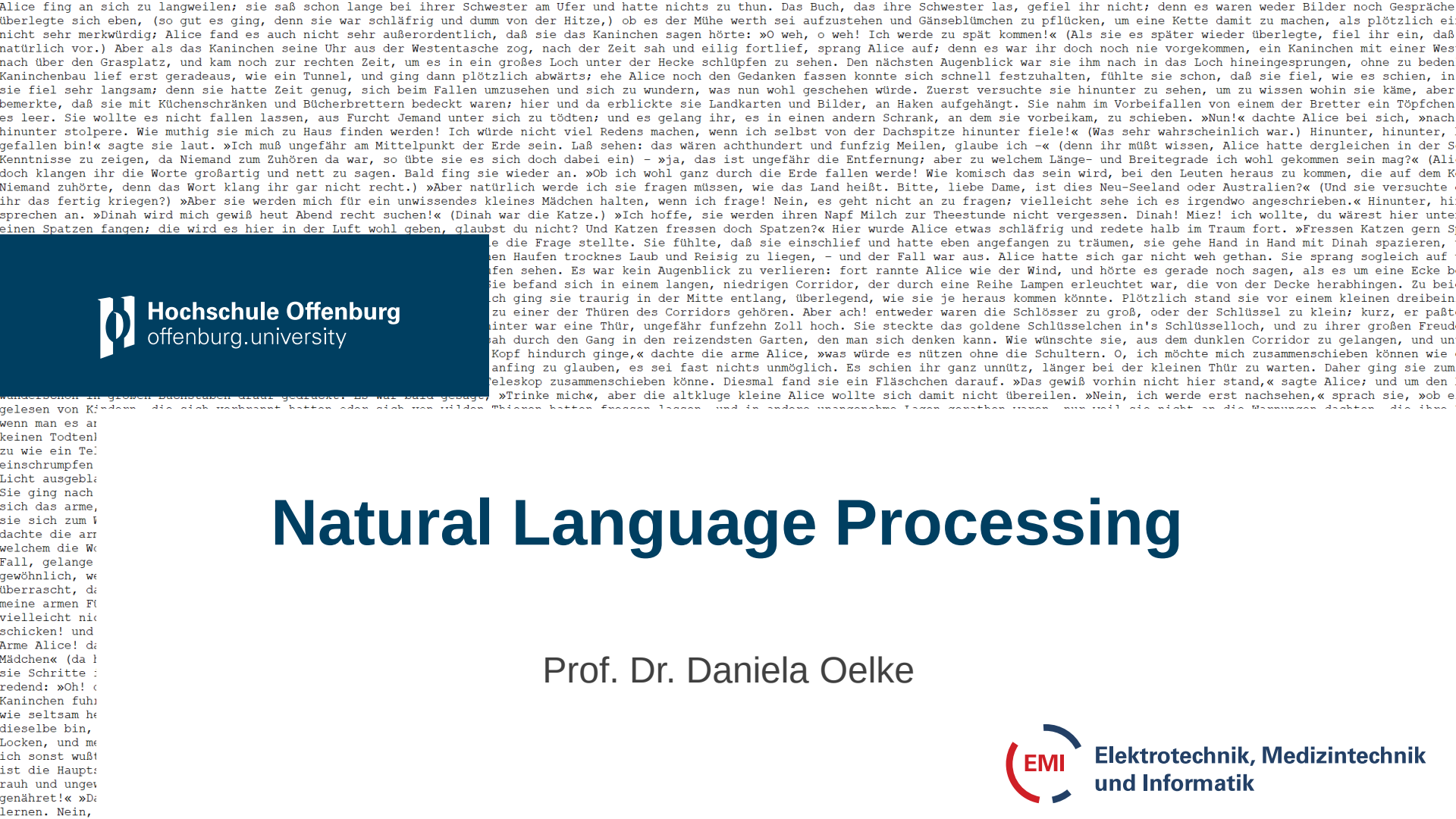


Hochschule Offenburg
offenburg.university

Natural Language Processing

Prof. Dr. Daniela Oelke

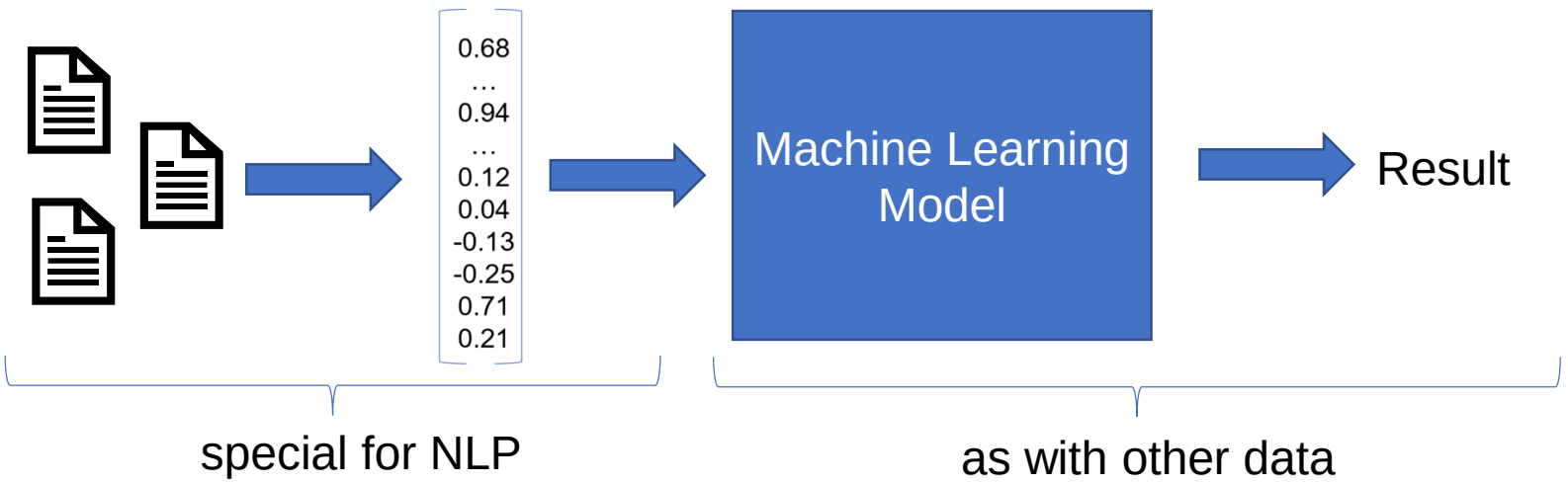
EMI Elektrotechnik, Medizintechnik und Informatik

Hochschule Offenburg
offenburg.university

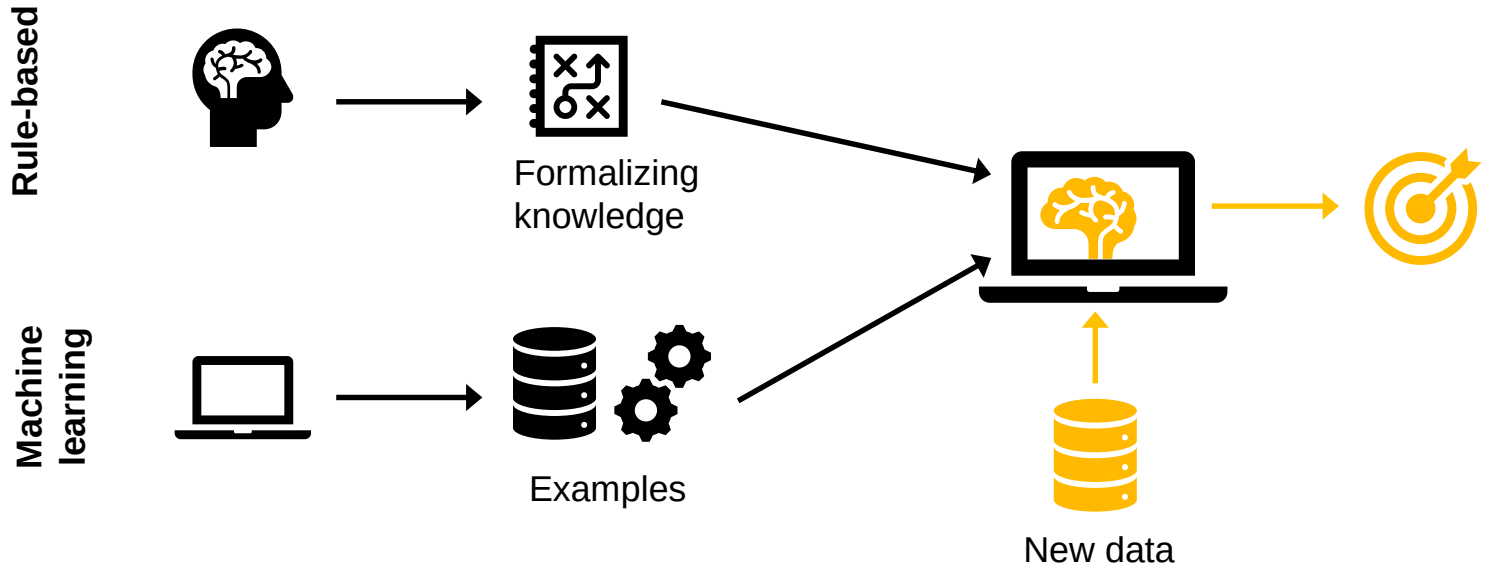
Natural Language Processing & Machine Learning

