technology (LT)
- Development Frameworks
- Example GATE Pipeline

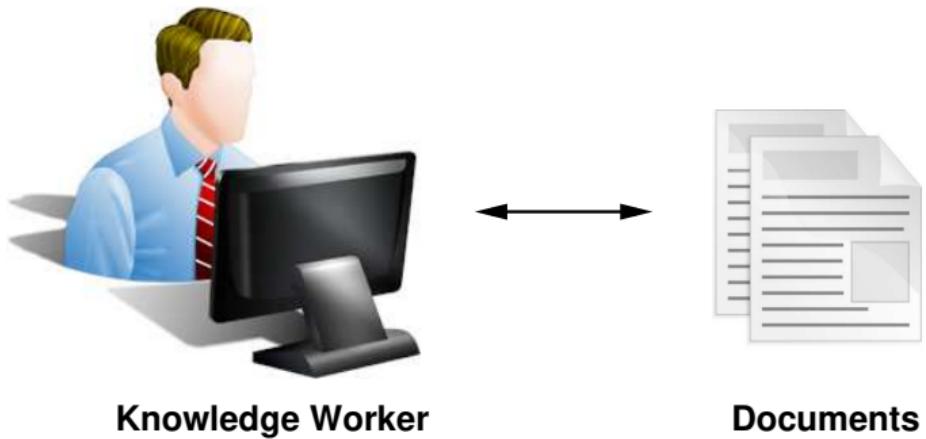
### NLP

- Language Models
- Tokenization
- Sentence Splitting
- Morphology
- Part-of-Speech (POS) Tagging
- Chunking and Parsing
- Named Entity Recognition
- Entity Linking
- Pipelines

### Text Mining

- Example: Scientific Literature Mining
- Summary

### Notes and Further Reading



## Introduction

- Text Mining in Science
- Text Mining Applications
- Language Technology (LT)
- Development Frameworks
- Example GATE Pipeline

## NLP

- Language Models
- Tokenization
- Sentence Splitting
- Morphology
- Part-of-Speech (POS) Tagging
- Chunking and Parsing
- Named Entity Recognition
- Entity Linking
- Pipelines

## Text Mining

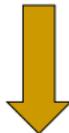
- Example: Scientific Literature Mining
- Summary

## Notes and Further Reading

## Example: Tumor Board KB Curation

The deletion mutation on exon-19 of **EGFR** gene was present in 16 patients, while the **L858E** point mutation on exon-21 was noted in 10.

All patients were treated with **gefitinib** and showed a partial response.



Gefitinib can treat tumors w. EGFR-L858E mutation

# PubMed

27 million abstracts

Two new abstracts every minute

Adds over one million every year



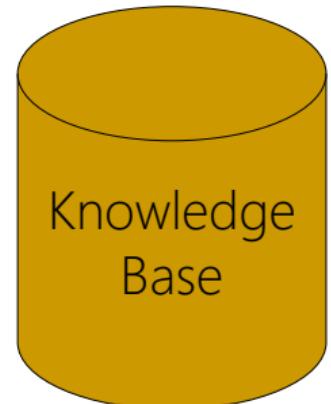
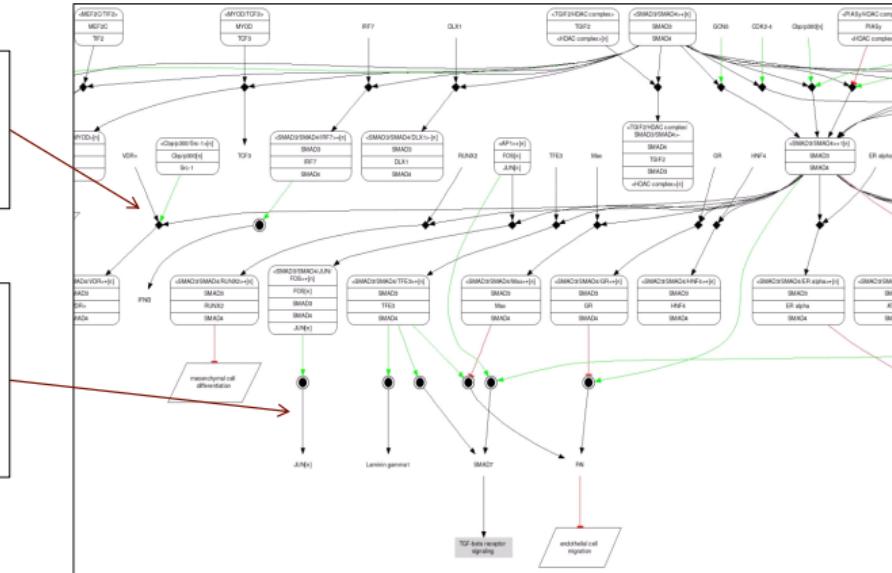
# Machine Reading

PMID: 123

...  
VDR+ binds to  
SMAD3 to form  
...  
...

PMID: 456

...  
JUN expression  
is induced by  
SMAD3/4  
...  
...



## Main Menu

- [Top Stories](#)
- [24 Hours Overview](#)
- [Events Detection](#)
- [Most Active Themes](#)
- [Help about EMM](#)
- [Overview](#)
- [Advanced Search](#)
- [Sources list](#)
- [Web Site Map](#)

## EU Focus

## EU Policy Areas

Agriculture and Rural Development

Better Regulation, Inter-Institutional Relations, the Rule of Law and the Charter of Fundamental Rights

Budget and Human Resources

Climate Action and Energy

Competition

Digital Economy and Society

Economic and Financial Affairs, Taxation and Customs

Education, Culture, Youth and Sport

Environment, Maritime Affairs and Fisheries

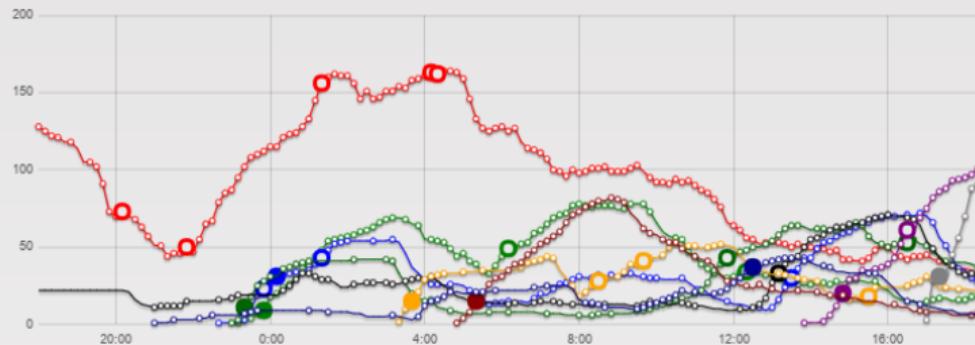
EU Foreign Affairs and Security Policy

DG Neighbourhood and Enlargement Negotiations

Financial Stability, Financial

## Biggest 10 Stories Over Last 24h

Language: en Period: Nov 6, 2018 12:20 AM - Nov 7, 2018 12:20 AM



### Oil prices drop over one percent on Iran sanctions waivers

Articles : 1527 | Last update : Nov 6, 2018 11:22:00 PM | Start : Nov 4, 2018 10:53:00 PM | Sources : 352

#### Oil market cautious as US punitive sanctions on Iran become effective

 mercopress Tuesday, November 6, 2018 10:03:00 PM CET | info [other]

Oil prices were mixed on Monday after a steep five-day fall, as the United States formally imposed punitive sanctions on Iran but granted eight countries temporary waivers allowing them to keep buying oil from the Islamic Republic. The sanctions are part of U.S....

[More articles...](#)

### Facebook blocks 115 accounts ahead of US midterm elections

Articles : 211 | Last update : Nov 7, 2018 12:18:00 AM | Start : Nov 6, 2018 5:12:00 AM | Sources : 165

#### Facebook blocks 115 accounts ahead of US midterm elections

 ohio Tuesday, November 6, 2018 9:00:00 PM CET | info [other]

Facebook, Twitter and other companies have been fighting misinformation and election meddling on their services for the past two years. LONDON Facebook said it blocked 115 accounts for suspected "coordinated inauthentic behavior" linked to foreign groups attempting to interfere in Tuesday's U.S. midterm elections....

## Tools

Wednesday, November 7, 2018  
12:48:00 AM CET

[RSS | MAP](#)

[Facebook](#)

[manage](#)

[info](#)



## Languages

Select top stories in other languages.

ar	bg	cs	da	de	el
en	es	et	fi	fr	hr
hu	it	lt	lv	mt	nl
pl	pt	ro	ru	sk	sl
sv	sw	tr	zh		

[Show additional languages](#)

Interface: en - English

[Legend](#)

## Country Watch

The country most in the news at the moment.



## Main Menu

- [Top Stories](#)
- [24 Hours Overview](#)
- [Events Detection](#)
- [Most Active Themes](#)
- [Help about EMM](#)
- [Overview](#)
- [Advanced Search](#)
- [Sources list](#)
- [Web Site Map](#)

## EU Focus

- [President Ursula von der Leyen](#)
- [Commission Vice-Presidents](#)
- [Commissioners](#)
- [EC News](#)
- [EP Presidency - David Sassoli](#)
- [Council President](#)
- [Portuguese Presidency of the Council of the EU](#)

## EU Policy Areas

- [Agriculture and Rural Development](#)

# Anthony Fauci

Last updated on 2019-10-29T05:06+0100.

