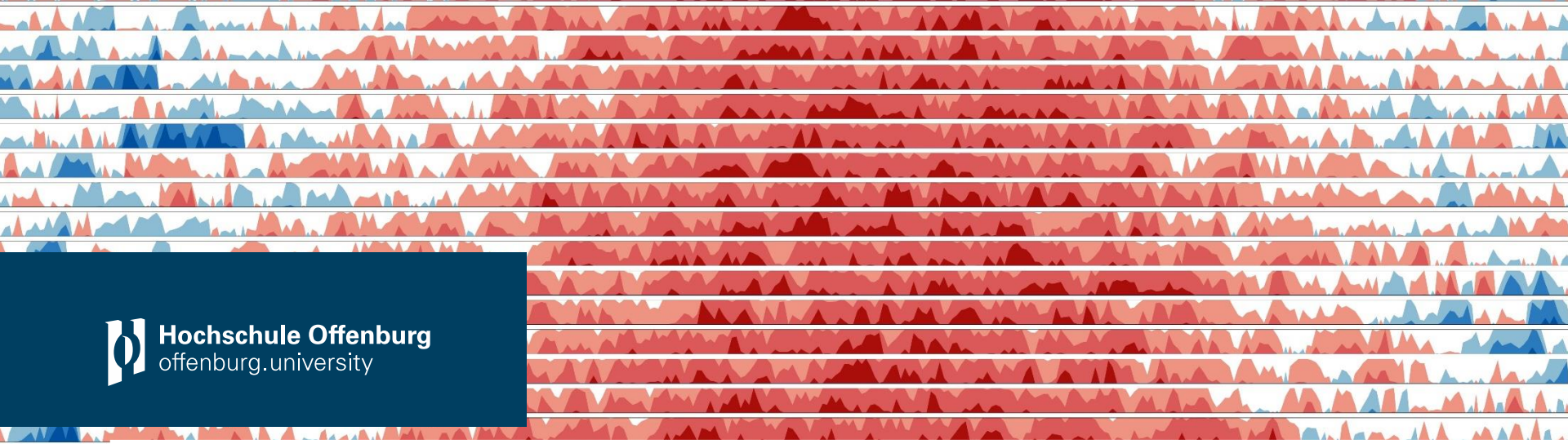


Hochschule Offenburg
offenburg.university

Natural Language Processing & Visual Analytics

Prof. Dr. Daniela Oelke

EMI Elektrotechnik, Medizintechnik und Informatik

Outline

Hochschule Offenburg
offenburg.university

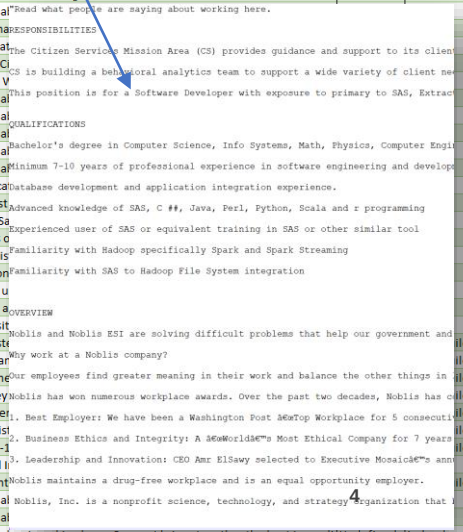
- Text Classification
- Normalization of Natural Language Texts
- Feature Extraction from Natural Language Texts
- Information Extraction
- NLP with spaCy
- Visual Analytics

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

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 (Midtc	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 (Midtc	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 (Midtc	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 Ci		
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 Sei		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



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 (Midtc	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 (Midtc	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 (Midtc	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 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 Ci		
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										
06.02.2019 08:06	Data Scientist / Data Engineer										
06.02.2019 08:03	Data Scientist with Exp in Data Modelling										
06.02.2019 08:06	Data Scientist										
06.02.2019 08:04	Data Scientist needed with solid R and Python skill										
06.02.2019 08:08	Level I Solutions Engineer - Data Scientist										
06.02.2019 08:40	Data Scientist	military									
06.02.2019 08:30	Data Scientist	business and financial operation									
06.02.2019 08:30	Senior Data Scientist - Tallahassee, FL - $150k-$170	business and financial operation									
06.02.2019 08:28	Data Scientist	business and financial operation									
06.02.2019 08:27	Data Scientist	business and financial operation									
06.02.2019 08:27	Data Scientist	business and financial operation									
06.02.2019 08:28	Data Scientist	business and financial operation									
06.02.2019 08:31	Data Scientist Tampa, FL $110-130K	business and financial operation									
06.02.2019 07:48	Data Scientist										
06.02.2019 08:29	Clinical Data Scientist	business and financial operation									
06.02.2019 09:37	Sr. Mgr, Data Scientist	Engineering/Architecture									
06.02.2019 09:37	Data Scientist Intern, Engineering - Software Devel	Engineering/Architecture									
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

SAS or equivalent training in SAS or other similar tool

adoop specifically Spark and Spark Streaming

AS to Hadoop File System integration

ESI are solving difficult problems that help our government and

is company?