EMI Elektrotechnik, Medizintechnik und Informatik

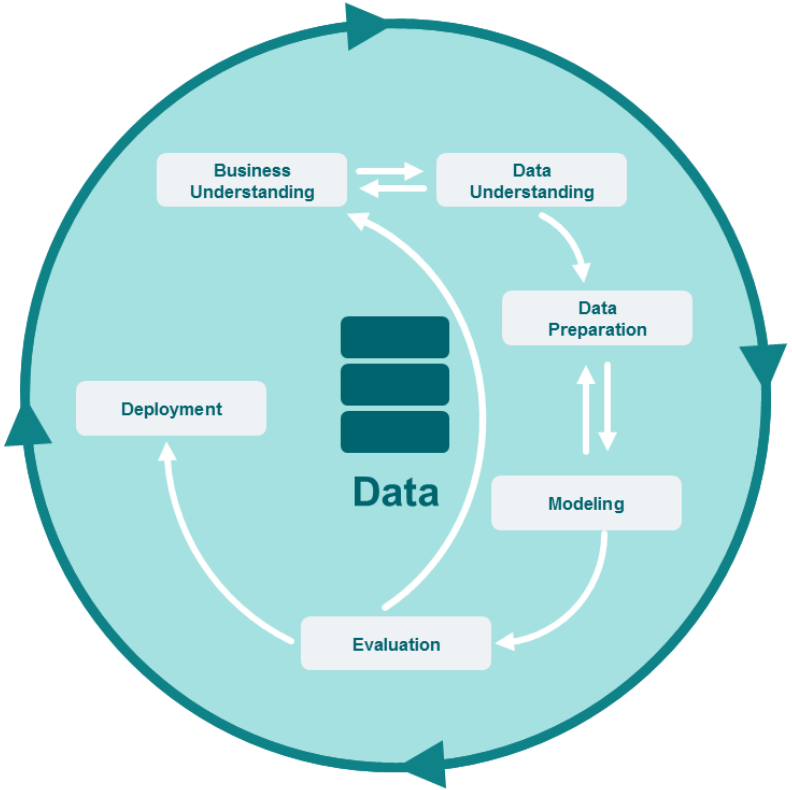
Classification / Clustering of Text Data



Coding of linguistic knowledge vs. machine learning



Cross Industry Standard Process for Data Mining (CRISP-DM)



Hochschule Offenburg
offenburg.university

Text Classification

Source: https://www.kaggle.com/jobspikr/data-scientist-job-postings-from-the-usa
Provided by: JobsPikr

Example: Data Science Job Postings

Hochschule Offenburg
offenburg.university

crawl_timestamp	job_title	category	company_name	city	state	country	inferred_city	post_date	job_description	job_type	job_board
06.02.2019 06:26	Enterprise Data Scientist I	Accounting/Finance	Farmers Insurance Group	Woodland Hills	CA	Usa	Woodland hills	06.02.2019	Read what people are saying about working here. We are Farmers!Join a team of diverse pr	Undefined	indeed
06.02.2019 06:33	Data Scientist		Luxoft USA Inc	Middletown	NJ	Usa	Middletown	05.02.2019	We have an immediate opening for a Sharp Data Scientist with a strong Mathematical/Statis	Undefined	dice
06.02.2019 06:33	Data Scientist		Cincinnati Bell Technology Solutions	New York	NY	Usa	New york	05.02.2019	Candidates should have the following background, skills and characteristics: Â Experience d	Full Time	dice
06.02.2019 06:33	Data Scientist, Aladdin Wealth Tech, Associate (Mc	Accounting/Finance	BlackRock	New York	NY 10055 (Midto	Usa	New york	06.02.2019	Read what people are saying about working here. About BlackRockBlackRock helps investors	Undefined	indeed
06.02.2019 06:48	Senior Data Scientist	biotech	CyberCoders	Charlotte	NC	Usa	Charlotte	05.02.2019	We are seeking an extraordinary Data Scientist in Charlotte to join our fast growing healthca	Full Time	monster
06.02.2019 06:36	CIB â€“ Fixed Income Research â€“ Machine Learni	Accounting/Finance	JP Morgan Chase	New York	NY 10179 (Midto	Usa	New york	05.02.2019	Read what people are saying about working here. OpportunityThe opportunity is to join our	Undefined	indeed
06.02.2019 06:34	Data Scientist, Licensing Operations	Accounting/Finance	Spotify	New York	NY 10011 (Chels	Usa	New york	06.02.2019	Read what people are saying about working here. At Spotify our mission is to provide the w	Undefined	indeed
06.02.2019 06:52	Sr. Data Scientist (Can work on Xoriant W2)		Xoriant Corporation	Santa Clara	CA	Usa	Santa clara	06.02.2019	Job Title: - Sr. Data Science Consultant Duration: 1+ Yrs (will get extend) Mode of interview -	Contract	dice
06.02.2019 06:34	Data Scientist, Aladdin Wealth Tech, Associate	Accounting/Finance	BlackRock	New York	NY 10055 (Midto	Usa	New york	06.02.2019	Read what people are saying about working here. About BlackRockBlackRock helps investors	Undefined	indeed
06.02.2019 07:03	Data Scientist		Adroit Resources	San Francisco	CA	Usa	San francisco	05.02.2019	â€¢ 3+ years related a professional experienceÂ â€¢ Proven achievements resulting from pr	Contract	dice
06.02.2019 07:19	Data Scientist	Computer/Internet	Northrop Grumman	Monterey	CA 93940	Usa	Monterey	06.02.2019	Read what people are saying about working here. At Northrop Grumman, innovation isn't ju	Undefined	indeed
06.02.2019 07:24	Data Scientist		Envoy Consulting Group, Inc	Reston	VA	Usa	Reston	05.02.2019	Data Scientist EnvoyIT is looking for a Data Scientist for a Full-time position with one of our	Full Time	dice
06.02.2019 07:21	ETL Developer / Data Scientist	Computer/Internet	Noblis	Reston	VA 20191	Usa	Reston	06.02.2019	Read what people are saying about working here. RESPONSIBILITIESThe Citizen Services Mis	Undefined	indeed
06.02.2019 07:22	Research Data Scientist	Computer/Internet	ARUP Laboratories	Salt Lake City	UT	Usa	Salt lake city	06.02.2019	Read what people are saying al		
06.02.2019 07:33	Data Scientist		Perspecta	Washington	DC	Usa	Washington	05.02.2019	Every day at Perspecta, we ena		
06.02.2019 07:38	Senior Data Scientist :: VK (9)		Pyramid Consulting, Inc.	Mclean	VA	Usa	Mclean	05.02.2019	Immediate need for Senior Dat		
06.02.2019 07:39	Senior Data Scientist - W2		Orpine.com	Boston	MA	Usa	Boston	05.02.2019	Work Authorization: Only US C		
06.02.2019 07:39	Data Scientist		Apeiro Technologies	Mclean	VA	Usa	Mclean	05.02.2019	Hi, Hope you are doing great. V		
06.02.2019 07:27	Data Scientist	Computer/Internet	Bank of America	Seattle	WA 98104 (First	Usa	Seattle	06.02.2019	Read what people are saying al		
06.02.2019 07:27	Data Scientist	Computer/Internet	Sallie Mae	Newark	DE	Usa	Newark	06.02.2019	Read what people are saying al		
06.02.2019 07:31	Sr. Data Scientist	Arts/Entertainment/Publishing	Taboola	Los Angeles	CA	Usa	Los angeles	06.02.2019	Read what people are saying al		
06.02.2019 07:27	Data Scientist	Computer/Internet	VideoAmp	Santa Monica	CA	Usa	Santa monica	06.02.2019	Read what people are saying al		
06.02.2019 07:28	Data Scientist	Computer/Internet	Apple	Cambridge	MA	Usa	Cambridge	06.02.2019	Read what people are saying al		
06.02.2019 07:50	Data Scientist		Amtex Enterprises	Plano	TX	Usa	Plano	06.02.2019	Data Scientist 2 years Qualifica		
06.02.2019 07:53	Senior Data Scientist		Sysmind, LLC	Dallas	TX	Usa	Dallas	05.02.2019	Position : Senior Data Scientist		
06.02.2019 07:51	Data Scientist with Sagemaker Experience | NO OF		FocuzMindz	Dallas	TX	Usa	Dallas	05.02.2019	Job Title: Data Scientist with Sa		
06.02.2019 08:06	Data Scientist / Data Engineer		Digital Intelligence Systems, LLC	Houston	TX	Usa	Houston	05.02.2019	Incorporated in 1994, DISYS is c		
06.02.2019 08:03	Data Scientist with Exp in Data Modelling		Advent Global Solutions, Inc.	Bothell	WA	Usa	Bothell	05.02.2019	urgent need for a Data Scientis		
06.02.2019 08:06	Data Scientist		Resource Informatics Group	Houston	TX	Usa	Houston	06.02.2019	We have below urgent positon		
06.02.2019 08:04	Data Scientist needed with solid R and Python skill		Viri Technology	Seattle	WA	Usa	Seattle	06.02.2019	Viri Technology is seeking an u		
06.02.2019 08:08	Level I Solutions Engineer - Data Scientist		Kforce Technology Staffing	Irvine	CA	Usa	Irvine	07.02.2019	RESPONSIBILITIES: Kforce has a		
06.02.2019 08:40	Data Scientist	military	L3 Technologies	Chantilly	VA	Usa	Chantilly	03.02.2019	Data Scientist Â - Â Requisit		
06.02.2019 08:30	Data Scientist	business and financial operations	Apex Systems	West Chester Towns	OH	Usa	West chester	05.02.2019	Description Building large syste		ilder
06.02.2019 08:30	Senior Data Scientist - Tallahassee, FL - $150k-$170	business and financial operations	Jefferson Frank	Tallahassee	FL	Usa	Tallahassee	05.02.2019	My client is a leader in the Mar		ilder
06.02.2019 08:28	Data Scientist	business and financial operations	5 Star Global Recruitment Partners,	Dallas	TX	Usa	Dallas	05.02.2019	Data Scientist Client: A Fortune		ilder
06.02.2019 08:27	Data Scientist	business and financial operations	IQVIA	Plymouth Meeting	PA	Usa	Plymouth meeting	05.02.2019	Join us on our exciting journey		ilder
06.02.2019 08:27	Data Scientist	business and financial operations	Grant Thornton LLP	Alexandria	VA	Usa	Alexandria	05.02.2019	Grant Thornton is seeking a Sen		ilder
06.02.2019 08:28	Data Scientist	business and financial operations	Experis	Seattle	WA	Usa	Seattle	05.02.2019	We are seeking a Data Scientist		ilder
06.02.2019 08:31	Data Scientist Tampa, FL $110-130K	business and financial operations	Jefferson Frank	Tampa	FL	Usa	Tampa	05.02.2019	Data Scientist Tampa, FL $110-1		ilder
06.02.2019 07:48	Data Scientist		Staffing Technologies	Boston	MA	Usa	Boston	01.02.2019	The Team The AIML (Artificial I		
06.02.2019 08:29	Clinical Data Scientist	business and financial operations	Aerotek	Waltham	MA	Usa	Waltham	05.02.2019	POSITION SUMMARY: The Cont		ilder
06.02.2019 09:37	Sr. Mgr, Data Scientist	Engineering/Architecture	Kraft Heinz Company	San Francisco	CA	Usa	San francisco	06.02.2019	Read what people are saying al		
06.02.2019 09:37	Data Scientist Intern, Engineering - Software Devel	Engineering/Architecture	Criteo	Palo Alto	CA 94301 (Profe	Usa	Palo alto	06.02.2019	Read what people are saying al		
06.02.2019 09:49	Data Scientist	Manufacturing/Mechanical	Comcast	Philadelphia	PA 19103	Usa	Philadelphia	06.02.2019	Read what people are saying about working here. Comcast brings together the best in medi	Undefined	indeed
06.02.2019 09:45	Principal Data Scientist (REMOTE)	Engineering/Architecture	FireEye, Inc.	Boston	VA 20190	Usa	Boston	06.02.2019	Read what people are saying about working here. Job DescriptionIf you dream of a job worth	Undefined	indeed