**ABOUT THIS IMAGE****LICENSE:** PUBLIC**AUTHOR:** WHITE HOUSE PHOTO BY**SHEALAH CRAIGHEAD**[Key Titles and Phrases \(Last 30\)](#)[Names \(Top 30\)](#)

NAMES	LANG	COUNT
-------	------	-------

Anthony Fauci	EN	60.64%
Anthony Fauci	ES	11.75%
Anthony Fauci	FR	8.08%

## Tools

Tuesday, March 23, 2021  
5:30:00 PM CET

[Facebook](#)[manage](#)

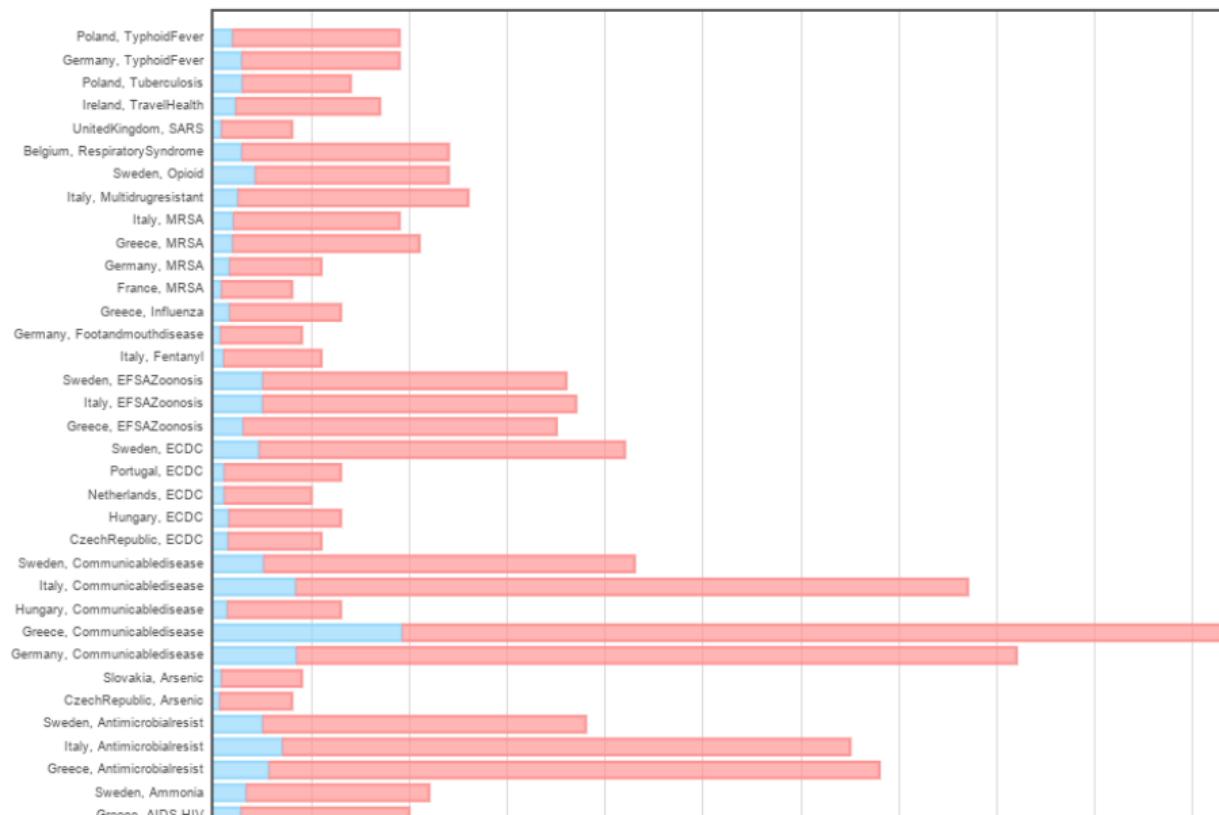
## Languages

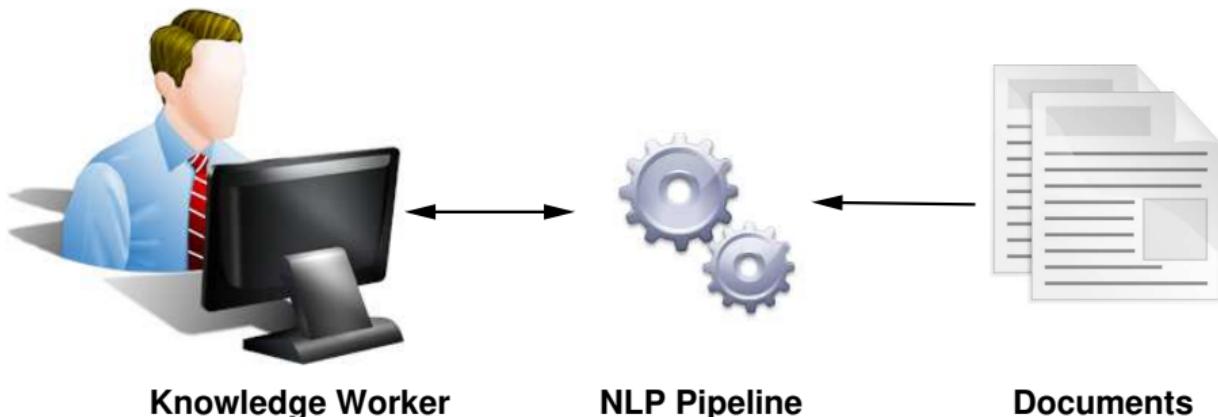
Select your languages

ab	ar	az	be	bg	bs
ca	cs	da	de	el	en
eo	es	et	fa	fi	fr
ga	gl	ha	he	hi	hr
hu	hy	id	is	it	ja
ka	km	ko	ku	ky	lb
lo	lt	lv	mk	ml	ms
mt	my	ne	nl	no	os
pap	pl	ps	pt	ro	ru
si	sk	sl	so	sq	sr
sv	sw	ta	th	tr	uk
ur	vi	zh			
all					

**Top Stories**
[Event Extraction](#)
[Recent Disease Incidents](#)
[Alert Statistics >](#)
[Communicable Diseases >](#)
[Symptoms >](#)
[Bioterrorism >](#)
[Nuclear >](#)
[Chemical >](#)
[ECDC >](#)
[EFSA >](#)
[EMCDDA >](#)
[ENV\\_RISKS >](#)
[Food Security >](#)
[SAM >](#)
[Medical Devices >](#)
[Vaccination >](#)
[Other >](#)
[Continents >](#)
[Official Sources >](#)
[Sources List](#)

## Today's Alert Statistics for European Union





→ Worksheet #8: Task 1

## Introduction

Text Mining in Science

## Text Mining Applications

Language Technology (LT)

Development Frameworks

Example GATE Pipeline

## NLP

Language Models

Tokenization

Sentence Splitting

Morphology

Part-of-Speech (POS) Tagging

Chunking and Parsing

Named Entity Recognition

Entity Linking

Pipelines

## Text Mining

Example: Scientific Literature Mining

Summary

## Notes and Further Reading



AI & MACHINE LEARNING

## Four ways to apply NLP in financial services

Apr 28, 2020  6 min



Jo Stichbury

Freelance Technical Writer



Natural language processing (NLP) is increasingly used to review unstructured content or spot trends in markets. How is Refinitiv Labs applying NLP in financial services to meet challenges around investment decision-making and risk management?

### Subscribe to Newsletter

 Your email address Select a country/territory I would like to receive the Refinitiv Perspectives newsletter.

By submitting your details, you are agreeing to receive communications about Refinitiv resources, events, products, or services. You also acknowledge that you have read and understood our [privacy statement](#).

## Solutions

### Refinitiv Labs

Refinitiv™ Labs collaborate with customers around the world to solve big problems and rapidly prototype and validate solutions using data science and lean techniques

# So you want to build a Text Mining system... .

René Witte



## Requirements

An NLP system requires a large amount of infrastructure work:

- Document handling, in various formats (plain text, HTML, XML, PDF, ...), from various sources (files, DBs, email, ...)
- Annotation handling (stand-off markup)
- Component implementations for standard tasks, like Tokenizers, Sentence Splitters, Part-of-Speech (POS) Taggers, Finite-State Transducers, Full Parsers, Classifiers, Noun Phrase Chunkers, Lemmatizers, Entity Taggers, Coreference Resolution Engines, Summarizers, ...

As well as *resources* for concrete tasks and languages:

- Lexicons, WordNets
- Grammar files and Language models
- Machine Learning Algorithms
- Evaluation Metrics
- etc.

### Introduction

[Text Mining in Science](#)

[Text Mining Applications](#)

### Language Technology (LT)

[Development Frameworks](#)

[Example GATE Pipeline](#)

### NLP

[Language Models](#)

[Tokenization](#)

[Sentence Splitting](#)

[Morphology](#)

[Part-of-Speech \(POS\) Tagging](#)

[Chunking and Parsing](#)

[Named Entity Recognition](#)

[Entity Linking](#)

[Pipelines](#)

### Text Mining

[Example: Scientific Literature Mining](#)

[Summary](#)

### Notes and Further Reading

## Introduction

[Text Mining in Science](#)

[Text Mining Applications](#)

## Language Technology (LT)

[Development Frameworks](#)

[Example GATE Pipeline](#)

## NLP

[Language Models](#)

[Tokenization](#)

[Sentence Splitting](#)

[Morphology](#)

[Part-of-Speech \(POS\) Tagging](#)

[Chunking and Parsing](#)

[Named Entity Recognition](#)

[Entity Linking](#)

[Pipelines](#)

## Text Mining

[Example: Scientific Literature Mining](#)

[Summary](#)

## Notes and Further Reading

## Fortunately, you don't have to start from scratch

Many (open source) tools and resources are available:

**NLP Tools:** programs performing a single task, like classifiers, parsers, or NP chunkers

**Frameworks:** integration architectures for combining and controlling all components and resources of an NLP system

**Resources:** for various languages, like lexicons, wordnets, or grammars

## Major Frameworks

Two important frameworks are:

- GATE (*General Architecture for Text Engineering*), under development since 1995 at University of Sheffield, UK
- UIMA (*Unstructured Information Management Architecture*), developed by IBM; open-sourced in 2007 (Apache project)

Both frameworks are open source (GATE: LGPL, UIMA: Apache)

## Libraries

- Numerous NLP libraries: NLTK (Python), Stanford CoreNLP, OpenNLP...
- Various integrations (e.g., CoreNLP has GATE wrapper, Python bindings)

## Current Trends

- Increasing use of Deep Learning tools/frameworks for NLP
- Keras, TensorFlow, PyTorch etc.

### Introduction

Text Mining in Science

Text Mining Applications

Language Technology (LT)

### Development Frameworks

Example GATE Pipeline

### NLP

Language Models

Tokenization

Sentence Splitting

Morphology

Part-of-Speech (POS) Tagging

Chunking and Parsing

Named Entity Recognition

Entity Linking

Pipelines

### Text Mining

Example: Scientific

Literature Mining

Summary

### Notes and Further Reading

# Unstructured Information Management Architecture (UIMA)

René Witte



UIIMA Ruta CDE - CDE/data/features/kdml12.pdfbox.txt.xmi (CDEdescriptor/uima/ruta/example/CDETypeSystem.xml) - Eclipse Platform

File Edit Navigate Search Project Run Window Help

UIIMA Ruta C... Team Synchron... UIIMA Ruta E... UIIMA Ruta

Script Explorer

CDE

- script
  - uima.ruta.example
  - AddReferences.ruta
  - CDE.ruta
  - Convert.ruta
  - Features.ruta
  - RemovePlainTextSt...
  - Test.ruta
  - UnmarkAllBut.ruta
- data
- descriptor
- input
- output
- resources
- test
- Interpreter Libraries [UIIMA]
- DKProExample
- ExampleProject [uima/sandbox]
- ExtensionsExample [uima/sanc...
- MedicalReports
- ruta-example-dkpro
- ruta-example-sandbox
- SandboxExample
- Tests
- TextRulerExample [uima/sandb...
- TM-Gutenberg [code/ruta/TM]

kFML12.pdfbox.txt.xmi

Novi Quadrianto, Alex J. Smola, Tibrío S. Caetano, and Quoc V. Le. **Estimating label proportions from label proportions.** *Journal of Machine Learning Research*, 10:2349–2374, Oct 2009.