greater meaning in their work and balance the other things in

arous workplace awards. Over the past two decades, Noblis has c

We have been a Washington Post â€œTop Workplace for 5 consecuti

and Integrity: A â€œWorldâ€™s Most Ethical Company for 7 years

Innovation: CEO Amr ElSawy selected to Executive Mosaicâ€™s ann

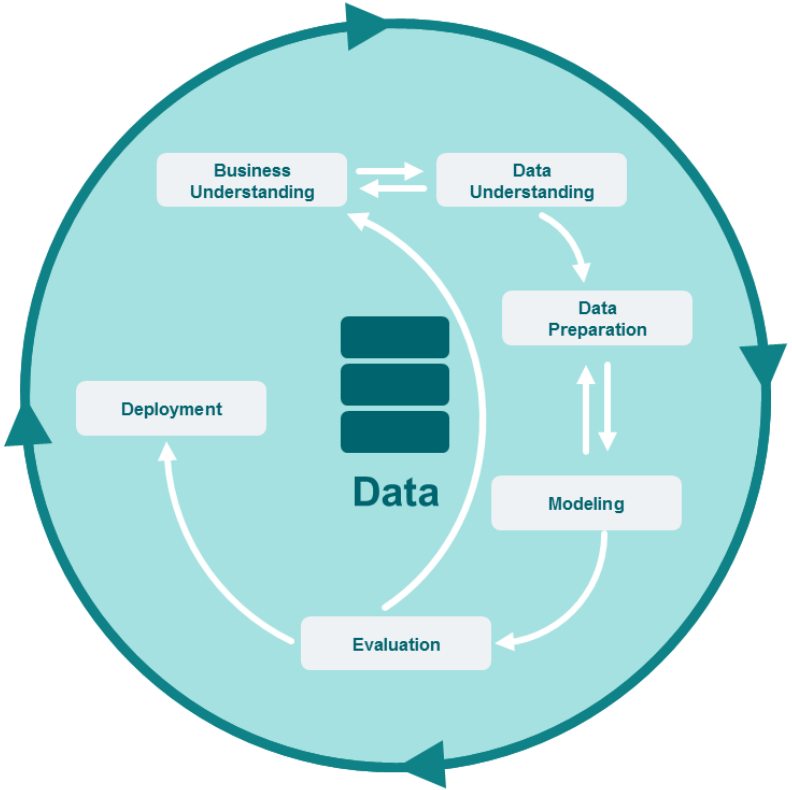
drug-free workplace and is an equal opportunity employer.

nonprofit science, technology, and strategy organization that

Using this dataset:
Which analysis goals could a data science student have?
Which analysis goals could the HR department of a company have?

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 (Midtc	Usa	New york	06.02.2019	Read what people are saying about working here. About BlackRockBlackRock helps investor	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 (Midtc	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 (Midtc	Usa	New york	06.02.2019	Read what people are saying about working here. About BlackRockBlackRock helps investor	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 about working here. Job DetailsDescriptionSchedule:Monday -	Undefined	indeed
06.02.2019 07:33	Data Scientist		Perspecta	Washington	DC	Usa	Washington	05.02.2019	E	Full Time	dice
			Pyramid Consulting, Inc.	Mclean	VA	Usa	Mclean	05.02.2019	Ir	Mont	Undefined	dice
			Orpine.com	Boston	MA	Usa	Boston	05.02.2019	W		Contract	dice
					VA	Usa	Mclean	05.02.2019	H	of ou	Contract	dice
						WA 98104 (First	Usa	Seattle	06.02.2019	R	ata Sc	Undefined	indeed
					DE	Usa	Newark	06.02.2019	R		Undefined	indeed

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



Text Classification

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 OF		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 o	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

Step 1: Feature Extraction

job_description
Read what people are saying about working here. We are Farmers!Join a team of diverse p
We have an immediate opening for a Sharp Data Scientist with a strong Mathematical/Stat
Candidates should have the following background, skills and characteristics: Â Experience
Read what people are saying about working here. About BlackRockBlackRock helps investc
We are seeking an extraordinary Data Scientist in Charlotte to join our fast growing healthc
Read what people are saying about working here. OpportunityThe opportunity is to join ou
Read what people are saying about working here. At Spotify our mission is to provide the v
Job Title: - Sr. Data Science Consultant Duration: 1+ Yrs (will get extend) Mode of interview
Read what people are saying about working here. About BlackRockBlackRock helps investc
â€¢ 3+ years related a professional experienceÂ â€¢ Proven achievements resulting from
Read what people are saying about working here. At Northrop Grumman, innovation isn't j
Data Scientist EnvoyIT is looking for a Data Scientist for a Full-time position with one of ou
Read what people are saying about working here. RESPONSIBILITIESThe Citizen Services Mi
Read what people are saying about working here. Job DetailsDescriptionSchedule:Monday
Every day at Perspecta, we enable hundreds of thousands of people to take on our nation
Immediate need for Senior Data Scientist with experience in the IT Industry. This is 12 Mor
Work Authorization: Only US Citizen's & Green CardÂ Â Job description:Â Â â€¢ Business
Hi, Hope you are doing great. We are having requirement forÂ Data Scientist with one of o
Read what people are saying about working here. Job Description:Position SummaryData S
Read what people are saying about working here. At Operations strategy and Analytics, we
Read what people are saying about working here. Read something interesting online today
Read what people are saying about working here. About UsVideoAmpâ€™s mission is to br
Read what people are saying about working here. SummaryPosted: Feb 5, 2019Weekly Hou
Data Scientist 2 years Qualifications and Skills: Â·Â Â Â Â Â Â Â 3+ years work experience in a

How do we get from plain text to a feature vector?