"Read what people are saying about working here.

RESPONSIBILITIES

The Citizen Services Mission Area (CS) provides guidance and support to its client

CS is building a behavioral analytics team to support a wide variety of client ne

This position is for a Software Developer with exposure to primary to SAS, Extract

QUALIFICATIONS

Bachelor's degree in Computer Science, Info Systems, Math, Physics, Computer Engi

Minimum 7-10 years of professional experience in software engineering and develop

Database development and application integration experience.

Advanced knowledge of SAS, C ##, Java, Perl, Python, Scala and r programming

Experienced user of SAS or equivalent training in SAS or other similar tool

Familiarity with Hadoop specifically Spark and Spark Streaming

Familiarity with SAS to Hadoop File System integration

OVERVIEW

Noblis and Noblis ESI are solving difficult problems that help our government and

Why work at a Noblis company?

Our employees find greater meaning in their work and balance the other things in

Noblis has won numerous workplace awards. Over the past two decades, Noblis has

1. Best Employer: We have been a Washington Post â€œTop Workplace for 5 consecuti

2. Business Ethics and Integrity: A â€œWorldâ€™s Most Ethical Company for 7 years

3. Leadership and Innovation: CEO Amr ElSawy selected to Executive Mosaicâ€™s ann

Noblis maintains a drug-free workplace and is an equal opportunity employer.

Noblis, Inc. is a nonprofit science, technology, and strategy organization that

Finding the right category

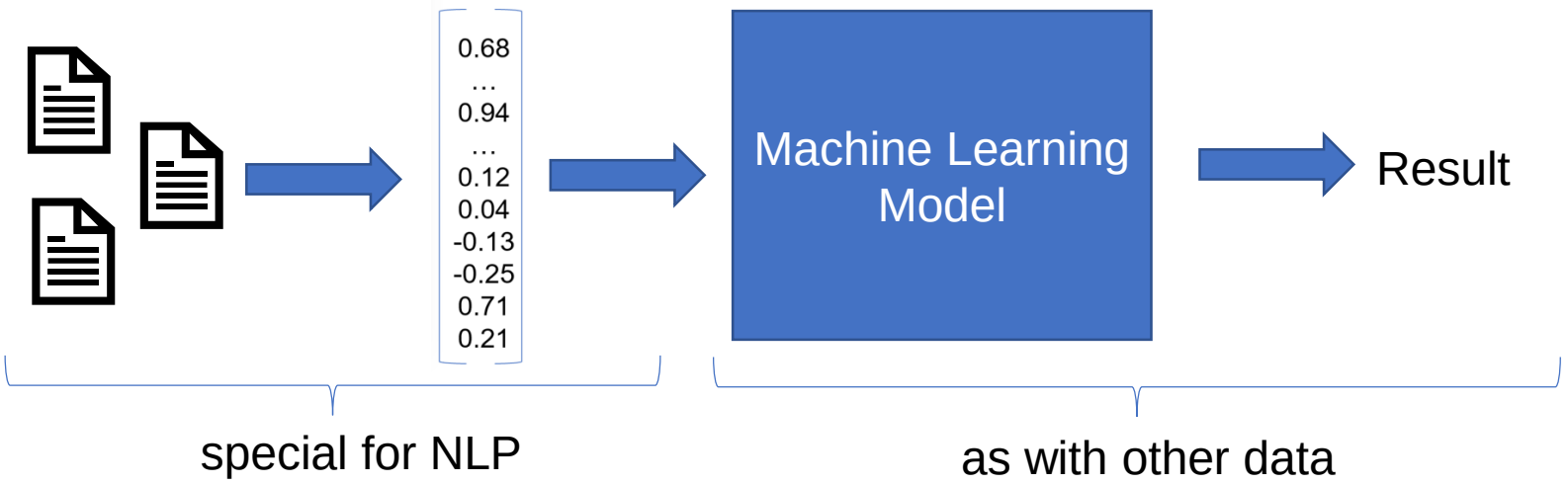
Hochschule Offenburg
offenburg.university