Stefan Rüping. **A simple method for estimating conditional probabilities in SVMs.** In A. Abecker, S. Bickel, U. Brefeld, I. Drost, N. Herold, M. Minor, T. Scheffer, L. Stojanovic, and S. Weibel, editors, *LWA 2004 - Lernen - Wissensentdeckung - Adaptivitäät*. Humboldt-Universität zu Berlin, 2004.

S. Tong and D. Koller. **Restricted bayes optimal classifiers.** In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI 2000)*.

V. Vapnik. **Statistical Learning Theory.** Wiley, Chichester, GB, 1998.

Bianca Zadrozny and Charles Elkan. **Transforming classifier scores into accurate multiclass probability estimates.** In *Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining*, 2002.

CDE Cons

Selection TextRuler Annotation Ruta Query

Constraint	Weight
Reference(OR(STARTSWITH(Author), STARTSWITH(Editor)));	1
Author,-CONTAINS(NUM);	1
Author (Date   Title);	1
Author(CONTAINS(CW,1,100));	1
Author(CONTAINS(W,2,200));	1
Author,-CONTAINS(EditorMarker);	1
Author(STARTSWITH(Reference));	1

CDE Document

Documents: t-textmarker\CDE\data\features

Test Data: ace-textmarker\CDE\data\gold\_author

Type System: ma/ruta/example/CDETypeSystem.xml

mse=8.0E-4 spearmans=0.6932 pearsons=0.7373 cosine=0.9997

Document	CDE	F1
kdml12.pdfbox.txt.xmi	0.952	0.8936
A97-1010.txt.xmi	0.958	0.9371
mlmd_2_2_80-99.pdfbox.t...	0.9657	0.9444
A00-2002.txt.xmi	0.978	0.9474
A88-1009.txt.xmi	0.987	0.9636
A94-1026.txt.xmi	0.9881	1.0
J05-4002.txt.xmi	0.9881	0.9571
C02-1020.txt.xmi	0.9898	0.9048
J05-2005.txt.xmi	0.9907	0.9664
mlmd_2_1_3-22.pdfbox.tx...	0.994	0.9782
1471-2105-12-36.pdfbox.t...	0.9947	0.9923
J05-1003.txt.xmi	0.9947	0.9875
1471-2105-12-43.pdfbox.t...	0.997	0.9934
1471-2105-12-37.pdfbox.t...	1.0	1.0
A00-1042.txt.xmi	1.0	1.0
C02-1035.txt.xmi	1.0	1.0
C04-1024.txt.xmi	1.0	1.0

CDE Result

Constraint	Result
Reference(OR(STARTSWITH(Author), STARTSWITH(Editor)));	0.846153846153846
Author,-CONTAINS(NUM);	1.0
Author (Date   Title);	0.909090909090909
Author(CONTAINS(CW,1,100));	1.0
Author(CONTAINS(W,2,200));	0.909090909090909
Author,-CONTAINS(EditorMarker);	1.0
Author(STARTSWITH(Reference));	1.0

## Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
**Development Frameworks**

## NLP

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS) Tagging  
Chunking and Parsing  
Named Entity Recognition  
Entity Linking  
Pipelines

## Text Mining

Example: Scientific Literature Mining  
Summary

## Notes and Further Reading

# General Architecture for Text Engineering (GATE)

René Witte

The screenshot shows the GATE Developer 6.1-snapshot build 3809 interface. The left sidebar contains a tree view of applications, language resources, processing resources, and datastores. The main window displays a document titled "26eval.xml\_0002..." with annotations. A central panel shows an annotation set for a "Person" entity, listing attributes like gender (male), matches ([9653, 9656]), rule (PersonFinal), and rule1 (PersonTitle). Below this is an "Open Search & Annotate tool". The right side features a sidebar with various development frameworks and a notes section.

**GATE**

**File Options Tools Help**

**Applications**

- ANNIE**

**Language Resources**

- Corpus for 26eval.xml\_0002E**
- 26eval.xml\_0002E**

**Processing Resources**

- ANNIE OrthoMatcher**
- ANNIE NE Transducer**
- ANNIE POS Tagger**
- ANNIE Sentence Splitter**
- ANNIE Gazetteer**
- ANNIE English Tokeniser**
- Document Reset PR**

**Datastores**

**Annotations**

**Annotations List**

**Annotations Stack**

**Co-reference Editor**

**OAT**

**RAT-C RAT-I Text**

**Messages ANNIE**

Chief among these is retired Air Force partner, **Albert Hakim**, has had a long-standing principal, attorney **Willard Zucker**. CSF, which is at the heart of the private network, was assisting the Contras and later to implement yesterday.

Within the **White House**, former **NSC** aide operations. At the same time, Col. North maintained encouragement, were raising money for

Tax deductible contributions were collected by the **Preservation of Liberty** and funneled through to Col. North and CSF, according to an interviewee.

**Mr. Channell**, who established the endowment in part to help **Mr. Reagan's** Central America policies, has repeatedly said he hadn't any intention to provide military aid. But this flow of money to Switzerland coincides roughly with the transfers back from **Europe to Panama** and **ACE** and raises the question of whether private funds were used.

**Person**

gender	male
matches	[9653, 9656]
rule	PersonFinal
rule1	PersonTitle

**Open Search & Annotate tool**

**Context**

red Air Force Maj. Gen. Richard Secord, whose business partner, Albert

**MatchesAnnots**: [null=([9603, 9604], [9605, 9606])]

**MimeType**: text/xml

**gate.SourceURL**: file:/home/kresten/GATE/26eval.xml\_0002E

**Document Editor**

**Initialisation Parameters**

ANNIE run in 2.892 seconds

**Development Frameworks**

- Date
- FirstPerson
- JobTitle
- Location
- Lookup
- Money
- Organization
- Person
- Sentence
- SpaceToken
- Split
- Temp
- TempDate
- Title
- Token
- Unknown

**NLP**

- Language Models
- Tokenization
- Sentence Splitting
- Morphology
- Part-of-Speech (POS) Tagging
- Chunking and Parsing
- Named Entity Recognition
- Entity Linking
- Pipelines

**Text Mining**

- Example: Scientific Literature Mining
- Summary

**Notes and Further Reading**



## Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks

## Example GATE Pipeline

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS) Tagging  
Chunking and Parsing  
Named Entity Recognition  
Entity Linking  
Pipelines

## Text Mining

Example: Scientific Literature Mining  
Summary

## Notes and Further Reading

# NLP Pipeline in GATE

René Witte



Messages ANNIE

Loaded Processing resources

Name	Type
LODeXporter 00018	LODeXporter

Selected Processing resources

!	Name	Type
Document Reset PR	Document Reset PR	
ANNIE English Tokeniser	ANNIE English Tokeniser	
ANNIE Gazetteer	ANNIE Gazetteer	
ANNIE Sentence Splitter	ANNIE Sentence Splitter	
ANNIE POS Tagger	ANNIE POS Tagger	
NE ANNIE NE Transducer	ANNIE NE Transducer	
ANNIE OrthoMatcher	ANNIE OrthoMatcher	

Run "ANNIE POS Tagger"?

Yes  No  If value of feature  [ ] is [ ]

Corpus: Corpus for GATE Document\_00013

Runtime Parameters for the "ANNIE POS Tagger" ANNIE POS Tagger:

Name	Type	Required	Value
baseSentenceAnnotationType	String	✓	Sentence
baseTokenAnnotationType	String	✓	Token
failOnMissingInputAnnotations	Boolean		true
inputASName	String		
outputASName	String		
outputAnnotationType	String	✓	Token
posTagAllTokens	Boolean		true

Run this Application

Serial Application Editor Initialisation Parameters About...

## Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks

## Example GATE Pipeline

### NLP

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS) Tagging  
Chunking and Parsing  
Named Entity Recognition  
Entity Linking  
Pipelines

### Text Mining

Example: Scientific Literature Mining Summary

### Notes and Further Reading

## Example Tokenisation Rules

```
#numbers#
// a number is any combination of digits
"DECIMAL_DIGIT_NUMBER"+ >Token;kind=number;

#whitespace#
(SPACE_SEPARATOR) >SpaceToken;kind=space;
(CONTROL) >SpaceToken;kind=control;
```

## Example Output

Type	Set	Start	End	Features
Token		158	163	{kind=word, length=5, orth=lowercase, string=years}
SpaceToken		163	164	{kind=space, length=1, string= }
Token		164	167	{kind=word, length=3, orth=lowercase, string=ago}
Token		167	168	{kind=punctuation, length=1, string=,}
SpaceToken		168	169	{kind=space, length=1, string= }
Token		169	180	{kind=word, length=11, orth=lowercase, string=researchers}
SpaceToken		180	181	{kind=space, length=1, string= }

1417 Annotations (0 selected)

[Introduction](#)[Text Mining in Science](#)[Text Mining Applications](#)[Language Technology \(LT\)](#)[Development Frameworks](#)[Example GATE Pipeline](#)[NLP](#)[Language Models](#)[Tokenization](#)[Sentence Splitting](#)[Morphology](#)[Part-of-Speech \(POS\) Tagging](#)[Chunking and Parsing](#)[Named Entity Recognition](#)[Entity Linking](#)[Pipelines](#)[Text Mining](#)[Example: Scientific Literature Mining](#)[Summary](#)[Notes and Further Reading](#)

## Producing POS Annotations

POS-Tagging assigns a part-of-speech-tag (POS tag) to each Token.

- GATE comes with the Hepple tagger for English, which is a modified version of the Brill tagger

### Example output

ILL., THE OWNER OF NEW YORK-based Loews Corp. that makes Kent cigarettes, stopped using crocidolite in its Micronite cigarette filters in 1956.

Although preliminary findings were reported more than a year ago, the latest results appear in today's New England Journal of

Type	Set	Start	End	Features
Token		485	494	{category=NN, kind=word, length=9, orth=upperInit}
Token		495	504	{category=NN, kind=word, length=9, orth=lowercase}
Token		505	512	{category=NNS, kind=word, length=7, orth=lowercase}
Token		513	515	{category=IN, kind=word, length=2, orth=lowercase}
Token		516	520	{category=CD, kind=number, length=4, string=1956}
Token		520	521	{category=., kind=punctuation, length=1, string=.}
Token		522	521	{category=IN, kind=word, length=8, orth=upperInit}

### Introduction

- Text Mining in Science
- Text Mining Applications
- Language Technology (LT)
- Development Frameworks

### Example GATE Pipeline

#### NLP

- Language Models
- Tokenization
- Sentence Splitting
- Morphology
- Part-of-Speech (POS) Tagging
- Chunking and Parsing
- Named Entity Recognition
- Entity Linking
- Pipelines

#### Text Mining

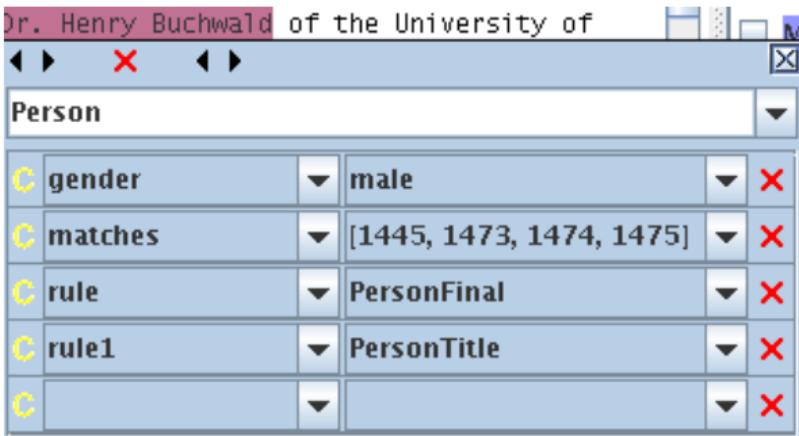
- Example: Scientific Literature Mining
- Summary

#### Notes and Further Reading

## Transducer-based NE Detection

Using all the information obtained in the previous steps (Tokens, Gazetteer lookups, POS tags), ANNIE now runs a sequence of JAPE-Transducers to detect Named Entities (NE)s.