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

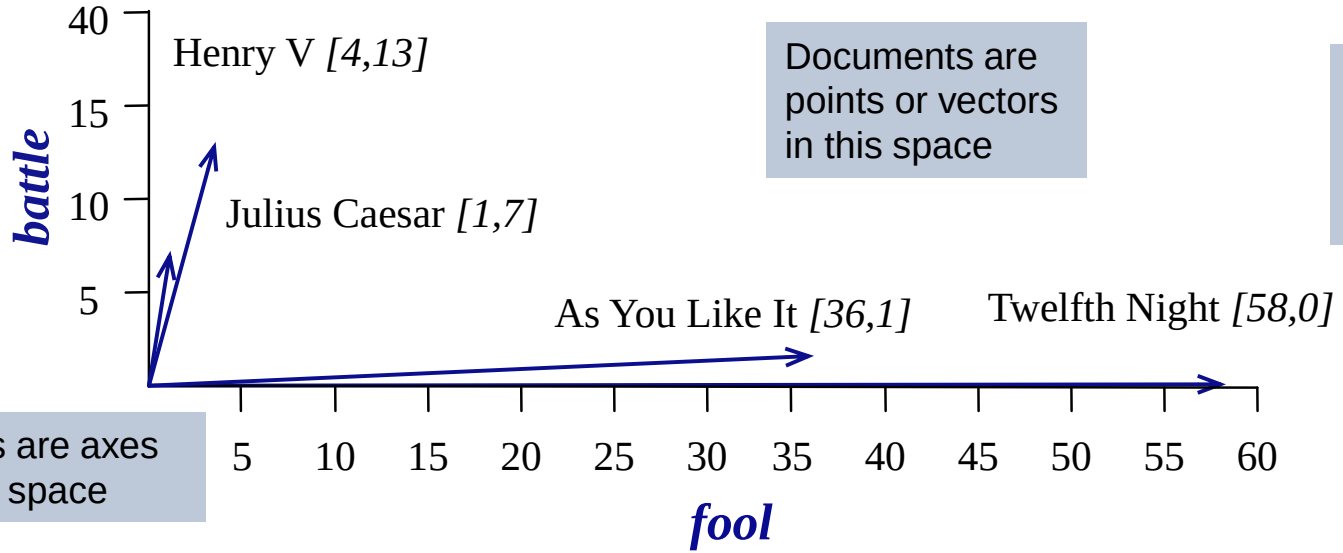
Bag of words model

- 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

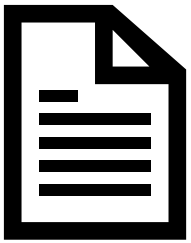
High-dimensional space! (here: only 2 dimensions shown)

Terms are axes of the space

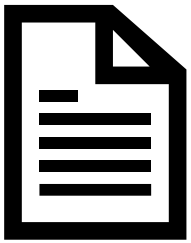
Term frequency tf

Hochschule Offenburg
offenburg.university

- 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



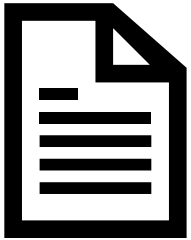
search term
count = 1



search term
count = 10

More relevant 10x more relevant?

Rare terms are more informative



Term frequency (arachnophobia) = 1

Term frequency (actor) = 1

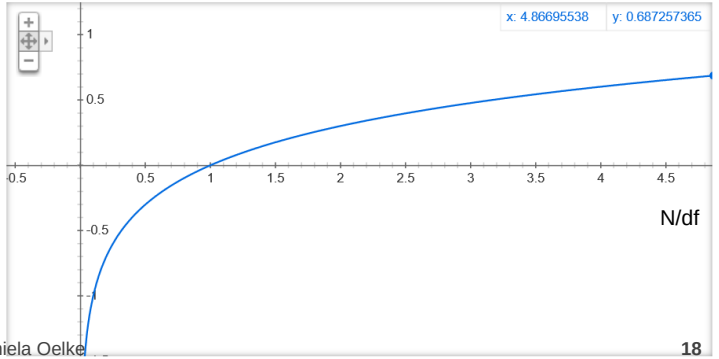
- Search term: *arachnophobia* → document most likely relevant
- Search term: *actor* → document not necessarily relevant

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.



idf example, suppose N = 1 million

Hochschule Offenburg
offenburg.university

term	df_t	idf_t
calpurnia	1	6
animal	100	4
sunday	1,000	3
fly	10,000	2
under	100,000	1
the	1,000,000	0

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

There is one idf value for each term t in a collection.

TF-IDF weighting

Hochschule Offenburg
offenburg.university

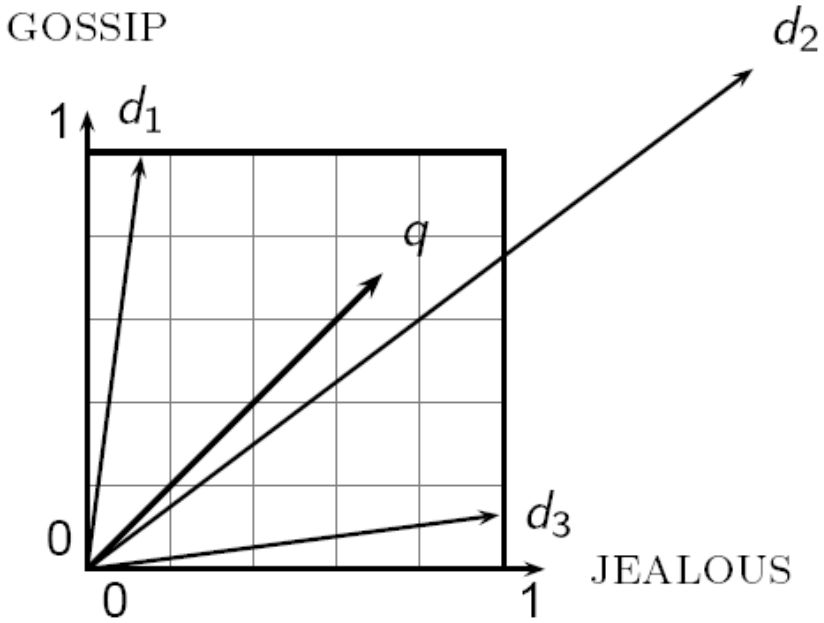
$$w_{t,d} = tf_{t,d}\ x \log_{10}(\frac{N}{df_t})$$

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

Calculating the similarity between two document vectors



How would you calculate the similarity between these document vectors?