crawl_timestamp	job_title	category	company_name	city	state	country	inferred_city	post_date	job_description	job_type	job_board
06.02.2019 06:26	Enterprise Data Scientist I	Accounting/Finance	Farmers Insurance Group	Woodland Hills	CA	Usa	Woodland hills	06.02.2019	Read what people are saying about working here. We are Farmers!Join a team of diverse pr	Undefined	indeed
06.02.2019 06:33	Data Scientist		Luxoft USA Inc	Middletown	NJ	Usa	Middletown	05.02.2019	We have an immediate opening for a Sharp Data Scientist with a strong Mathematical/Statis	Undefined	dice
06.02.2019 06:33	Data Scientist		Cincinnati Bell Technology Solutions	New York	NY	Usa	New york	05.02.2019	Candidates should have the following background, skills and characteristics: Â Experience de	Full Time	dice
06.02.2019 06:33	Data Scientist, Aladdin Wealth Tech, Associate (Mc	Accounting/Finance	BlackRock	New York	NY 10055 (Midto	Usa	New york	06.02.2019	Read what people are saying about working here. About BlackRockBlackRock helps investors	Undefined	indeed
06.02.2019 06:48	Senior Data Scientist	biotech	CyberCoders	Charlotte	NC	Usa	Charlotte	05.02.2019	We are seeking an extraordinary Data Scientist in Charlotte to join our fast growing healthca	Full Time	monster
06.02.2019 06:36	CIB â€“ Fixed Income Research â€“ Machine Learni	Accounting/Finance	JP Morgan Chase	New York	NY 10179 (Midto	Usa	New york	05.02.2019	Read what people are saying about working here. OpportunityThe opportunity is to join our	Undefined	indeed
06.02.2019 06:34	Data Scientist, Licensing Operations	Accounting/Finance	Spotify	New York	NY 10011 (Chels	Usa	New york	06.02.2019	Read what people are saying about working here. At Spotify our mission is to provide the wo	Undefined	indeed
06.02.2019 06:52	Sr. Data Scientist (Can work on Xoriant W2)		Xoriant Corporation	Santa Clara	CA	Usa	Santa clara	06.02.2019	Job Title: - Sr. Data Science Consultant Duration: 1+ Yrs (will get extend) Mode of interview -	Contract	dice
06.02.2019 06:34	Data Scientist, Aladdin Wealth Tech, Associate	Accounting/Finance	BlackRock	New York	NY 10055 (Midto	Usa	New york	06.02.2019	Read what people are saying about working here. About BlackRockBlackRock helps investors	Undefined	indeed
06.02.2019 07:03	Data Scientist		Adroit Resources	San Francisco	CA	Usa	San francisco	05.02.2019	â€¢ 3+ years related a professional experienceÂ â€¢ Proven achievements resulting from pr	Contract	dice
06.02.2019 07:19	Data Scientist	Computer/Internet	Northrop Grumman	Monterey	CA 93940	Usa	Monterey	06.02.2019	Read what people are saying about working here. At Northrop Grumman, innovation isn't ju	Undefined	indeed
06.02.2019 07:24	Data Scientist		Envoy Consulting Group, Inc	Reston	VA	Usa	Reston	05.02.2019	Data Scientist EnvoyIT is looking for a Data Scientist for a Full-time position with one of our k	Full Time	dice
06.02.2019 07:21	ETL Developer / Data Scientist	Computer/Internet	Noblis	Reston	VA 20191	Usa	Reston	06.02.2019	Read what people are saying about working here. RESPONSIBILITIESThe Citizen Services Mis	Undefined	indeed
06.02.2019 07:22	Research Data Scientist	Computer/Internet	ARUP Laboratories	Salt Lake City	UT	Usa	Salt lake city	06.02.2019	Read what people are saying about working here. Job DetailsDescriptionSchedule:Monday -	Undefined	indeed
06.02.2019 07:33	Data Scientist		Perspecta	Washington	DC	Usa	Washington	05.02.2019	Every day at Perspecta, we enable hundreds of thousands of people to take on our nation s	Full Time	dice
06.02.2019 07:38	Senior Data Scientist :: VK (9)		Pyramid Consulting, Inc.	Mclean	VA	Usa	Mclean	05.02.2019	Immediate need for Senior Data Scientist with experience in the IT Industry. This is 12 Mont	Undefined	dice
06.02.2019 07:39	Senior Data Scientist - W2		Orpine.com	Boston	MA	Usa	Boston	05.02.2019	Work Authorization: Only US Citizen's & Green CardÂ Â Job description:Â Â â€¢ Business In	Contract	dice
06.02.2019 07:39	Data Scientist		Apeiro Technologies	Mclean	VA	Usa	Mclean	05.02.2019	Hi, Hope you are doing great. We are having requirement forÂ Data Scientist with one of ou	Contract	dice
06.02.2019 07:27	Data Scientist	Computer/Internet	Bank of America	Seattle	WA 98104 (First	Usa	Seattle	06.02.2019	Read what people are saying about working here. Job Description:Position SummaryData Sc	Undefined	indeed
06.02.2019 07:27	Data Scientist	Computer/Internet	Sallie Mae	Newark	DE	Usa	Newark	06.02.2019	Read what people are saying about working here. At Operations strategy and Analytics, we a	Undefined	indeed
06.02.2019 07:31	Sr. Data Scientist	Arts/Entertainment/Publishing	Taboola	Los Angeles	CA	Usa	Los angeles	06.02.2019	Read what people are saying about working here. Read something interesting online today?	Undefined	indeed
06.02.2019 07:27	Data Scientist	Computer/Internet	VideoAmp	Santa Monica	CA	Usa	Santa monica	06.02.2019	Read what people are saying about working here. About UsVideoAmpâ€™s mission is to bri	Undefined	indeed
06.02.2019 07:28	Data Scientist	Computer/Internet	Apple	Cambridge	MA	Usa	Cambridge	06.02.2019	Read what people are saying about working here. SummaryPosted: Feb 5, 2019Weekly Hour	Undefined	indeed
06.02.2019 07:50	Data Scientist		Amtex Enterprises	Plano	TX	Usa	Plano	06.02.2019	Data Scientist 2 years Qualifications and Skills: Â·Â Â Â Â Â Â Â 3+ years work experience in a b	Contract	dice
06.02.2019 07:53	Senior Data Scientist		Sysmind, LLC	Dallas	TX	Usa	Dallas	05.02.2019	Position : Senior Data Scientist Location : Dallas, TX Job Description: 8-10 Years ExperienceÂ	Undefined	dice
06.02.2019 07:51	Data Scientist with Sagemaker Experience | NO OP		FocuzMindz	Dallas	TX	Usa	Dallas	05.02.2019	Job Title: Data Scientist with Sagemaker Experience | NO OPTâ€™s please Work Location & F	Contract	dice
06.02.2019 08:06	Data Scientist / Data Engineer		Digital Intelligence Systems, LLC	Houston	TX	Usa	Houston	05.02.2019	Incorporated in 1994, DISYS is one of the largest IT staffing firms in the US; DISYS has nearly c	Contract	dice
06.02.2019 08:03	Data Scientist with Exp in Data Modelling		Advent Global Solutions, Inc.	Bothell	WA	Usa	Bothell	05.02.2019	urgent need for a Data Scientist, Essentially, the guidance was, send resumes and weâ€™ll p	Contract	dice
06.02.2019 08:06	Data Scientist		Resource Informatics Group	Houston	TX	Usa	Houston	06.02.2019	We have below urgent positon in TX. Below are the more details on it. Let me know your int	Contract	dice
06.02.2019 08:04	Data Scientist needed with solid R and Python skill		Viri Technology	Seattle	WA	Usa	Seattle	06.02.2019	Viri Technology is seeking an up and coming Data Scientist for an immediate FTE role with a	Full Time	dice
06.02.2019 08:08	Level I Solutions Engineer - Data Scientist		Kforce Technology Staffing	Irvine	CA	Usa	Irvine	07.02.2019	RESPONSIBILITIES: Kforce has a client seeking a Level I Solutions Engineer - Data Scientist in	Full Time	dice
06.02.2019 08:40	Data Scientist	military	L3 Technologies	Chantilly	VA	Usa	Chantilly	03.02.2019	Data Scientist Â - Â Requisition ID Â 105274 USA-Virginia-Chantilly Â Â Description	Undefined	monster
06.02.2019 08:30	Data Scientist			West Chester Towns	OH	Usa	West chester	05.02.2019	Description Building large systematic reports and one-off small pieces of exploratory analys	Full Time	careerbuilder
06.02.2019 08:30	Senior Data Scientis			Tallahassee	FL	Usa			has operations in multiple states acro	Full Time	careerbuilder
06.02.2019 08:28	Data Scientist			rtners, I Dallas	TX	Usa			national motor vehicle industry leader.	Full Time	careerbuilder
06.02.2019 08:27	Data Scientist			Plymouth Meeting	PA	Usa			an Data Science Companyâ„¢, focused	Full Time	careerbuilder
06.02.2019 08:27	Data Scientist	business and financial operations	Grant Thornton LLP	Alexandria	VA	Usa			s Advanced Digital Technology and Ana	Full Time	careerbuilder
06.02.2019 08:28	Data Scientist	business and financial operations	Experis	Seattle	WA	Usa			to hire role with a leading client in Se	Contract	careerbuilder
06.02.2019 08:31	Data Scientist Tampa, FL $110-130K	business and financial operations	Jefferson Frank	Tampa	FL	Usa	Tampa	05.02.2019	Data Scientist Tampa, FL $110-130K Job Description - A highly respected client is seeking a sk	Full Time	careerbuilder
06.02.2019 07:48	Data Scientist		Staffing Technologies	Boston	MA	Usa	Boston	01.02.2019	The Team The AIML (Artificial Intelligence and Machine Learning) team contributes to the vi	Full Time	dice
06.02.2019 08:29	Clinical Data Scientist	business and financial operations	Aerotek	Waltham	MA	Usa	Waltham	05.02.2019	POSITION SUMMARY: The Contract Clinical Data Scientist in CDS is responsible for supporting	Full Time	careerbuilder
06.02.2019 09:37	Sr. Mgr, Data Scientist	Engineering/Architecture	Kraft Heinz Company	San Francisco	CA	Usa	San francisco	06.02.2019	Read what people are saying about working here. Our CompanyThe Kraft Heinz Company is	Undefined	indeed
06.02.2019 09:37	Data Scientist Intern, Engineering - Software Devel	Engineering/Architecture	Criteo	Palo Alto	CA 94301 (Profe	Usa	Palo alto	06.02.2019	Read what people are saying about working here. InternshipWho we areAt Criteo, we conne	Undefined	indeed
06.02.2019 09:49	Data Scientist	Manufacturing/Mechanical	Comcast	Philadelphia	PA 19103	Usa	Philadelphia	06.02.2019	Read what people are saying about working here. Comcast brings together the best in medi	Undefined	indeed
06.02.2019 09:45	Principal Data Scientist (REMOTE)	Engineering/Architecture	FireEye, Inc.	Boston	VA 20190	Usa	Boston	06.02.2019	Read what people are saying about working here. Job DescriptionIf you dream of a job worth	Undefined	indeed

Target: category

Given: job description

Classification / Clustering of Text Data



Term-document count matrices

- Consider the number of occurrences of a term in a document:
  - Each document is a count vector in $\mathbb{N}^V$: a column below

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
Caesar	232	227	0	2	1	1
Calpurnia	0	10	0	0	0	0
Cleopatra	57	0	0	0	0	0
mercy	2	0	3	5	5	1
worser	2	0	1	1	1	0