### Example for a detected Person



The screenshot shows the ANNIE interface with a sentence "Dr. Henry Buchwald of the University of" at the top. Below it is a detailed view of the detected entity "Person". The entity has five attributes:

C	gender	male	X
C	matches	[1445, 1473, 1474, 1475]	X
C	rule	PersonFinal	X
C	rule1	PersonTitle	X
C			X

We can now look at the grammar rules that found this person.

[Introduction](#)[Text Mining in Science](#)[Text Mining Applications](#)[Language Technology \(LT\)](#)[Development Frameworks](#)[Example GATE Pipeline](#)[NLP](#)[Language Models](#)[Tokenization](#)[Sentence Splitting](#)[Morphology](#)[Part-of-Speech \(POS\)](#)[Tagging](#)[Chunking and Parsing](#)[Named Entity Recognition](#)[Entity Linking](#)[Pipelines](#)[Text Mining](#)[Example: Scientific](#)[Literature Mining](#)[Summary](#)[Notes and Further](#)[Reading](#)

## Strategy

A JAPE grammar rule combines information obtained from POS-tags with Gazetteer lookup information

- although the last name in the example is not in any list, it can be found based on its POS tag and an additional first name/last name rule (not shown)
- many additional rules for other Person patterns, as well as Organizations, Dates, Addresses, ...

## Persons with Titles

```
Rule: PersonTitle
Priority: 35
(
  {Token.category == DT} |
  {Token.category == PRP} |
  {Token.category == RB}
)?
(
  (TITLE) +
  ((FIRSTNAME | FIRSTNAMEAMBIG
    | INITIALS2)
)?
  (PREFIX) *
  (UPPER)
  (PERSONENDING) ?
)
:person --> ...
```

### Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks

### Example GATE Pipeline

### NLP

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS) Tagging  
Chunking and Parsing  
Named Entity Recognition  
Entity Linking  
Pipelines

### Text Mining

Example: Scientific Literature Mining  
Summary

### Notes and Further Reading

→ Worksheet #8: Task 2

## 1 Introduction

### Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks  
Example GATE Pipeline

## 2 Natural Language Processing (NLP)

### NLP

- Language Models
- Tokenization
- Sentence Splitting
- Morphology
- Part-of-Speech (POS) Tagging
- Chunking and Parsing
- Named Entity Recognition
- Entity Linking
- Pipelines

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS) Tagging  
Chunking and Parsing  
Named Entity Recognition  
Entity Linking  
Pipelines

### Text Mining

Example: Scientific Literature Mining  
Summary

## 3 Text Mining

### Notes and Further Reading

## 4 Notes and Further Reading

Search	Web	Documents	Autocomplete
Editing	Spelling	Grammar	Style
Dialog	Chatbot	Assistant	Scheduling
Writing	Index	Concordance	Table of contents
Email	Spam filter	Classification	Prioritization
Text mining	Summarization	Knowledge extraction	Medical diagnoses
Law	Legal inference	Precedent search	Subpoena classification
News	Event detection	Fact checking	Headline composition
Attribution	Plagiarism detection	Literary forensics	Style coaching
Sentiment analysis	Community morale monitoring	Product review triage	Customer care
Behavior prediction	Finance	Election forecasting	Marketing
Creative writing	Movie scripts	Poetry	Song lyrics

## Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks  
Example GATE Pipeline

## NLP

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS) Tagging  
Chunking and Parsing  
Named Entity Recognition  
Entity Linking  
Pipelines

## Text Mining

Example: Scientific Literature Mining  
Summary

## Notes and Further Reading

# Industrial-Strength Natural Language Processing

IN PYTHON

## Get things done

spaCy is designed to help you do real work – to build real products, or gather real insights. The library respects your time, and tries to avoid wasting it. It's easy to install, and its API is simple and productive.

[GET STARTED](#)

## Blazing fast

spaCy excels at large-scale information extraction tasks. It's written from the ground up in carefully memory-managed Cython. If your application needs to process entire web dumps, spaCy is the library you want to be using.

[FACTS & FIGURES](#)

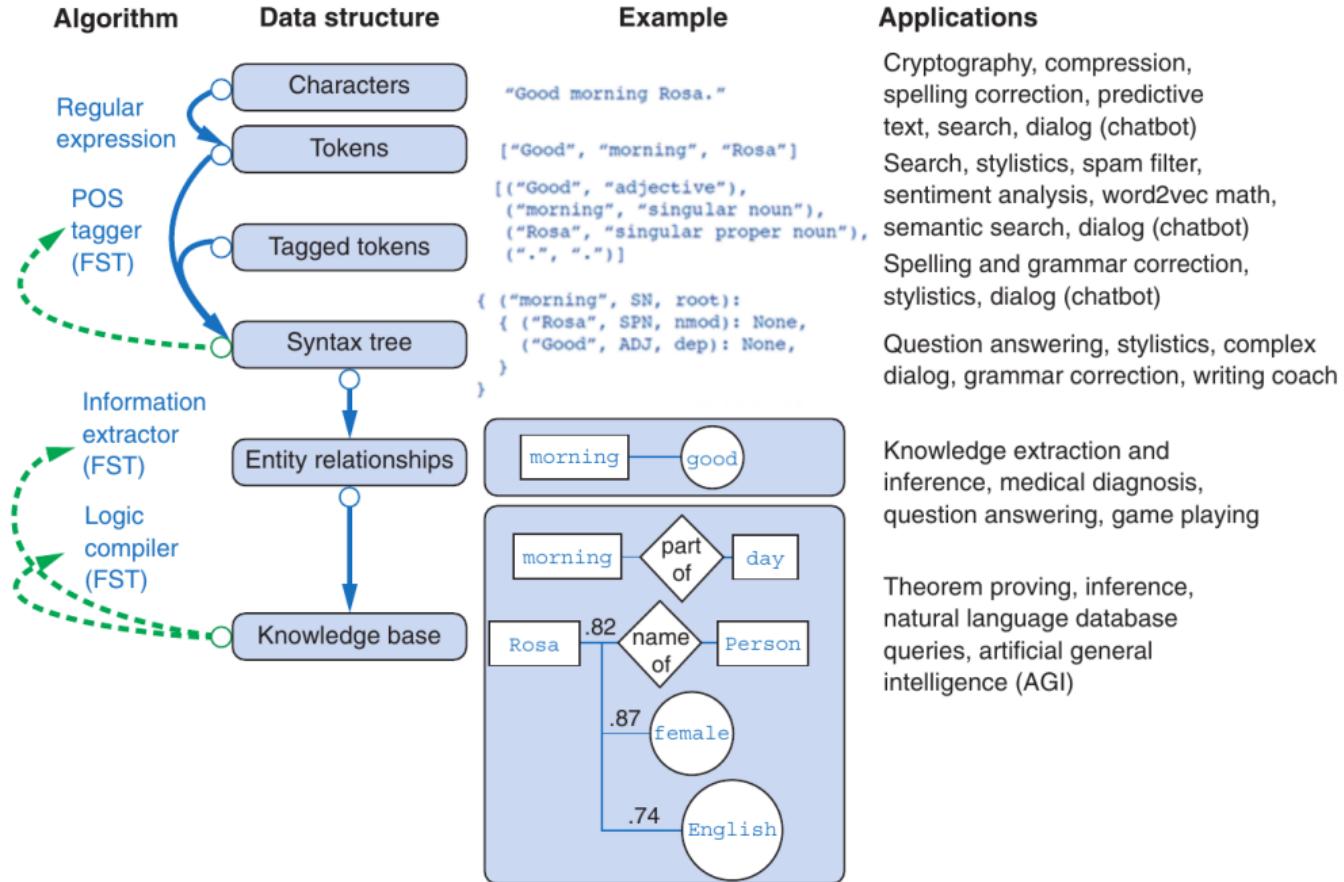
## Awesome ecosystem

In the five years since its release, spaCy has become an industry standard with a huge ecosystem. Choose from a variety of plugins, integrate with your machine learning stack and build custom components and workflows.

[READ MORE](#)

# Example NLP Pipeline

René Witte



## Introduction

- Text Mining in Science
- Text Mining Applications
- Language Technology (LT)
- Development Frameworks
- Example GATE Pipeline

## NLP

- Language Models
- Tokenization
- Sentence Splitting
- Morphology
- Part-of-Speech (POS) Tagging
- Chunking and Parsing
- Named Entity Recognition
- Entity Linking
- Pipelines

## Text Mining

- Example: Scientific Literature Mining
- Summary

## Notes and Further Reading

## Language dependent code

- Some parts of spaCy work language-independent
- But many steps require **language-specific data**, such as rules or pre-trained ML models

Need to load a **language model** to start, e.g., for English:

```
import spacy
nlp = spacy.load("en_core_web_sm")
```

LANGUAGE	CODE	LANGUAGE DATA	PIPELINES
Chinese	zh	lang/zh < >	4 packages 
Danish	da	lang/da < >	3 packages 
Dutch	nl	lang/nl < >	3 packages 
English	en	lang/en < >	4 packages 
French	fr	lang/fr < >	4 packages 
German	de	lang/de < >	4 packages 
Greek	el	lang/el < >	3 packages 
Italian	it	lang/it < >	3 packages 
Japanese	ja	lang/ja < >	3 packages 

### Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks  
Example GATE Pipeline

### NLP

#### Language Models

Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS) Tagging  
Chunking and Parsing  
Named Entity Recognition  
Entity Linking  
Pipelines

#### Text Mining

Example: Scientific Literature Mining Summary

#### Notes and Further Reading

## Tokenization

Text is split into basic units called *Tokens*:

- word tokens
- number tokens
- space tokens
- ...