Cosine similarity

Hochschule Offenburg
offenburg.university

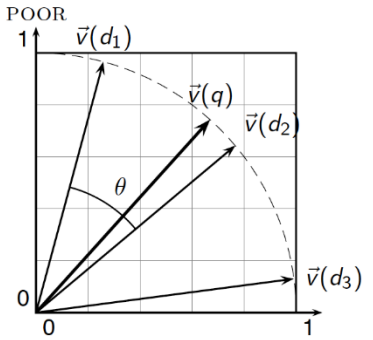
$$\text{cosine}(\vec{v},\vec{w}) = \frac{\vec{v}\cdot\vec{w}}{|\vec{v}||\vec{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

If the vectors are length-normalized:

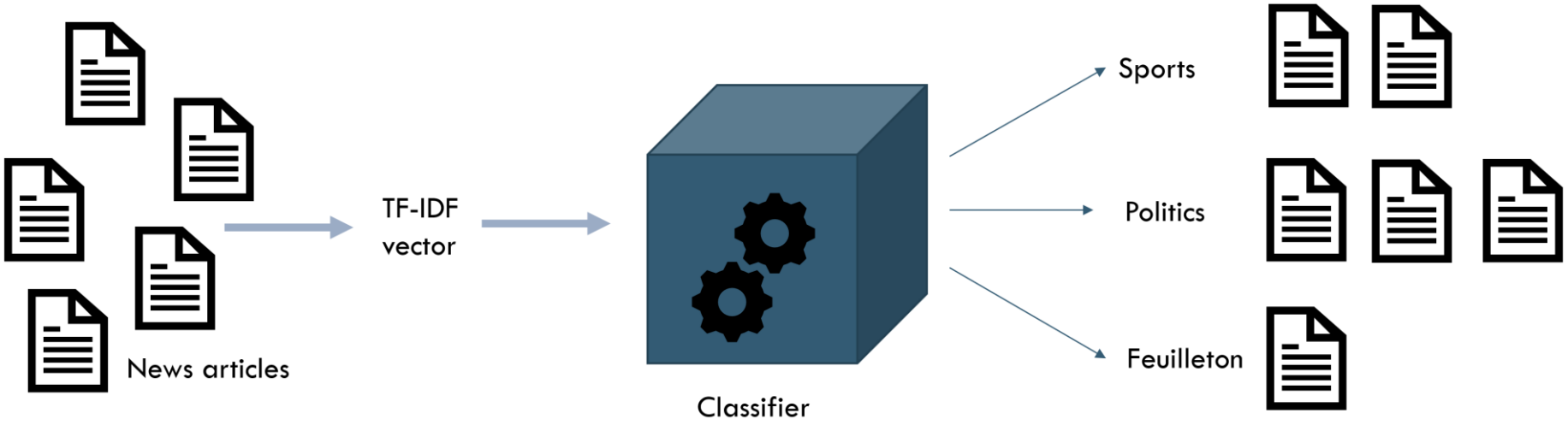
$$\text{cosine}(\vec{v},\vec{w}) = \sum_{i=1}^{N} v_i w_i$$

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*



Application: TF-IDF for classification

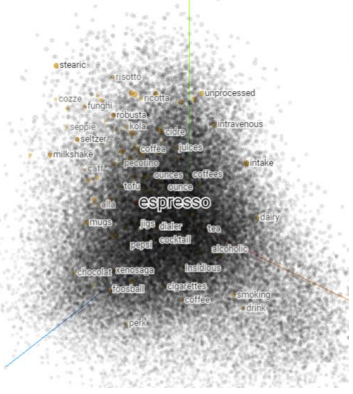


Exercise 1: Text Classification

What‘s next?
Word Embeddings & Co.

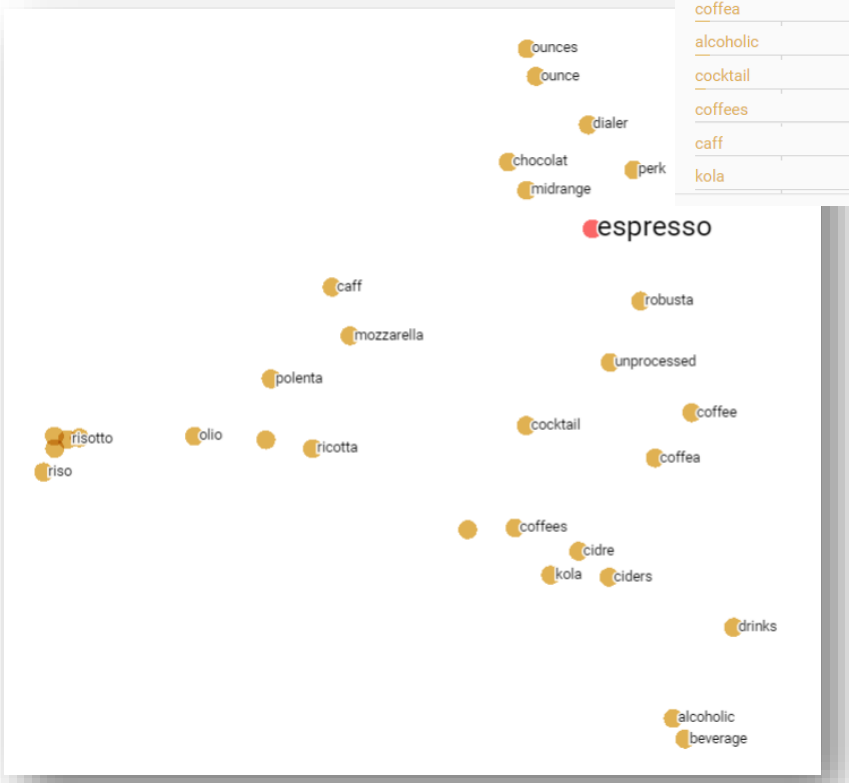
Word2Vec

- 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

https://projector.tensorflow.org/

espresso

$$\begin{bmatrix} 0.68 \\ \dots \\ 0.94 \\ \dots \\ 0.12 \\ 0.04 \\ -0.13 \\ -0.25 \\ 0.71 \\ 0.21 \end{bmatrix}$$