Bag of words model

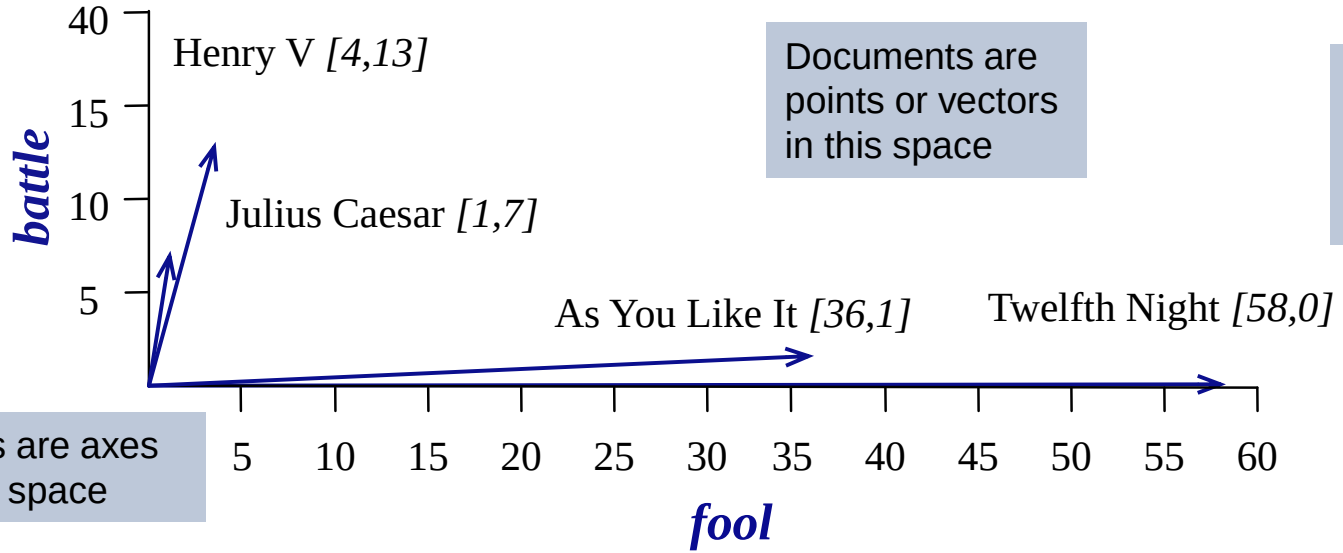
Hochschule Offenburg
offenburg.university

- Vector representation doesn't consider the ordering of words in a document
- *John is quicker than Mary* and *Mary is quicker than John* have the same vectors
- This is called the bag of words model.

Document vectors visualized

Hochschule Offenburg
offenburg.university

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3



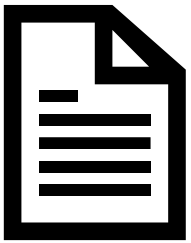
Documents are points or vectors in this space

High-dimensional space! (here: only 2 dimensions shown)

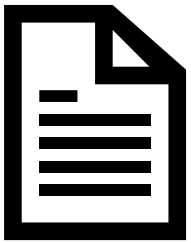
Terms are axes of the space

Term frequency tf

- The term frequency $tf_{t,d}$ of term *t* in document *d* is defined as the number of times that *t* occurs in *d*.
- Compute query-document match score



search term
count = 1



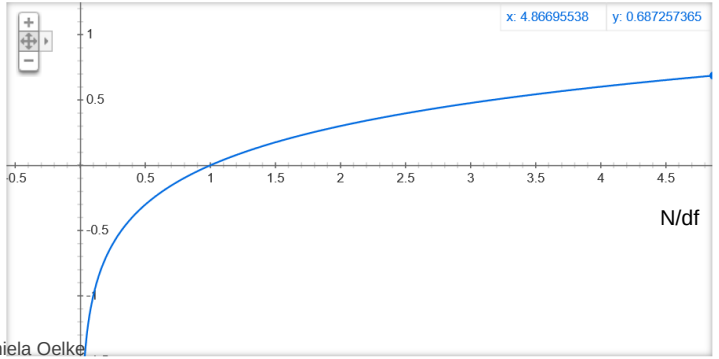
search term
count = 10

More relevant 10x more relevant?

idf weight

- *N* is the number of documents in the collection
- df_t is the document frequency of *t*: the number of documents that contain *t*
  - df_t is an inverse measure of the informativeness of *t*
  - $df_t \leq N$
- We define the idf (inverse document frequency) of *t* by

$$\text{idf}_t = \log_{10}(N/\text{df}_t)$$



  - We use log (N/df_t) instead of N/df_t to "dampen" the effect of idf.

TF-IDF weighting

$$w_{t,d} = tf_{t,d}\ x \log_{10}(\frac{N}{df_t})$$

When does the tf-idf value of a term increase?

- with the number of occurrences of the term within a document
- with the rarity of the term in the collection

Variants of how TF-IDF values are calculated exist

Cosine similarity

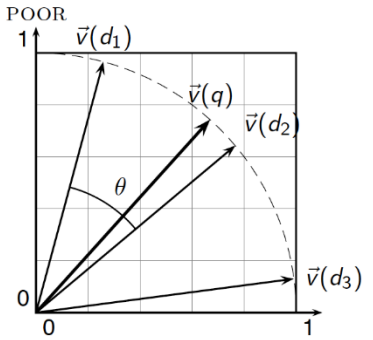
$$\text{cosine}(\vec{v},\vec{w}) = \frac{\vec{v}\cdot\vec{w}}{|\vec{v}||\vec{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

If the vectors are length-normalized:

$$\text{cosine}(\vec{v},\vec{w}) = \sum_{i=1}^{N} v_i w_i$$

v_i is the tf-idf value for word i in document v
w_i is the tf-idf value for word i in document w

Cos(*v*,*w*) is the cosine similarity of *v* and *w*



Hochschule Offenburg
offenburg.university

What‘s next?
Word Embeddings & Co.

How similar are the following sentences?

1. Text is read in as an array of characters.
2. Sentences are passed as a list of strings.
3. Peter is known as a master of chess.

Fill in the missing word

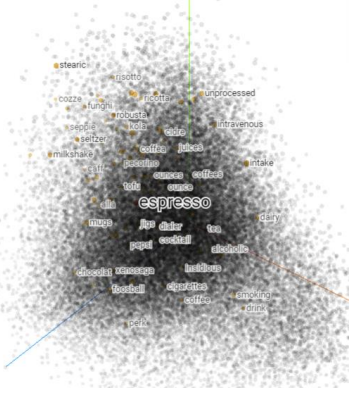
Hochschule Offenburg
offenburg.university

- Ich bin mit dem … nach Offenburg gereist.

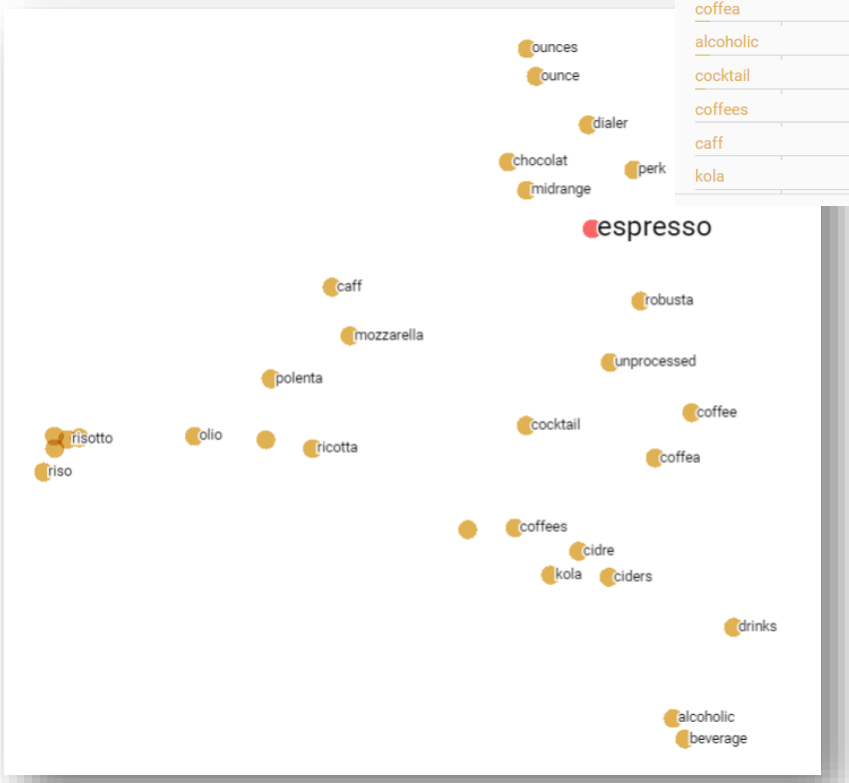
Word2Vec

- Word Embedding = Representation of words as vectors in an n-dimensional space
- Similar Words are located close to each other in the feature space

Caveat:
Projection from 200D → 3D!



Nearest neighbors of „espresso“



Nearest points in the original space:

coffee	0.885
robusta	0.927
coffea	0.956
alcoholic	0.958
cocktail	0.970
coffees	1.010
caff	1.019
kola	1.029

espresso

0.68
…
0.94
…
0.12
0.04
-0.13
-0.25
0.71
0.21

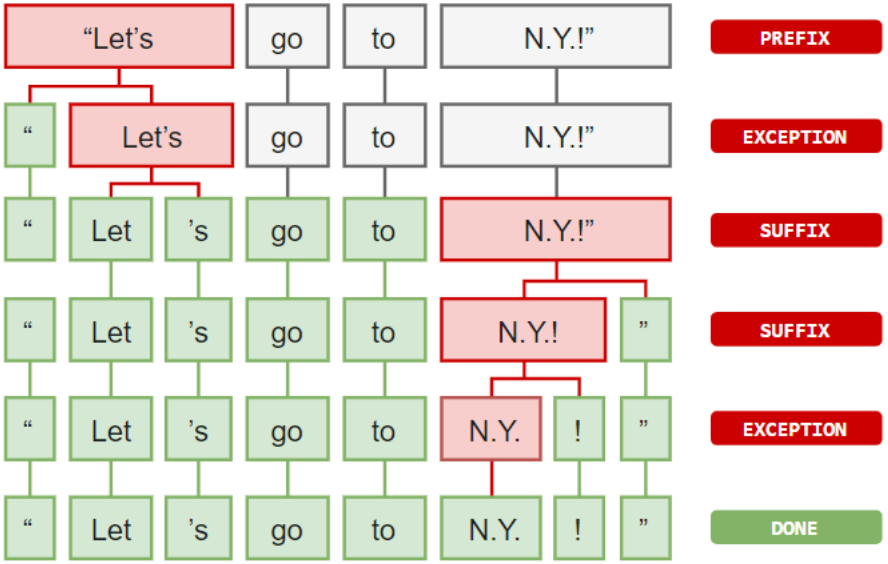
https://projector.tensorflow.org/

Hochschule Offenburg
offenburg.university

Normalization of Natural Language Texts

Tokenization

- Tokenization: The task of converting a text from a single string to a list of tokens.
- It is harder that it seems!