Consistent tokenization is important for all later processing steps

## What is a word?

Unfortunately, even tokenization can be difficult:

- Is “John’s” in *John’s sick* one token or two?  
If one → problems in parsing (where’s the verb?)  
If two → what do we do with *John’s house*?
- What to do with hyphens?  
E.g., *database* vs. *data-base* vs. *data base*
- what to do with “C++”, “A/C”, “:-)”, “...”?

### Introduction

- Text Mining in Science
- Text Mining Applications
- Language Technology (LT)
- Development Frameworks
- Example GATE Pipeline

### NLP

- Language Models
- Tokenization**
- Sentence Splitting
- Morphology
- Part-of-Speech (POS) Tagging
- Chunking and Parsing
- Named Entity Recognition
- Entity Linking
- Pipelines

### Text Mining

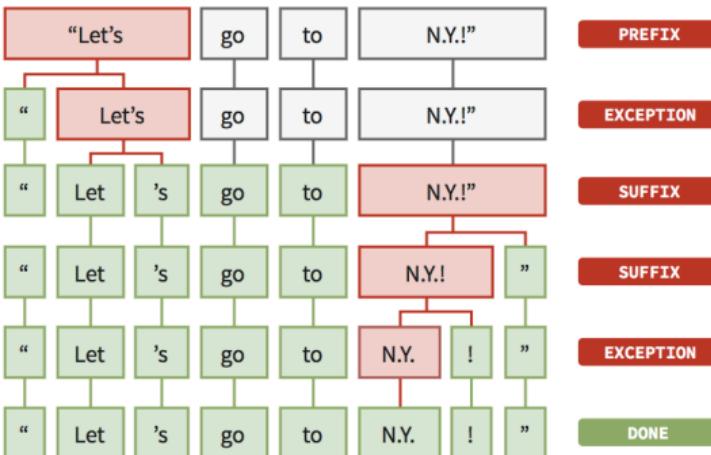
- Example: Scientific Literature Mining
- Summary

### Notes and Further Reading

# Tokenization



```
doc = nlp(u"Let's go to N.Y.!")
print([token.text for token in doc])
```



## Introduction

- Text Mining in Science
- Text Mining Applications
- Language Technology (LT)
- Development Frameworks
- Example GATE Pipeline

## NLP

- Language Models

### Tokenization

- Sentence Splitting
- Morphology
- Part-of-Speech (POS) Tagging
- Chunking and Parsing
- Named Entity Recognition
- Entity Linking
- Pipelines

## Text Mining

- Example: Scientific Literature Mining
- Summary

## Notes and Further Reading

## Mark Sentence Boundaries

Detects sentence units. Easy case:

- often, sentences end with “.”, “!”, or “?”

Hard (or annoying) cases:

- difficult when a “.” do not indicate an EOS:  
“*MR. X*”, “*3.14*”, “*Y Corp.*”, ...
- we can detect common abbreviations (“U.S.”), but what if a sentence ends with one?  
“...announced today by the U.S. The...”
- Sentences can be *nested* (e.g., within quotes)

## Correct sentence boundary is important

for many downstream analysis tasks:

- POS-Taggers maximize probabilities of tags within a sentence
- Most Parsers work on individual sentences

See [https://en.wikipedia.org/wiki/Sentence\\_boundary\\_disambiguation](https://en.wikipedia.org/wiki/Sentence_boundary_disambiguation)

### Introduction

Text Mining in Science

Text Mining Applications

Language Technology (LT)

Development Frameworks

Example GATE Pipeline

### NLP

Language Models

Tokenization

### Sentence Splitting

Morphology

Part-of-Speech (POS)

Tagging

Chunking and Parsing

Named Entity Recognition

Entity Linking

Pipelines

### Text Mining

Example: Scientific

Literature Mining

Summary

### Notes and Further Reading

## Some difficult examples for sentence splitting

René Witte



*I live in the U.S. but I commute to work in Mexico on S.V. Australis for a woman from St. Bernard St. on the Gulf of Mexico.*

*I went to G.T. You?*

*She yelled “It’s right here!” but I kept looking for a sentence boundary anyway.*

*I stared dumbfounded on as things like “How did I get here?,” “Where am I?,” “Am I alive?” flittered across the screen.*

*The author wrote “I don’t think it’s conscious.’ Turing said.”*

[https://www.tm-town.com/natural-language-processing#golden\\_rules](https://www.tm-town.com/natural-language-processing#golden_rules)

→ Worksheet #8: Task 4

### Introduction

- Text Mining in Science
- Text Mining Applications
- Language Technology (LT)
- Development Frameworks
- Example GATE Pipeline

### NLP

- Language Models
- Tokenization
- Sentence Splitting
- Morphology
- Part-of-Speech (POS) Tagging
- Chunking and Parsing
- Named Entity Recognition
- Entity Linking
- Pipelines

### Text Mining

- Example: Scientific Literature Mining
- Summary

### Notes and Further Reading

## Morphological Variants

Words are changed through a morphological process called *inflection*:

- typically indicates changes in case, gender, number, tense, etc.
- example *car* → *cars*, *give* → *gives*, *gave*, *given*

Goal: “normalize” words

## Stemming and Lemmatization

Two main approaches to normalization:

Stemming reduce words to a *base form*

Lemmatization reduce words to their *lemma*

Main difference: stemming just finds **any** base form, which doesn't even need to be a word in the language! Lemmatization find the actual *root* of a word, but requires morphological analysis.

### Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks  
Example GATE Pipeline

### NLP

Language Models  
Tokenization  
Sentence Splitting

### Morphology

Part-of-Speech (POS) Tagging  
Chunking and Parsing  
Named Entity Recognition  
Entity Linking  
Pipelines

### Text Mining

Example: Scientific Literature Mining  
Summary

### Notes and Further Reading

## Stemming

Commonly used in Information Retrieval:

- Can be achieved with rule-based algorithms, usually based on suffix-stripping
- Standard algorithm for English: the *Porter* stemmer
- Advantages: simple & fast
- Disadvantages:
  - Rules are language-dependent
  - Can create words that do not exist in the language, e.g., *computers* → *comput*
  - Often reduces different words to the same stem, e.g.,  
*army, arm* → *arm*  
*stocks, stockings* → *stock*
- Stemming for other languages: *Lucene* and *Snowball* stemmer have rule files for many languages

### Introduction

- Text Mining in Science
- Text Mining Applications
- Language Technology (LT)
- Development Frameworks
- Example GATE Pipeline

### NLP

- Language Models
- Tokenization
- Sentence Splitting

### Morphology

- Part-of-Speech (POS) Tagging
- Chunking and Parsing
- Named Entity Recognition
- Entity Linking
- Pipelines

### Text Mining

- Example: Scientific Literature Mining
- Summary

### Notes and Further Reading

## Lemmatization

Lemmatization is the process of deriving the base form, or *lemma*, of a word from one of its inflected forms. This requires a morphological analysis, which in turn typically requires a *lexicon*.

- Advantages:
  - identifies the *lemma* (root form), which is an actual word
  - less errors than in stemming
- Disadvantages:
  - more complex than stemming, slower
  - requires additional language-dependent resources
- While stemming is good enough for Information Retrieval, Text Mining often requires lemmatization
  - Semantics is more important (we need to distinguish an *army* and an *arm!*)
  - Errors in low-level components can multiply when running downstream

### Introduction

- Text Mining in Science
- Text Mining Applications
- Language Technology (LT)
- Development Frameworks
- Example GATE Pipeline

### NLP

- Language Models
- Tokenization
- Sentence Splitting

### Morphology

- Part-of-Speech (POS) Tagging
- Chunking and Parsing
- Named Entity Recognition
- Entity Linking
- Pipelines

### Text Mining

- Example: Scientific Literature Mining
- Summary

### Notes and Further Reading

## Where are we now?

So far, we splitted texts into *tokens* and *sentences* and performed some *normalization*.

- Still a long way to go to an *understanding* of natural language...

Typical approach in text mining: deal with the complexity of language by applying intermediate processing steps to acquire more and more structure. Next stop: *POS-Tagging*.

## POS-Tagging

A statistical POS Tagger scans tokens and assigns **POS Tags**.

*A black cat plays...* → *A/DT black/JJ cat/NN plays/VB...*

- relies on different word order probabilities
- needs a manually tagged corpus for machine learning

Note: *this is not parsing!*

### Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks  
Example GATE Pipeline

### NLP

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
**Part-of-Speech (POS) Tagging**  
Chunking and Parsing  
Named Entity Recognition  
Entity Linking  
Pipelines

### Text Mining

Example: Scientific  
Literature Mining  
Summary

### Notes and Further Reading

## Tagsets

A **tagset** defines the tags to assign to words. Main POS classes are:

**Noun** refers to entities like people, places, things or ideas

**Adjective** describes the properties of nouns or pronouns

**Verb** describes actions, activities and states

**Adverb** describes a verb, an adjective or another adverb

**Pronoun** word that can take the place of a noun

**Determiner** describes the particular reference of a noun

**Preposition** expresses spatial or time relationships

Note: real tagsets have from 45 (Penn Treebank) to 146 tags (C7).

## Introduction

[Text Mining in Science](#)

[Text Mining Applications](#)

[Language Technology \(LT\)](#)

[Development Frameworks](#)

[Example GATE Pipeline](#)

## NLP

[Language Models](#)

[Tokenization](#)

[Sentence Splitting](#)

[Morphology](#)

**Part-of-Speech (POS) Tagging**

[Chunking and Parsing](#)

[Named Entity Recognition](#)

[Entity Linking](#)

[Pipelines](#)

## Text Mining

[Example: Scientific](#)

[Literature Mining](#)

[Summary](#)

**Notes and Further Reading**

## Fundamentals

POS-Tagging generally requires:

Training phase where a **manually annotated** corpus is processed by a machine learning algorithm; and a

Tagging algorithm that processes texts using learned parameters.

Performance is generally good (around 96%) when staying in the same domain.