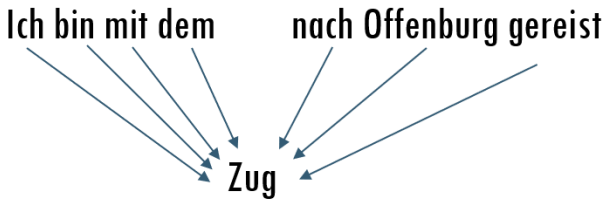
How do we get the vectors?

Hochschule Offenburg
offenburg.university

„You shall know a word by the company it keeps“

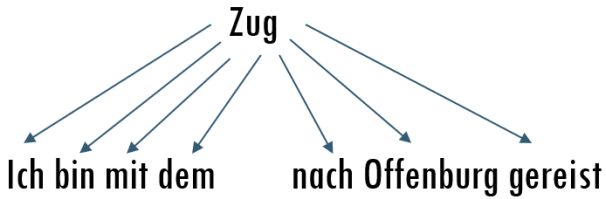
Method 1:
Continuous Bag of Words (CBOW)

Task: Using the context, predict the missing word

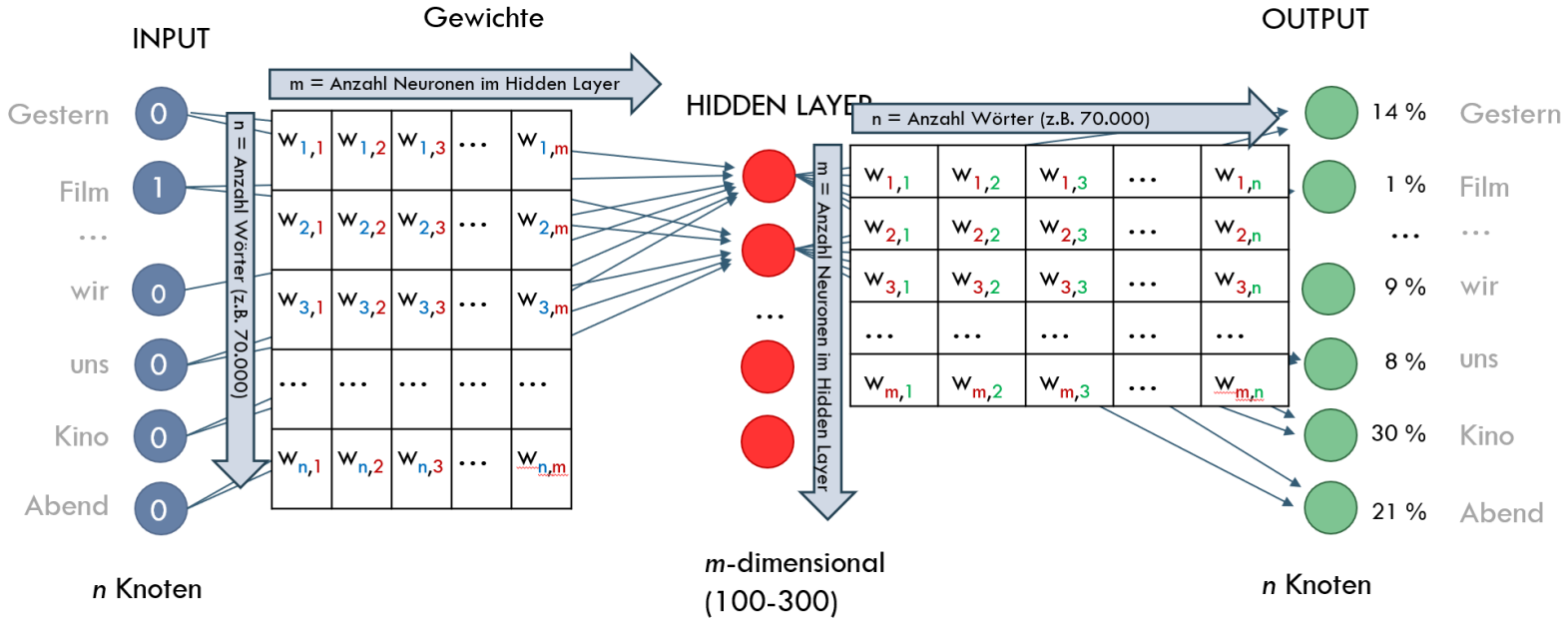


Method 2:
Skip-gram

Task: Given a word, predicts its context words



Learning the word vectors



Disadvantages of Word2Vec

Hochschule Offenburg
offenburg.university

- Using Word2Vec, any given word in a vocabulary (e.g., "go") has its own word vector, and those vectors are effectively stored in a lookup table or dictionary
- Word2Vec **does not address polysemy**, or the co-existence of many possible meanings for a given word or phrase
  - *"go"* is a verb and it is also a board game
- The **meaning** of a given word type **varies according to its context**

- In recent years, several models have been presented that learn **context-dependent** word vectors (e.g., ELMo, BERT)

Context Dependency

Hochschule Offenburg
offenburg.university

- **Context independent models** (e.g., Word2vec, Glove)
  - these models output just one vector (embedding) for each word regardless of where the words occur in a sentence and regardless of the different meanings they may have

- **Context dependent models** (e.g., ELMo, BERT)
  - these models can generate different word embeddings for a word that **captures the context** of a word
  - they take the entire input sentence into equation for calculating the word embeddings