Source: Spacy.io

Stopword Removal

Hochschule Offenburg
offenburg.university

```
i
me
my
myself
we
our
ours
ourselves
you
your
yours
yourself
yourselves
he
him
his
himself
she
her
hers
herself
it
its
itself
```

```
they
them
their
theirs
themselves
what
which
who
whom
this
that
these
those
am
is
are
was
were
be
been
being
have
has
had
```

```
having
do
does
did
doing
a
an
the
and
but
if
or
because
as
until
while
of
at
by
for
with
about
against
between
```

```
until
while
of
at
by
for
with
about
against
between
into
through
during
before
after
above
below
to
from
up
down
in
out
on
```

```
off
over
under
again
further
then
once
here
there
when
where
why
how
all
any
both
each
few
more
most
other
some
such
no
```

```
nor
not
only
own
same
so
than
too
very
s
t
can
will
just
don
should
now
```

Customization of stopword lists may be important for some applications.

English stopwords in NLTK
https://gist.github.com/sebleier/554280

Word normalization

- Putting words / tokens in a standard format
  e.g.
  Usa, US, USA → USA
  uh-huh, uhhuh → uhhuh

- Mapping everything to lower case

Stem and Lemma

Hochschule Offenburg
offenburg.university

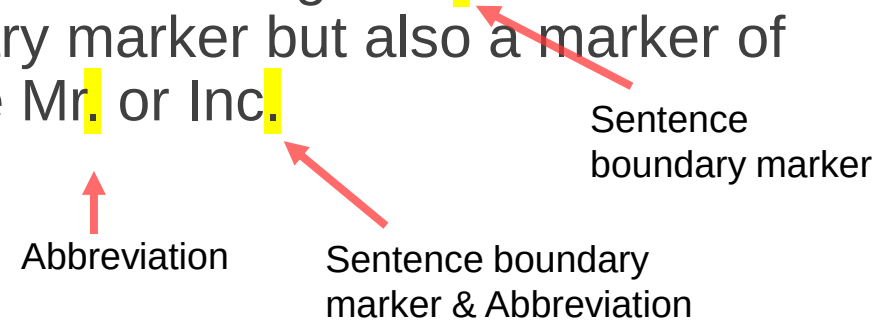
Word	Inflection	Stem	Morphological information	Lemma
study	-y	stud	Infinitive of the verb “study”	study
studies	-ies	stud	Third person, singular, Present Simple of the verb “study”	study
studying	-ing	stud	Gerund of the verb “study”	study

https://chatbotsmagazine.com/how-to-use-nlp-for-building-a-chatbot-fac05476a58e?source=---------9----------&gi=12a9d7ac3f50

Sentence Segmentation

- Cues are punctuation, like periods, question marks, and exclamation marks
- But: Periods are ambiguous!

The period character is ambiguous. It can be a sentence boundary marker but also a marker of abbreviations like Mr. or Inc.



Hochschule Offenburg
offenburg.university

Feature Extraction from Natural Language Text

Parts of speech (POS)
aka word classes, syntactic categories

Hochschule Offenburg
offenburg.university

- Noun
- Verb
- Adjective / Adverb

open class types

- Pronoun
- Preposition
- Conjunction
- Participle
- Article

closed class types

Terminology

- **Tagging**
  - The process of associating labels with each token in a text
- **Tags**
  - The labels
- **Tag Set**
  - The collection of tags used for a particular task
  - Tag Sets are defined by linguists and not unambiguous. Different languages require different tag sets -> difficult to compare tags of different languages.

Examples

The/DT grand/JJ jury/NN commented/VBD on/IN a/DT number/NN of/IN other/JJ topics/NNS ./.

There/EX are/VBP 70/CD children/NNS there/RB

Preliminary/JJ findings/NNS were/VBD reported/VBN in/IN today/NN 's/POS New/NNP England/NNP Journal/NNP of/IN Medicine/NNP ./.

Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coordinating conjunction	*and, but, or*	PDT	predeterminer	*all, both*	VBP	verb non-3sg present	*eat*
CD	cardinal number	*one, two*	POS	possessive ending	*'s*	VBZ	verb 3sg pres	*eats*
DT	determiner	*a, the*	PRP	personal pronoun	*I, you, he*	WDT	wh-determ.	*which, that*
EX	existential 'there'	*there*	PRP$	possess. pronoun	*your, one's*	WP	wh-pronoun	*what, who*
FW	foreign word	*mea culpa*	RB	adverb	*quickly*	WP$	wh-possess.	*whose*
IN	preposition/ subordin-conj	*of, in, by*	RBR	comparative adverb	*faster*	WRB	wh-adverb	*how, where*
JJ	adjective	*yellow*	RBS	superlatv. adverb	*fastest*	$	dollar sign	*$*
JJR	comparative adj	*bigger*	RP	particle	*up, off*	#	pound sign	*#*
JJS	superlative adj	*wildest*	SYM	symbol	*+,%, &*	"	left quote	' or "
LS	list item marker	*1, 2, One*	TO	"to"	*to*	"	right quote	' or "
MD	modal	*can, should*	UH	interjection	*ah, oops*	(	left paren	[, (, {, <
NN	sing or mass noun	*llama*	VB	verb base form	*eat*	)	right paren	],), }, >
NNS	noun, plural	*llamas*	VBD	verb past tense	*ate*	,	comma	,
NNP	proper noun, sing.	*IBM*	VBG	verb gerund	*eating*	.	sent-end punc	. ! ?
NNPS	proper noun, plu.	*Carolinas*	VBN	verb past part.	*eaten*	:	sent-mid punc	: ; ... – -

Figure 8.1 Penn Treebank part-of-speech tags (including punctuation).

Some of the best-known Tagsets

- Brown corpus: 87 tags
  - (more when tags are combined)
- Penn Treebank: 45 tags
- Lancaster UCREL C5 (used to tag the BNC): 61 tags
- Lancaster C7: 145 tags

For the German language:

- Stuttgart-Tübingen Tagset (STTS): 54 tags

Syntactic parsing

Hochschule Offenburg
offenburg.university

Syntactic parsing is the task of recognizing a sentence and assigning a syntactic structure to it.

Comparison of POS tagging, chunking and full parsing

Hochschule Offenburg
offenburg.university

He reckons the current account deficit will narrow to only 2 billion in September.

Output of POS tagger:

He\PRP reckons\VBZ the\DT current\JJ account\NN deficit\NN will\MD narrow\VB to\TO only\RB 2\CD billion\CD in\IN September\NNP .\.

Comparison of POS tagging, chunking and full parsing

Hochschule Offenburg
offenburg.university

He reckons the current account deficit will narrow to only 2 billion in September.

Output of chunker:

[NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only 2 billion] [PP in] [NP September].

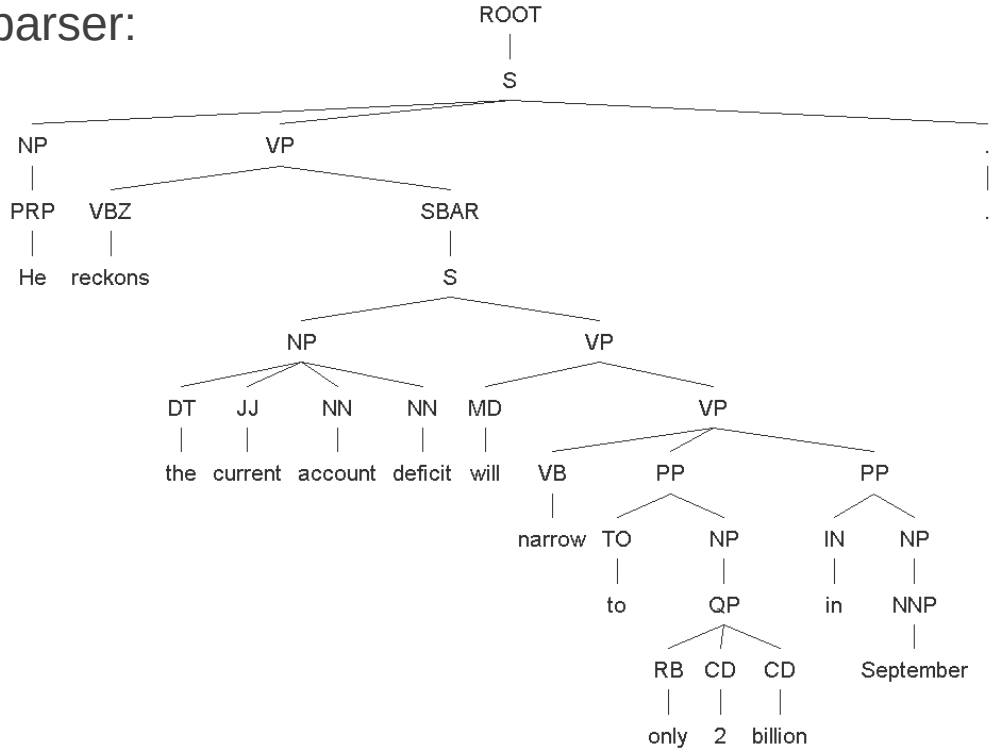
What about German?
Determine the constituents *in the following sentence:*