## Algorithms used in POS-Tagging

There is a multitude of approaches, commonly used are:

- Decision Trees
- Hidden Markov Models (HMMs)
- Support Vector Machines (SVM)
- Transformation-based Taggers (e.g., the **Brill** tagger)

### Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks  
Example GATE Pipeline

### NLP

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS) Tagging  
Chunking and Parsing  
Named Entity Recognition  
Entity Linking  
Pipelines

### Text Mining

Example: Scientific Literature Mining  
Summary

### Notes and Further Reading

# POS Tagging in spaCy

René Witte



```
import spacy

nlp = spacy.load("en_core_web_sm")
doc = nlp("Apple_is_looking_at_buying_U.K._startup_for_$1_billion")

for token in doc:
    print(token.text, token.lemma_, token.pos_, token.tag_, token.dep_,
          token.shape_, token.is_alpha, token.is_stop)
```

TEXT	LEMMA	POS	TAG	DEP	SHAPE	ALPHA	STOP
Apple	apple	PROPN	NNP	nsubj	Xxxxx	True	False
is	be	VERB	VBD	aux	xx	True	True
looking	look	VERB	VBG	ROOT	xxxx	True	False
at	at	ADP	IN	prep	xx	True	True
buying	buy	VERB	VBG	pcomp	xxxx	True	False
U.K.	u.k.	PROPN	NNP	compound	X.X.	False	False
startup	startup	NOUN	NN	dobj	xxxx	True	False
for	for	ADP	IN	prep	xxx	True	True
\$	\$	SYM	\$	quantmod	\$	False	False

## Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks  
Example GATE Pipeline

## NLP

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS) Tagging

Chunking and Parsing  
Named Entity Recognition  
Entity Linking Pipelines

## Text Mining

Example: Scientific Literature Mining Summary

## Notes and Further Reading

## Understanding POS Tags

- There are different tagsets used by different tools
- spaCy has a built-in explanation method:

```
spacy.explain("NNP")
> noun, proper singular
```

- spaCy uses the [Universal Dependency Scheme](https://universaldependencies.org/u/pos/)  
(<https://universaldependencies.org/u/pos/>)

## → Worksheet #8: Task 5

POS	DESCRIPTION	EXAMPLES
ADJ	adjective	big, old, green, incomprehensible, first
ADP	adposition	in, to, during
ADV	adverb	very, tomorrow, down, where, there
AUX	auxiliary	is, has (done), will (do), should (do)
CONJ	conjunction	and, or, but

### Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks  
Example GATE Pipeline

### NLP

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS) Tagging  
Chunking and Parsing  
Named Entity Recognition  
Entity Linking  
Pipelines

### Text Mining

Example: Scientific Literature Mining Summary

Notes and Further Reading

## Finding Syntactic Structures

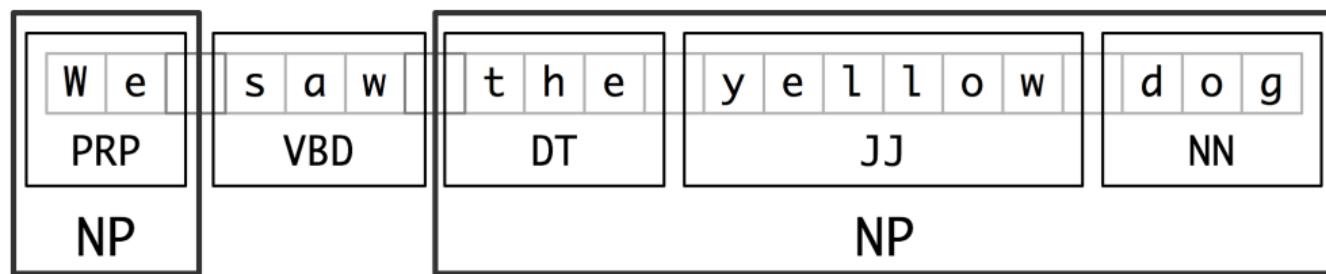
We can now start a **syntactic analysis** of a sentence using:

- Parsing      producing a *parse tree* for a sentence using a parser, a grammar, and a lexicon
- Chunking      finding syntactic constituents like *Noun Phrases (NPs)* or *Verb Groups (VGs)* within a sentence

## Chunking vs. Parsing

Producing a *full parse tree* often fails due to grammatical inaccuracies, novel words, bad tokenization, wrong sentence splits, errors in POS tagging, ...

Hence, *chunking* and *partial parsing* are more commonly used.



From Natural Language Processing with Python, by Steven Bird, Ewan Klein and Edward Loper, Copyright 2019 the authors. It is distributed with the Natural Language Toolkit (<http://nltk.org/>), Version 3.0, under the terms of the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 United States License (<http://creativecommons.org/licenses/by-nd/3.0/us/>).

### Introduction

- Text Mining in Science
- Text Mining Applications
- Language Technology (LT)
- Development Frameworks
- Example GATE Pipeline

### NLP

- Language Models
- Tokenization
- Sentence Splitting
- Morphology
- Part-of-Speech (POS) Tagging

### Chunking and Parsing

- Named Entity Recognition
- Entity Linking
- Pipelines

### Text Mining

- Example: Scientific Literature Mining
- Summary

### Notes and Further Reading

## NP Chunker

Rule-based approach for finding NPs

## Grammar Excerpt

```
(NP      (DET MOD HEAD))  
(MOD    (MOD-ingredients)  
        (MOD-ingredients MOD)  
        ())  
(HEAD   (NN)  ...)
```

## Example

"I couldn't believe what I saw," said McNeill, who also discovered bomb-making instructions and detailed maps of U.S. landmarks in the cave. "On top of all the destruction these people had already unleashed, plans were underway to harass the American people with a merciless assault of offers for everything from discounts on home DSL lines to pre-approved, low-interest credit cards."

For all the evidence collected by the CIA, the "smoking gun" in the investigation may turn out to be an alleged Osama bin Laden motivational videotape, currently in the possession of CNN. The controversial tape, which has never aired on the cable network, is rumored to feature bin Laden urging his followers to think positive and believe in the quality of the product they are pitching, closing on the grim slogan "Smile And Dial."

Type	Set	Start	End	Features
P	Default	3582	3596	{DET="", MOD="", HEAD="Guantanamo Bay "}
P	Default	776	791	{DET="the ", MOD="dinner ", HEAD="hour "}
P	Default	2259	2262	{DET="", MOD="", HEAD="out "}
P	Default	1806	1807	{DET="", MOD="", HEAD="I "}
P	Default	3849	3852	{DET="", MOD="", HEAD="one "}
P	Default	987	996	{DET="The ", MOD="", HEAD="video "}
P	Default	1487	1494	{DET="", MOD="", HEAD="McNeill "}
P	Default	2280	2318	{DET="", MOD="Osama bin Laden motivational ", HEAD="videotape "}
P	Default	894	910	{DET="", MOD="money ", HEAD="laundering "}

## Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks  
Example GATE Pipeline

## NLP

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS)  
Tagging

## Chunking and Parsing

Named Entity Recognition  
Entity Linking  
Pipelines

## Text Mining

Example: Scientific Literature Mining  
Summary

## Notes and Further Reading

```
import spacy

nlp = spacy.load("en_core_web_sm")
doc = nlp("Autonomous_cars_shift_insurance_liability_toward_manufacturers")
for chunk in doc.noun_chunks:
    print(chunk.text, chunk.root.text, chunk.root.dep_,
          chunk.root.head.text)
```

TEXT	ROOT.TEXT	ROOT.DEP_	ROOT.HEAD.TEXT
Autonomous cars	cars	nsubj	shift
insurance liability	liability	dobj	shift
manufacturers	manufacturers	pobj	toward

## Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks  
Example GATE Pipeline

## NLP

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS)  
Tagging

Chunking and Parsing  
Named Entity Recognition  
Entity Linking  
Pipelines

## Text Mining

Example: Scientific  
Literature Mining  
Summary

Notes and Further  
Reading

## What can we do with chunks?

(NP) chunks are very useful in finding **named entities** (NEs), e.g., *Persons*, *Companies*, *Locations*, *Patents*, *Organisms*, . . .

But additional methods are needed for finding **relations**:

- *Who* invented *X*?
- *What* company created product *Y* that is doomed to fail?
- *Which* organism is this protein coming from?

Parse trees can help in determining these relationships

## Parsing Challenges

Parsing is hard due to many kinds of ambiguities:

**PP-Attachment** which NP takes the PP? Compare:

*He ate spaghetti with a fork.*

*He ate spaghetti with tomato sauce.*

**NP Bracketing** *plastic cat food can cover*

### Introduction

Text Mining in Science

Text Mining Applications

Language Technology (LT)

Development Frameworks

Example GATE Pipeline

### NLP

Language Models

Tokenization

Sentence Splitting

Morphology

Part-of-Speech (POS)

Tagging

### Chunking and Parsing

Named Entity Recognition

Entity Linking

Pipelines

### Text Mining

Example: Scientific

Literature Mining

Summary

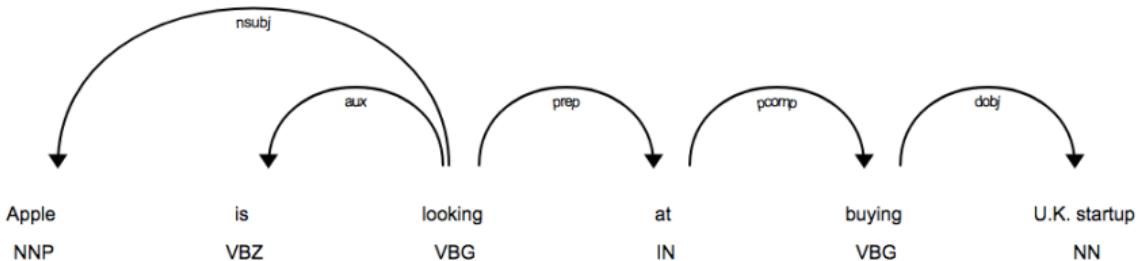
### Notes and Further Reading

# POS tags & dependencies



```
doc = nlp(u"Apple is looking at buying U.K. startup")

for token in doc:
    print(token.text, token.pos_, token.tag_)
```



## Introduction

- Text Mining in Science
- Text Mining Applications
- Language Technology (LT)
- Development Frameworks
- Example GATE Pipeline

## NLP

- Language Models
- Tokenization
- Sentence Splitting
- Morphology
- Part-of-Speech (POS) Tagging

## Chunking and Parsing

- Named Entity Recognition
- Entity Linking
- Pipelines

## Text Mining

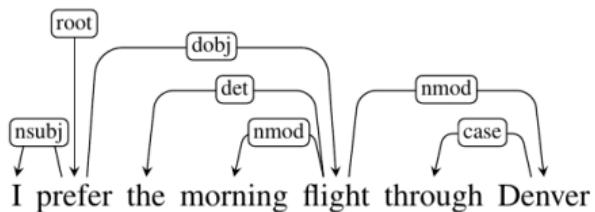
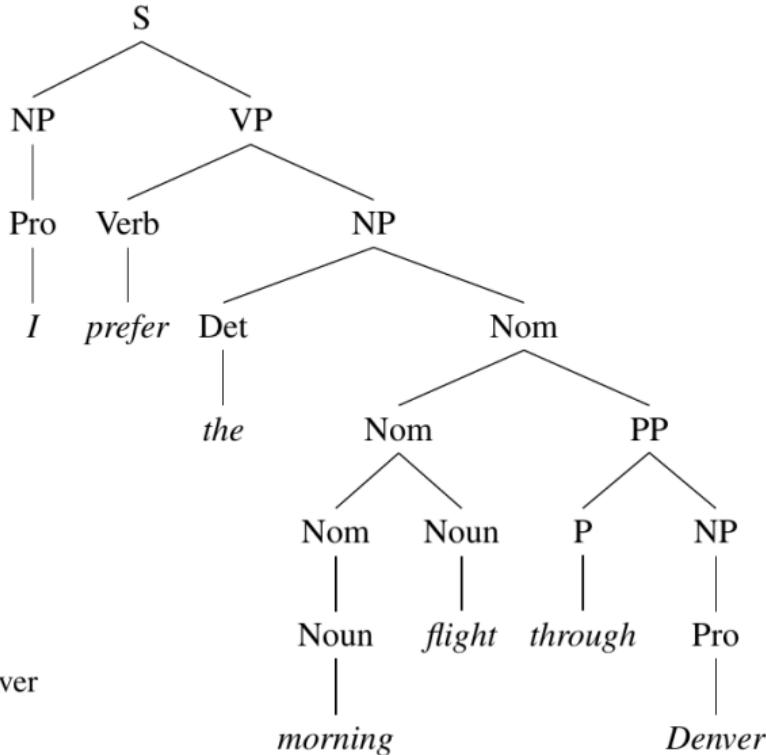
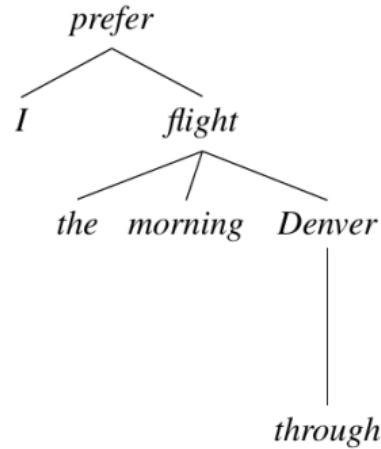
- Example: Scientific Literature Mining
- Summary

## Notes and Further Reading

# Constituent-based Parse Tree vs. Dependency Parsing

Parsing “I prefer the morning flight through Denver.”