Sentence Embeddings

Hochschule Offenburg
offenburg.university

- Simple baseline: Averaging Word Embeddings
  - The simplest way to calculate sentence embeddings is to calculate the average word embeddings among all words in a sentence
  - Use a *simple* model (e.g., Word2Vec, Glove) to get word embeddings
  - Even though the approach is simple, the results are satisfying
- Deep Learning Approaches

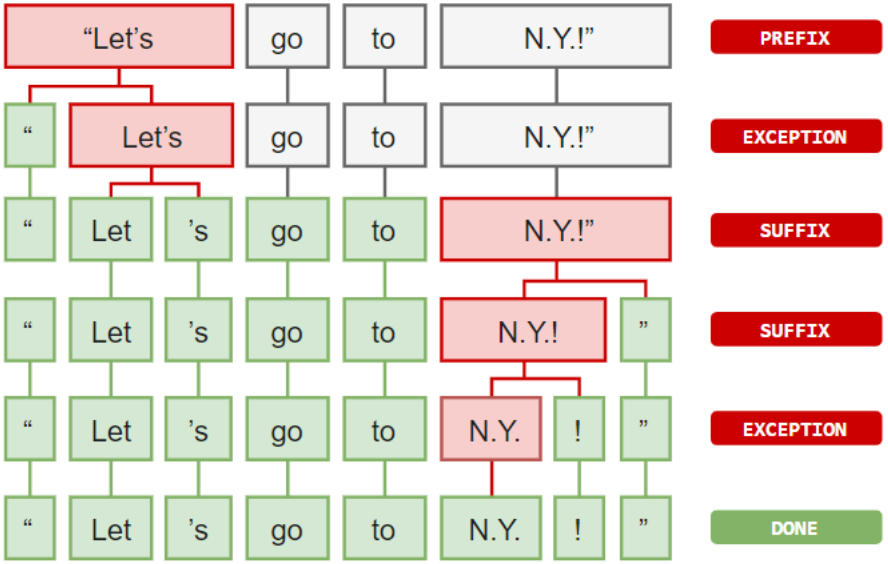
Normalization of Natural Language Texts

Tokenization

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

  I'll see you in New York.

  The aluminum-export ban.
- Approaches
  - separate words using whitespace
  - use regular expressions to specify which substrings are valid words



Source: Spacy.io

Stopword Removal

- Stopwords: very common words, „without meaning“ (in the specific context) or of little value for the specific task
- Benefit of removing:
  - dataset size decreases
  - can increase accuracy, because fewer and only meaningful tokens are left
  - can increase speed
- Avoid stopword removal for
  - machine translation
  - language modeling
  - text summarization
  - question-answering problems

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

Case folding

- Mapping everything to lower case
- Helpful in tasks like information retrieval or speech recognition
- Not recommended for tasks like text classification, information extraction or machine translation (case = important information!)

Stemming

- Naive version of morphological analysis
- Chops off word-final affixes
- Example:

```
Pierre Vinken , 61 years old , will join the board as a nonexecutive
director Nov. 29 . Mr. Vinken is chairman of Elsevier N.V. , the Dutch
publishing group . Rudolph Agnew , 55 years old and former chairman of
Consolidated Gold Fields PLC , was named a nonexecutive director of
this British industrial conglomerate . A form of asbestos once used to
make Kent cigarette filters has caused a high percentage of cancer
deaths among a group of workers exposed to it more than 30 years ago ,
researchers reported .
```

is turned into:

```
Pierr Vinken , 61 year old , will join the board as a nonexecut
director Nov. 29 . Mr. Vinken is chairman of Elsevi N.V. , the Dutch
publish group . Rudolph Agnew , 55 year old and former chairman of
Consolid Gold Field PLC , wa name a nonexecut director of thi British
industri conglomer . A form of asbesto onc use to make Kent cigarett
filter ha caus a high percentag of cancer death among a group of
worker expos to it more than 30 year ago , research report .
```

Example rules of Porter Stemmer:

```
sses → ss
ss → ss
s → _
ies → i
y → i
```

Lemmatization

- Determine if two words have the same root
- Examples:
  - am, are, is → be
  - dinner, dinners → dinner
  - He is reading detective stories → He be read detective story
- Lemmatization is done by morphological parsing
- Lemmatization algorithms can be complex!

Example

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.

Abbreviation

Sentence boundary marker & Abbreviation

Sentence boundary marker

- How solved?
  - Use rules or ML to decide whether a period is part of a word or a sentence-boundary marker.
  - Use abbreviation dictionaries

spaCy uses the dependency parser for sentence segmentation

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

What does Tagging do?

1. **Collapses Distinctions**
   - Lexical identity may be discarded
   - e.g., all personal pronouns tagged with PRP

2. **Introduces Distinctions**
   - Ambiguities may be resolved
   - e.g. *deal* tagged with NN or VB

3. **Helps in classification and prediction**

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

The Tagging Process

Hochschule Offenburg
offenburg.university

Training data
(Annotated text)

This	*sentence*	*serves*	*as*	*an*	*example*	*of*	*annotated*	*text…*
Det	N	V1	P	Det	N	P	V2	N

"This is a new sentence." → POS Tagger → *This* Det *is* Aux *a* Det *new* Adj *sentence.* N

Pick the most likely tag sequence.

$$p(w_1,...,w_k,t_1,...,t_k) = \begin{cases} p(t_1 \mid w_1)...p(t_k \mid w_k)\,p(w_1)...p(w_k) \\ \prod_{i=1}^{k} p(w_i \mid t_i)\,p(t_i \mid t_{i-1}) \end{cases}$$

Independent assignment
Most common tag

Partial dependency
(HMM)

(Adapted from ChengXiang Zhai, CS 397cxz – Fall 2003)

Training and Testing of Learning Algorithms

- Algorithms that “learn” from data see a set of examples and try to generalize from them.
- Training set:
  - Examples trained on
- Test set:
  - Also called held-out data and unseen data
  - Use this for evaluating your algorithm
  - Must be separate from the training set
    - Otherwise, you cheated!
- “Gold” standard
  - A test set that a community has agreed on and uses as a common benchmark.