„Ich sah gestern Abend einen schönen Film im neuen Kino unserer Stadt.“

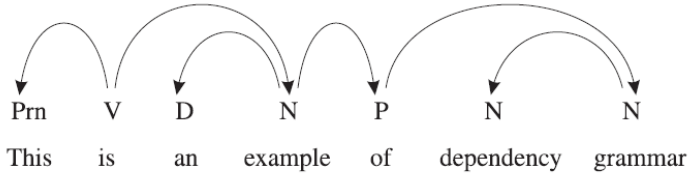
Comparison of POS tagging, chunking and full parsing

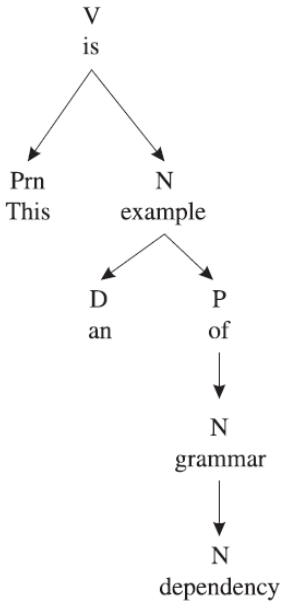
Hochschule Offenburg
offenburg.university

Output of full parser:



Dependency Parsing





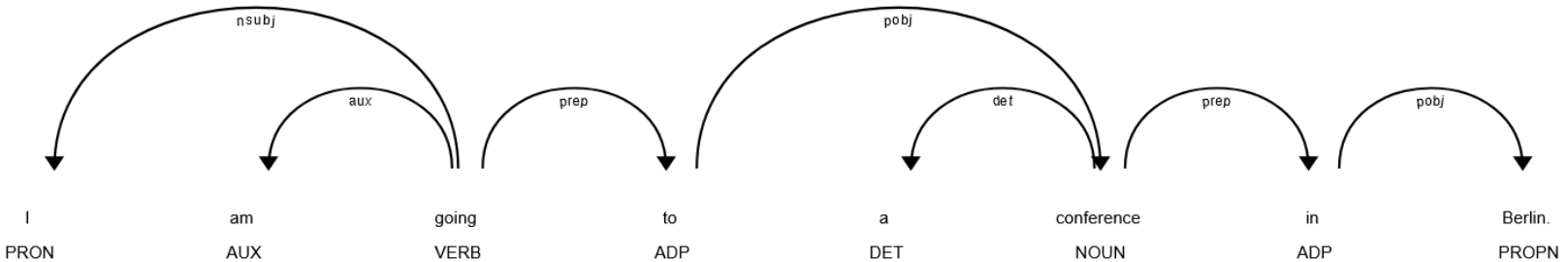
Example: Flight booking dialogue system

Hochschule Offenburg
offenburg.university

- I am going to a conference in Berlin

Dialogue system: Where do you want to fly to?

Hochschule Offenburg
offenburg.university

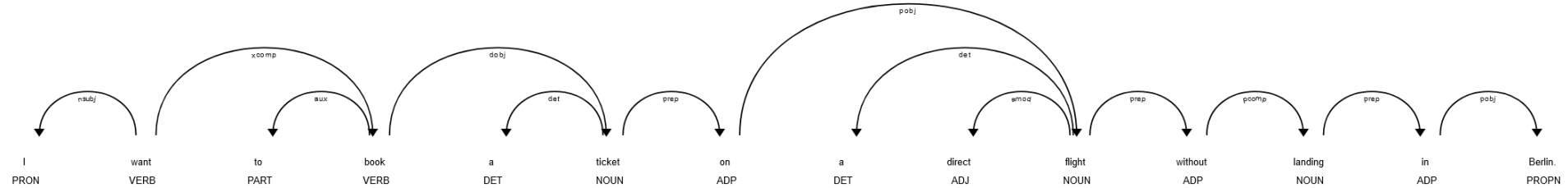


GPE

Dialogue system: Where do you want to fly to?

Hochschule Offenburg
offenburg.university

- I want to book a ticket on a direct flight without landing in Berlin

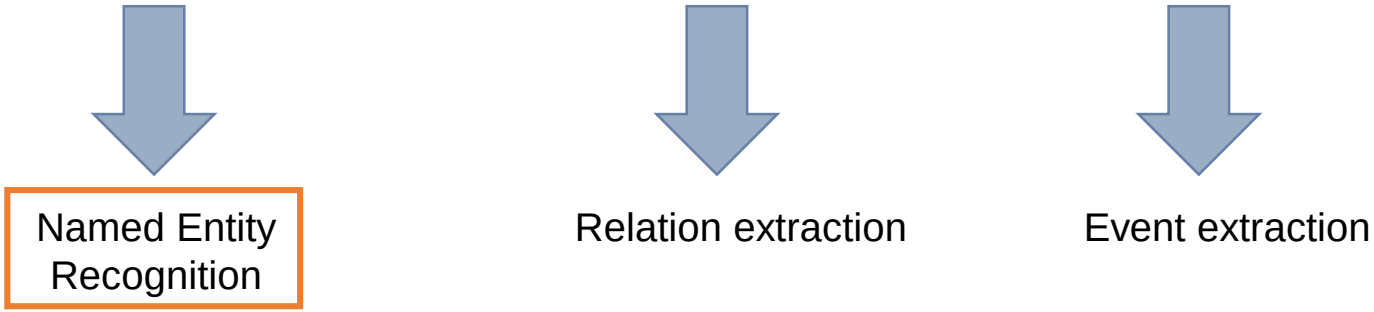


Hochschule Offenburg
offenburg.university

Information Extraction

Information Extraction

Information Extraction (IE) is the **process of extracting structured information** (e.g., database tables) from unstructured machine-readable documents (e.g., Web documents).



Named Entities

- Roughly speaking anything that can be referred to with a proper name
- E.g., person, location, organization
- Commonly also dates, times, numerical expressions, …
- Can also include domain-specific entities like protein names or commercial products

Example

Citing high fuel prices, [ORG **United Airlines**] said [TIME **Friday**] it has increased fares by [MONEY **$6**] per round trip on flights to some cities also served by lower-cost carriers. [ORG **American Airlines**], a unit of [ORG **AMR Corp.**], immediately matched the move, spokesman [PER **Tim Wagner**] said. [ORG **United**], a unit of [ORG **UAL Corp.**], said the increase took effect [TIME **Thursday**] and applies to most routes where it competes against discount carriers, such as [LOC **Chicago**] to [LOC **Dallas**] and [LOC **Denver**] to [LOC **San Francisco**].

A list of generic named entity types

Type	Tag	Sample Categories	Example sentences
People	PER	people, characters	**Turing** is a giant of computer science.
Organization	ORG	companies, sports teams	The **IPCC** warned about the cyclone.
Location	LOC	regions, mountains, seas	The **Mt. Sanitas** loop is in **Sunshine Canyon**.
Geo-Political Entity	GPE	countries, states, provinces	**Palo Alto** is raising the fees for parking.
Facility	FAC	bridges, buildings, airports	Consider the **Golden Gate Bridge**.
Vehicles	VEH	planes, trains, automobiles	It was a classic **Ford Falcon**.

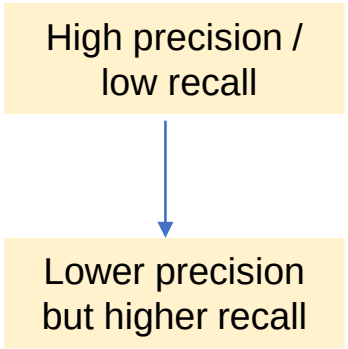
How would you approach the problem?

Hochschule Offenburg
offenburg.university

- If you would want to recognize names of **locations**, what would you be looking for?
- If you would want to recognize **product names**, like DCT-154, CT-9914, etc. what would you be looking for?

Rule-based NER

- Commercial approaches often are based on pragmatic combinations of lists and rules (with smaller amount of ML)
- Often repeated rule-based passes over a text
  → Output of one pass becomes input of next one



Hochschule Offenburg
offenburg.university

Natural Language Processing with spaCy

EMI Elektrotechnik, Medizintechnik und Informatik

Acknowledgements

The following slides are based on material provided by

- spacy.io

What‘s spaCy?

- a free, open-source library for NLP in Python
- can be used for
  - text preprocessing
  - information extraction
  - natural language understanding systems

spaCy vs. nltk

Hochschule Offenburg
offenburg.university

nltk

- created with focus on teaching and research
- user can choose between multiple algorithms with equivalent functionality

spaCy

- created with focus on production-use
- choice has been made by spaCy-developers (easier for developers + performance-optimization)

Features

NAME	DESCRIPTION
Tokenization	Segmenting text into words, punctuations marks etc.
Part-of-speech (POS) **Tagging**	Assigning word types to tokens, like verb or noun.
Dependency Parsing	Assigning syntactic dependency labels, describing the relations between individual tokens, like subject or object.
Lemmatization	Assigning the base forms of words. For example, the lemma of “was” is “be”, and the lemma of “rats” is “rat”.
Sentence Boundary Detection (SBD)	Finding and segmenting individual sentences.
Named Entity Recognition (NER)	Labelling named “real-world” objects, like persons, companies or locations.
Entity Linking (EL)	Disambiguating textual entities to unique identifiers in a Knowledge Base.
Similarity	Comparing words, text spans and documents and how similar they are to each other.
Text Classification	Assigning categories or labels to a whole document, or parts of a document.
Rule-based Matching	Finding sequences of tokens based on their texts and linguistic annotations, similar to regular expressions.
Training	Updating and improving a statistical model’s predictions.
Serialization	Saving objects to files or byte strings.