René Witte



→ Worksheet #8: Task 6

## Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks  
Example GATE Pipeline

## NLP

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS) Tagging

Chunking and Parsing  
Named Entity Recognition  
Entity Linking  
Pipelines

## Text Mining

Example: Scientific Literature Mining  
Summary

Notes and Further Reading

# Named Entity Recognition in spaCy

René Witte



```
import spacy

nlp = spacy.load("en_core_web_sm")
doc = nlp("Apple is looking at buying U.K. startup for $1 billion")

for ent in doc.ents:
    print(ent.text, ent.start_char, ent.end_char, ent.label_)
```

TEXT	START	END	LABEL	DESCRIPTION
Apple	0	5	ORG	Companies, agencies, institutions.
U.K.	27	31	GPE	Geopolitical entity, i.e. countries, cities, states.
\$1 billion	44	54	MONEY	Monetary values, including unit.

Apple **ORG** Is looking at buying **U.K. GPE** startup for **\$1 billion MONEY**

## Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks  
Example GATE Pipeline

## NLP

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS) Tagging  
Chunking and Parsing  
Named Entity Recognition

Entity Linking  
Pipelines

## Text Mining

Example: Scientific Literature Mining  
Summary

## Notes and Further Reading

## Which entities are detected?

- Depends on the **model** and its **training data**
- E.g., spaCy trained on the OntoNotes-5.0 corpus (<https://catalog.ldc.upenn.edu/LDC2013T19>)

## → Worksheet #8: Task 7

TYPE	DESCRIPTION
PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)

### Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks  
Example GATE Pipeline

### NLP

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS) Tagging  
Chunking and Parsing

### Named Entity Recognition

Entity Linking  
Pipelines

### Text Mining

Example: Scientific Literature Mining  
Summary

### Notes and Further Reading

## How to find a new kind of Named Entity (NE)?

Two general solutions:

**Rule-based:** write rules (regular expressions, transducers) that capture as many variations as possible, with as few false positives as possible

**Machine learning:** train a machine learning model, using manually annotated examples (supervised learning)

## Pros&Cons

- Rules can be developed quickly: good for proof-of-concept/demo, bootstrapping a ML corpus, easy (unambiguous) patterns
- Rules are “brittle”, they do not generalize well
- ML solutions generally perform better (more robust with respect to variations)
- But a ML approach requires significant effort for creating training data, as well as effort for feature engineering, training, evaluation, etc.

### Introduction

[Text Mining in Science](#)

[Text Mining Applications](#)

[Language Technology \(LT\)](#)

[Development Frameworks](#)

[Example GATE Pipeline](#)

### NLP

[Language Models](#)

[Tokenization](#)

[Sentence Splitting](#)

[Morphology](#)

[Part-of-Speech \(POS\) Tagging](#)

[Chunking and Parsing](#)

### Named Entity Recognition

[Entity Linking](#)

[Pipelines](#)

### Text Mining

[Example: Scientific](#)

[Literature Mining](#)

[Summary](#)

### Notes and Further Reading

## Text Mining in Science

### Text Mining Applications

### Language Technology (LT)

### Development Frameworks

### Example GATE Pipeline

## NLP

### Language Models

### Tokenization

### Sentence Splitting

### Morphology

### Part-of-Speech (POS)

### Tagging

### Chunking and Parsing

### Named Entity Recognition

### Entity Linking

### Pipelines

## Text Mining

### Example: Scientific

### Literature Mining

### Summary

## Notes and Further

### Reading

→ Worksheet #8: Task 8



## Rule-based Matcher Explorer

Test spaCy's rule-based **Matcher** by creating token patterns interactively and running them over your text. Each token can set multiple attributes like text value, part-of-speech tag or boolean flags. The token-based view lets you explore how spaCy processes your text – and why your pattern matches, or why it doesn't. For more details on rule-based matching, see the [documentation](#).

- POS ▾ ADJ

- OP ▾ ? ▾

**add attribute**

- LEMMA ▾ match

- POS ▾ NOUN

**add attribute**

### Text to check against pattern

A match is a tool for starting a fire. Typically, modern matches are made of small wooden sticks or stiff paper. One end is coated with a material that can be ignited by frictional heat generated by striking the match against a suitable surface.

### Model ?

English - en\_core\_web\_sm (v2.0.0)

Show tokens displaCy ? displaCy ENT ?

A **match is** a tool for starting a fire. Typically, **modern matches are** made of small wooden sticks or stiff paper. One end is coated with a material that can be ignited by frictional heat generated by striking the match against a suitable surface. **Wooden matches are** packaged in matchboxes, and paper **matches are** partially cut into rows and stapled into matchbooks.

# Entity Linking

René Witte



## Grounding to a Knowledge Base

spaCy provides an API for linking entities to a knowledge base, but (currently) no pre-trained models.

## Example Pipeline

See some experimental development code using Wikidata at  
<https://pypi.org/project/spacy-entity-linker/>

```
import spacy
from SpacyEntityLinker import EntityLinker
entityLinker = EntityLinker()
nlp = spacy.load("en_core_web_sm")
nlp.add_pipe(entityLinker, last=True, name="entityLinker")
doc = nlp("I watched the Pirates of the Caribbean last silvester")

#returns all entities in the whole document
all_linked_entities=doc._.linkedEntities
for sent in doc.sents:
    sent._.linkedEntities.pretty_print()

#OUTPUT:
#https://www.wikidata.org/wiki/Q194318          194318
    Pirates of the Caribbean      Series of fantasy adventure films
#https://www.wikidata.org/wiki/Q12525597  12525597
    Silvester   the day celebrated on 31 December (Roman Catholic Church) or 2 January (Easte
```

### Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks  
Example GATE Pipeline

### NLP

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS) Tagging  
Chunking and Parsing  
Named Entity Recognition

### Entity Linking

Pipelines

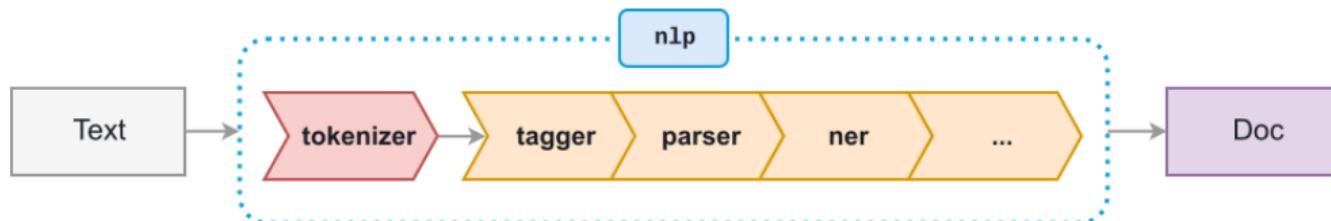
### Text Mining

Example: Scientific Literature Mining  
Summary

Notes and Further Reading

# Pipelines in spaCy

René Witte



NAME	COMPONENT	CREATES	DESCRIPTION
tokenizer	<a href="#">Tokenizer</a>	Doc	Segment text into tokens.
tagger	<a href="#">Tagger</a>	Doc[i].tag	Assign part-of-speech tags.
parser	<a href="#">DependencyParser</a>	Doc[i].head, Doc[i].dep, Doc.sents, Doc.noun_chunks	Assign dependency labels.
ner	<a href="#">EntityRecognizer</a>	Doc.ents, Doc[i].ent_iob, Doc[i].ent_type	Detect and label named entities.
textcat	<a href="#">TextCategorizer</a>	Doc.cats	Assign document labels.
...	<a href="#">custom components</a>	Doc._.xxx, Token._.xxx, Span._.xxx	Assign custom attributes, methods or properties.

## Introduction

[Text Mining in Science](#)  
[Text Mining Applications](#)  
[Language Technology \(LT\)](#)  
[Development Frameworks](#)  
[Example GATE Pipeline](#)

## NLP

[Language Models](#)  
[Tokenization](#)  
[Sentence Splitting](#)  
[Morphology](#)  
[Part-of-Speech \(POS\) Tagging](#)  
[Chunking and Parsing](#)  
[Named Entity Recognition](#)  
[Entity Linking](#)  
[Pipelines](#)

## Text Mining

[Example: Scientific Literature Mining](#)  
[Summary](#)

## Notes and Further Reading

# Pipelines in spaCy (contd.)

René Witte



## Working with pipelines

Loading a `model` defines the pipeline to be used in its metadata:

```
"pipeline": ["tagger", "parser", "ner"]
```

Processing a text will then apply each component in the pipeline in turn:

```
doc = nlp.make_doc("This_is_a_sentence")
for name, proc in nlp.pipeline:
    doc = proc(doc)
```

You can disable components you don't need:

```
nlp = spacy.load("en_core_web_sm", disable=["parser"])
```

And of course add your own components (here at the end):

```
nlp.add_pipe(my_component, name="My_new_component", last=True)
```

See <https://spacy.io/usage/processing-pipelines>

### Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks  
Example GATE Pipeline

### NLP

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS) Tagging  
Chunking and Parsing  
Named Entity Recognition  
Entity Linking  
**Pipelines**

### Text Mining

Example: Scientific Literature Mining  
Summary

Notes and Further Reading

# Writing a spaCy component

René Witte



## A simple “info” component

```
import spacy

def my_component(doc):
    print("After tokenization, this doc has {} tokens.".format(len(doc)))
    print("The part-of-speech tags are:", [token.pos_ for token in doc])
    if len(doc) < 10:
        print("This is a pretty short document.")
    return doc

nlp = spacy.load("en_core_web_sm")
nlp.add_pipe(my_component, name="print_info", last=True)
print(nlp.pipe_names)
doc = nlp("This is a sentence.")
```