Statistical (ML) models

Hochschule Offenburg
offenburg.university

- Some features require a **statistical model** to be loaded
  → **prediction** of linguistic annotations

Small, default models

Hochschule Offenburg
offenburg.university

Small, default models include

- **Binary weights** (to predict annotations in context)
- **Lexical entries** (words and context-independent attributes)
- **Data files** (e.g. lemmatization rules and lookup tables)
- **Word vectors** (meaning representations)
- **Configuration options**

Which statistical models are available?

Hochschule Offenburg
offenburg.university

- Different languages (as of 9/2020: 16)
- Trained on different corpora (web, news, …)
- Different sizes, speed, memory usage and accuracy

→You have to choose and load the right model!

https://spacy.io/usage/models

How do I use models?

1. Download and install the required model (It‘s a Python package) https://spacy.io/usage/models

```
# Download best-matching version of specific model for your spaCy installat
python -m spacy download en_core_web_sm

# Out-of-the-box: download best-matching default model and create shortcut
python -m spacy download en

# Download exact model version (doesn't create shortcut link)
python -m spacy download en_core_web_sm-2.2.0 --direct
```

How do I use models? (2)

Hochschule Offenburg
offenburg.university

2. Loading models

```
import spacy

# load model package "en_core_web_sm"
nlp = spacy.load("en_core_web_sm")

# load package from a directory
nlp = spacy.load("/path/to/en_core_web_sm")

# load model with shortcut link "en"
nlp = spacy.load("en")
```

How do I use models? (3)

Hochschule Offenburg
offenburg.university

3. Using the model

```
nlp = spacy.load(„...“)
```

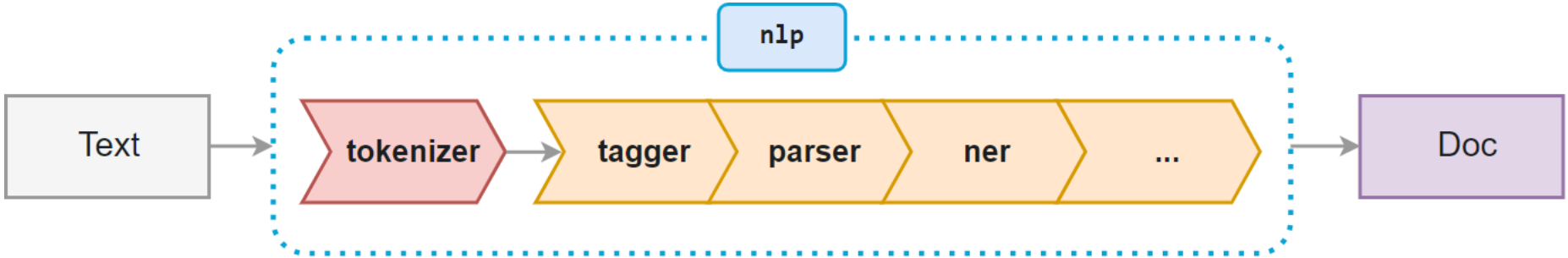
Returns a Language object, usually called „nlp“

```
doc = nlp("This is a sentence.")
```

Returns a Doc object

Can be called with a string.

Processing pipeline



Lemmatization

```
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_)
```

Text: The original word text.

Lemma: The base form of the word.

TEXT	LEMMA
Apple	apple
is	be
looking	look
at	at
buying	buy
U.K.	u.k.
startup	startup
for	for
$	$
1	1
billion	billion

Part-of-speech (POS) tagging


```python
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_)
```

TEXT	LEMMA	POS	TAG
Apple	apple	PROPN	NNP
is	be	AUX	VBZ
looking	look	VERB	VBG
at	at	ADP	IN
buying	buy	VERB	VBG
U.K.	u.k.	PROPN	NNP
startup	startup	NOUN	NN
for	for	ADP	IN
$	$	SYM	$
1	1	NUM	CD
billion	billion	NUM	CD

POS: The simple UPOS part-of-speech tag.

Tag: The detailed part-of-speech tag.

Dependency Parsing

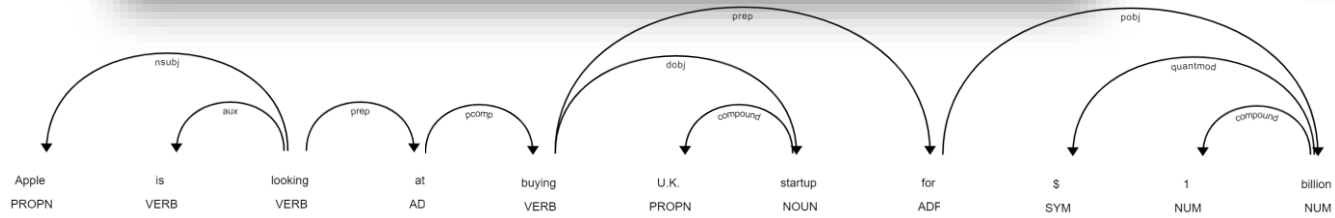
Hochschule Offenburg
offenburg.university

Dep: Syntactic dependency, i.e. the relation between tokens.

```
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_)
```




```
from spacy import displacy
displacy.render(doc, style = "dep")
```

TEXT	LEMMA	POS	TAG	DEP
Apple	apple	PROPN	NNP	nsubj
is	be	AUX	VBZ	aux
looking	look	VERB	VBG	ROOT
at	at	ADP	IN	prep
buying	buy	VERB	VBG	pcomp
U.K.	u.k.	PROPN	NNP	compound
artup	startup	NOUN	NN	dobj
r	for	ADP	IN	prep
	$	SYM	$	quantmod
	1	NUM	CD	compound
billion	billion	NUM	CD	pobj

More features


```
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)
```

Shape: The word shape – capitalization, punctuation, digits.

is alpha: Is the token an alpha character?

is stop: Is the token part of a stop list, i.e. the most common words of the language?

TEXT	LEMMA	POS	TAG	DEP	SHAPE	ALPHA	STOP
Apple	apple	PROPN	NNP	nsubj	Xxxxx	True	False
is	be	AUX	VBZ	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
1	1	NUM	CD	compound	d	False	False
billion	billion	NUM	CD	pobj	xxxx	True	False

Named Entities

Named Entities = „real-world objects“ that have been assigned a name (i.e. a person, a country, a product, a book title…)

Hochschule Offenburg
offenburg.university

```
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.

```
from spacy import displacy
displacy.render(doc, style = "ent")
```

Apple ORG is looking at buying U.K. GPE startup for $1 billion MONEY

Result visualized with displaCy visualizer

Word Vectors / Word embeddings

Hochschule Offenburg
offenburg.university

- multi-dimensional meaning representation of a word
- e.g. generated with word2vec
- only available in larger models!
- available on the level of Tokens, Documents and Spans
- out-of vocabulary words do not have a word vector

BANANA.VECTOR

```
array([2.02280000e-01,  -7.66180009e-02,   3.70319992e-01,
       3.28450017e-02,  -4.19569999e-01,   7.20689967e-02,
      -3.74760002e-01,   5.74599989e-02,  -1.24009997e-02,
       5.29489994e-01,  -5.23800015e-01,  -1.97710007e-01,
      -3.41470003e-01,   5.33169985e-01,  -2.53309999e-02,
       1.73800007e-01,   1.67720005e-01,   8.39839995e-01,
       5.51070012e-02,   1.05470002e-01,   3.78719985e-01,
       2.42750004e-01,   1.47449998e-02,   5.59509993e-01,
       1.25210002e-01,  -6.75960004e-01,   3.58420014e-01,
       # ... and so on ...
       3.66849989e-01,   2.52470002e-03,  -6.40089989e-01,
      -2.97650009e-01,   7.89430022e-01,   3.31680000e-01,
      -1.19659996e+00,  -4.71559986e-02,   5.31750023e-01], dtype=float32)
```

Word Similarity

Hochschule Offenburg
offenburg.university

- Doc, Span and Token come with a method .similarity()

```
import spacy

nlp = spacy.load("en_core_web_md")  # make sure to use larger model!
tokens = nlp("dog cat banana")

for token1 in tokens:
    for token2 in tokens:
        print(token1.text, token2.text, token1.similarity(token2))
```

```
dog dog 1.0
dog cat 0.80168545
dog banana 0.24327643
cat dog 0.80168545
cat cat 1.0
cat banana 0.28154364
banana dog 0.24327643
banana cat 0.28154364
banana banana 1.0
```

Credits & Recommended reading

Hochschule Offenburg
offenburg.university

- Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Daniel Jurafsky, James H. Martin, 2019. https://web.stanford.edu/~jurafsky/slp3/edbook_oct162019.pdf
- Introduction to Information Retrieval. Manning, Raghavan, Schütze. Cambridge University Press, 2008. https://nlp.stanford.edu/IR-book/information-retrieval-book.html
- spaCy, http://spacy.io
- Some slides are (slightly adapted) taken from Dan Jurafsky, Stanford University in addition to screenshots and content from his book