## Output

```
['tagger', 'parser', 'ner', 'print_info']
After tokenization, this doc has 5 tokens.
The part-of-speech tags are: ['DET', 'AUX', 'DET', 'NOUN', 'PUNCT']
This is a pretty short document.
```

### Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks  
Example GATE Pipeline

### NLP

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS) Tagging  
Chunking and Parsing  
Named Entity Recognition  
Entity Linking  
**Pipelines**

### Text Mining

Example: Scientific Literature Mining  
Summary

Notes and Further Reading

# Summary: spaCy Architecture

René Witte



# spaCy



## Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks  
Example GATE Pipeline

## NLP

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS) Tagging  
Chunking and Parsing  
Named Entity Recognition  
Entity Linking  
**Pipelines**

## Text Mining

Example: Scientific Literature Mining  
Summary

## Notes and Further Reading

## 1 Introduction

### Introduction

- Text Mining in Science
- Text Mining Applications
- Language Technology (LT)
- Development Frameworks
- Example GATE Pipeline

## 2 Natural Language Processing (NLP)

### NLP

- Language Models
- Tokenization
- Sentence Splitting
- Morphology
- Part-of-Speech (POS) Tagging
- Chunking and Parsing
- Named Entity Recognition
- Entity Linking
- Pipelines

## 3 Text Mining

Example: Scientific Literature Mining  
Summary

### Text Mining

- Example: Scientific Literature Mining
- Summary

## 4 Notes and Further Reading

Notes and Further Reading

# Text Mining Biomedical Literature for Protein Mutations

René Witte

## Excerpts from PubMed journal PMID:14592457

... glutathione S-transferase (GST) fusion proteins in *Escherichia coli* and purified by GSH–agarose affinity chromatography. Mutant Q15K-W37R and mutant Q15R-W37R showed comparable activity for NAD and NADP with an increase in activity nearly 3fold over that of the wild type.

(Orange: Mutation, Red: Enzyme, Blue: Organism, Violet: Impact expression, Purple: Protein property, Green: Physical quantity)

## What we are looking for?

<b>Impact</b>	Mutant Q15K-W37R and mutant Q15R-W37R showed... an increase in activity 3fold over that of the wild type.
<b>Organism</b>	<i>Escherichia coli</i>
<b>Mutation</b>	Q15K/W37R,Q15R/W37R
<b>Enzyme</b>	glutathione S-transferase (GST)
<b>Protein property</b>	activity
<b>Physical Quantity</b>	3fold
<b>Impact Expression</b>	increase

### Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks  
Example GATE Pipeline

### NLP

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS) Tagging  
Chunking and Parsing  
Named Entity Recognition  
Entity Linking  
Pipelines

### Text Mining

Example: Scientific Literature Mining  
Summary

Notes and Further Reading

# Enzymatic mechanism of low-activity mouse alcohol dehydrogenase 2

Stroemberg, P.; Svensson, S.; Berst, K.B.; Plapp, B.V.; Höög, J.O.; *Biochemistry* 43, 1323-1328 (2004)

## Data extracted from this reference:

### Engineering

Amino acid exchange	Commentary	Organism
P47A	site-directed mutagenesis, about 100fold increased activity compared to the wild-type enzyme	Mus musculus
P47H	site-directed mutagenesis, about 100fold increased activity compared to the wild-type enzyme	Mus musculus
P47Q	site-directed mutagenesis, about 100fold increased activity compared to the wild-type enzyme	Mus musculus

### Inhibitors

Inhibitors	Commentary	Organism	Structure
cyclohexylformamide	dead-end inhibition pattern	Mus musculus	
Octanoic acid	dead-end inhibition pattern	Mus musculus	

### Metals/Ions

Metals/Ions	Commentary	Organism	Structure
Zn <sup>2+</sup>	catalytic zinc ion	Mus musculus	

### Organism

Organism	Primary Accession No. (UniProt)	Commentary	Textmining
Mus musculus	-	low-activity isozyme ADH2	-

### Introduction

- Text Mining in Science
- Text Mining Applications
- Language Technology (LT)
- Development Frameworks
- Example GATE Pipeline

### NLP

- Language Models
- Tokenization
- Sentence Splitting
- Morphology
- Part-of-Speech (POS) Tagging
- Chunking and Parsing
- Named Entity Recognition
- Entity Linking
- Pipelines

### Text Mining

- Example: Scientific Literature Mining
- Summary

### Notes and Further Reading

## Organism Examples

genus      old genus name      species  
Emericella      (Aspergillus)      nidulans  
organism mention

genus      species      strain  
Escherichia      coli      XLI-Blue  
organism mention

## Finding Organisms: Rule matching

Priority	Pattern	
5	(GENUS) (SPECIES) (SUBSPECIES) (STRAIN)?	Notes and Further Reading
4	(GENUS) ("(")(GENUS)("") (SPECIES) (STRAINKEYWORD)? (STRAIN) (STRAINKEYWORD)?	
3	(SPECIES) (STRAINKEYWORD) (STRAIN)	
2	(GENUS) (STRAINKEYWORD)? (STRAIN)	
1	(FULLNAME) (STRAINKEYWORD)? (STRAIN) (STRAINKEYWORD)?	Example: Scientific Literature Mining Summary

# Query Interface (<https://www.semanticsoftware.info/omm-query>)

"Show me all protein mutations that impacted the protein property *affinity*"

René Witte



Searching index: PMC-2012-06-18

{Impact} OVER affinity

Search

Results 1 - 10 of 6,993

[2796128 \(cached\)](#)

PIKKs (27). The caffeine resistance mutations in TOR1 decrease affinity for caffeine or confer increased TORC1 kinase activity. The W2176R mutation

[2796128 \(cached\)](#)

TORC1 kinase activity. The W2176R mutation in TOR1 decreased affinity for caffeine. Trp2176 is conserved in

[2855616 \(cached\)](#)

reported (7). The ParM variant, labeled with two tetramethylrhodamines, binds ADP with relatively weak affinity (dissociation constant 30 μM) but responds to ADP binding with ~15-fold signal increase. This means that the tetramethylrhodamine

[2855616 \(cached\)](#)

variant for tetramethylrhodamine labeling. The same mutations had been successful in decreasing ATP affinity in the MDCC-ParM biosensor (7). The new ParM mutant

## Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks  
Example GATE Pipeline

## NLP

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS) Tagging  
Chunking and Parsing  
Named Entity Recognition  
Entity Linking  
Pipelines

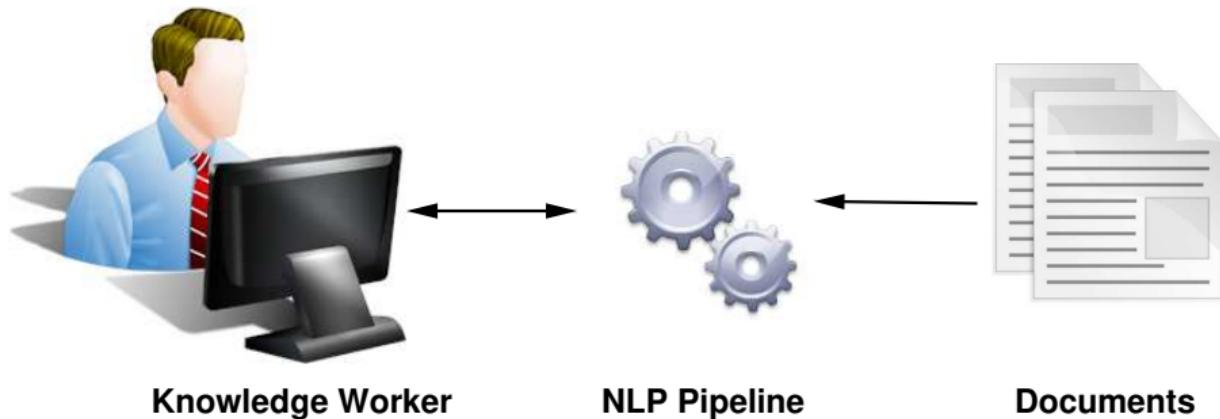
## Text Mining

Example: Scientific Literature Mining  
Summary

Notes and Further Reading

## Building a text mining system

- We have mature tools&resources to build robust, scalable systems now
- Prototyping a basic system demo can be done “in hours”
- Of course, some tasks are still complex R&D problems
- Cloud APIs are another option (convenient, but added cost and confidentiality concerns)
- Many businesses are still not aware of the potentials in analyzing their documents



### Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks  
Example GATE Pipeline

### NLP

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS) Tagging  
Chunking and Parsing  
Named Entity Recognition  
Entity Linking  
Pipelines

### Text Mining

Example: Scientific Literature Mining  
Summary

Notes and Further Reading

## 1 Introduction

### Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks  
Example GATE Pipeline

## 2 Natural Language Processing (NLP)

### NLP

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS) Tagging  
Chunking and Parsing  
Named Entity Recognition  
Entity Linking  
Pipelines

## 3 Text Mining

### Text Mining

Example: Scientific Literature Mining  
Summary

## 4 Notes and Further Reading

Notes and Further Reading

## Required

- [LHH19, Chapter 11] (Information extraction)

## Supplemental

- [PS12, Chapter 6] (Annotation and Adjudication)
- [JM, Chapter 14] (Dependency Parsing)

### Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks  
Example GATE Pipeline

### NLP

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS) Tagging  
Chunking and Parsing  
Named Entity Recognition  
Entity Linking  
Pipelines

### Text Mining

Example: Scientific Literature Mining  
Summary

Notes and Further Reading

# References

René Witte



- [JM] Daniel Jurafsky and James H. Martin.  
*Speech and Language Processing*.  
Third Edition draft, December 30, 2020.  
<https://web.stanford.edu/~jurafsky/slp3/>.
- [LHH19] Hobson Lane, Cole Howard, and Hannes Max Hapke.  
*Natural Language Processing in Action*.  
Manning Publications Co., 2019.  
<https://concordiauniversity.on.worldcat.org/oclc/1102387045>.
- [PS12] James Pustejovsky and Amber Stubbs.  
*Natural Language Annotation for Machine Learning*.  
O'Reilly, 2012.  
<https://concordiauniversity.on.worldcat.org/oclc/801812987>.

## Introduction

Text Mining in Science  
Text Mining Applications  
Language Technology (LT)  
Development Frameworks  
Example GATE Pipeline

## NLP

Language Models  
Tokenization  
Sentence Splitting  
Morphology  
Part-of-Speech (POS) Tagging  
Chunking and Parsing  
Named Entity Recognition  
Entity Linking  
Pipelines

## Text Mining

Example: Scientific  
Literature Mining  
Summary

Notes and Further  
Reading