Three main corpora tagged with POS for English: Brown (different genres), WSJ (Wall Street Journal), Switchboard (phone conversations)

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].

Constituency

Hochschule Offenburg
offenburg.university

- Constituents: groups of words behaving as single units
- Example:
  - *On September seventeenth*, I'd like to fly from Atlanta to Denver
  - I'd like to fly on *September seventeenth* from Atlanta to Denver
  - I'd like to fly from Atlanta to Denver *on September seventeenth*
- What about
  - On September, I'd like to fly seventeenth from Atlanta to Denver
  - On I'd like to fly September seventeenth from Atlanta to Denver
  - I'd like to fly on September from Altanta to Denver seventeenth

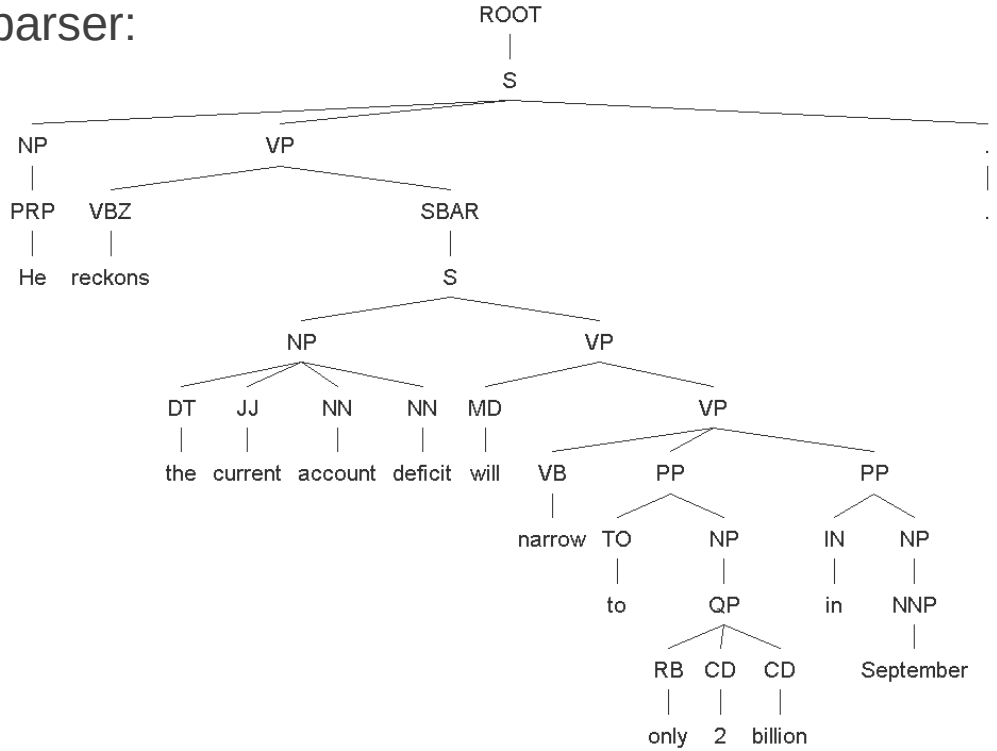
What about German?
Determine the constituents *in the following sentence:*

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

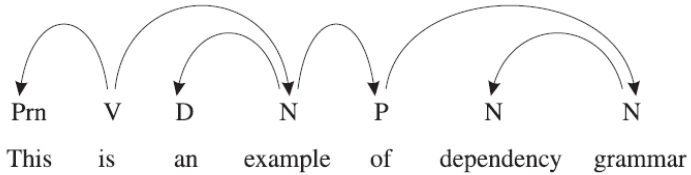
Comparison of POS tagging, chunking and full parsing

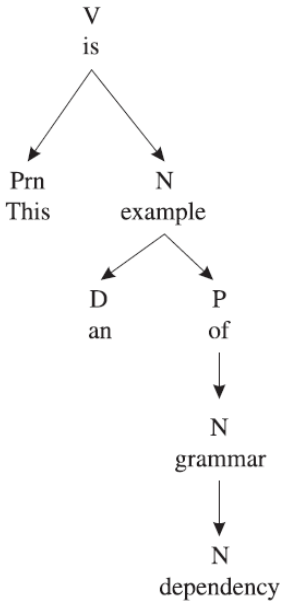
Hochschule Offenburg
offenburg.university

Output of full parser:

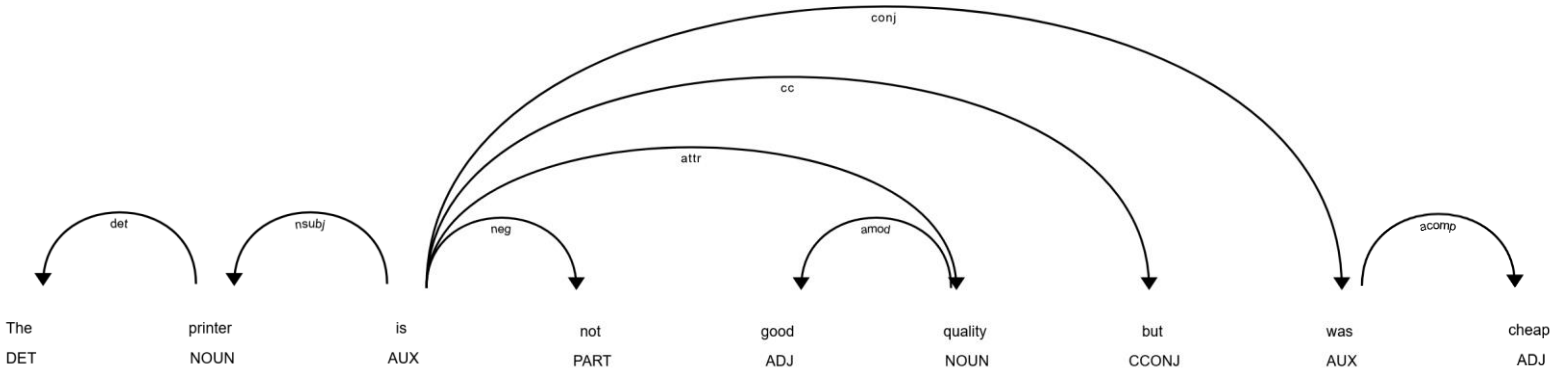


Dependency Parsing





Example with negation



Application 1: Grammar checking

- Check if sentence can be parsed (full parse tree)
- Yes: likely grammatical
- No: likely ungrammatical or at least difficult to read

Application 2: Question answering

- Example question:
  „What books were written by British women authors before 1800?“
- Parsing helps to determine if the user wants a list of books or authors

Application 3:
Information extraction systems

Hochschule Offenburg
offenburg.university

- With chunking, only segments in a text that contain valuable information are identified and classified.

Application 4: Information Retrieval

Hochschule Offenburg
offenburg.university

- Using chunking, Information Retrieval systems may index texts according to a subset of the constituents found in them

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 Entity Recognition

Relation extraction

Event extraction

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**.

Families of algorithms for NER tagging

- Feature based ML systems
- Neural (bi-LSTM)
- Rule-based

Input for feature based NER systems

Typical features for a feature-based NER system

identity of w_i, identity of neighboring words
embeddings for w_i, embeddings for neighboring words
part of speech of w_i, part of speech of neighboring words
base-phrase syntactic chunk label of w_i and neighboring words
presence of w_i in a **gazetteer**
w_i contains a particular prefix (from all prefixes of length ≤ 4)
w_i contains a particular suffix (from all suffixes of length ≤ 4)
w_i is all upper case
word shape of w_i, word shape of neighboring words
short word shape of w_i, short word shape of neighboring words
presence of hyphen

Deep dive: Word shape features

- Abstract lettern patterns of a word
- Mapping
  - lower-case letters → x
  - upper-case letters → X
  - numbers → d
  - punctuation → retained
- Example:
  - DC10-30 → XXdd-dd
- Short version: remove consecutive character types
  - DC10-30 → XXdd-dd → Xd-d

Deep-dive: Gazetteer

- List of place names
- Related: name-lists / corporations / commercial products…
- Typically implemented as binary feature (in / not in gazetteer)
- Problems:
  - can be difficult to create and maintain
  - usefulness varies considerably

Which features are best?

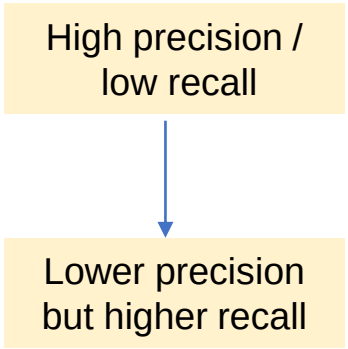
- Which features are best?
  This also depends on the application, genre, media, and language

identity of w_i, identity of neighboring words
embeddings for w_i, embeddings for neighboring words
part of speech of w_i, part of speech of neighboring words
base-phrase syntactic chunk label of w_i and neighboring words
presence of w_i in a **gazetteer**
w_i contains a particular prefix (from all prefixes of length ≤ 4)
w_i contains a particular suffix (from all suffixes of length ≤ 4)
w_i is all upper case
word shape of w_i, word shape of neighboring words
short word shape of w_i, short word shape of neighboring words
presence of hyphen

Which features would you use when applying NER to Twitter posts?

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



Natural Language Processing with spaCy

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

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(„...“)
```

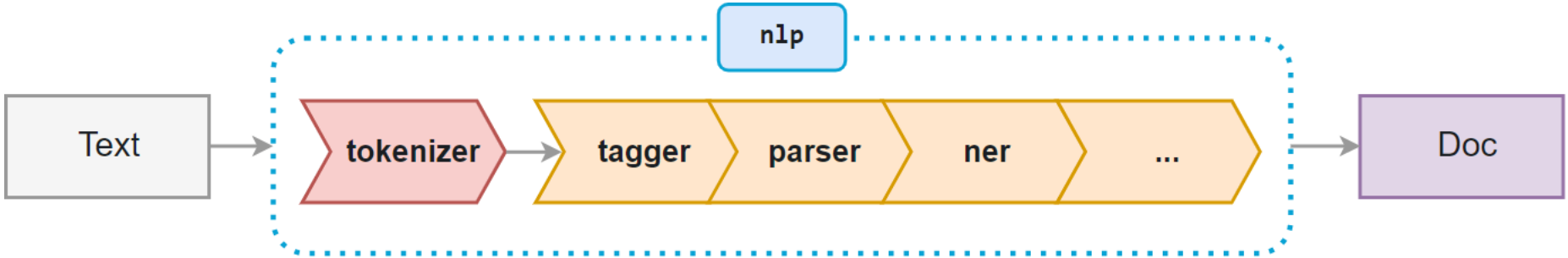
Returns a Language object, usually called „nlp“

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

Returns a Doc object

Can be called with a string.

Processing pipeline



Lemmatization

```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_)
```

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

Hochschule Offenburg
offenburg.university

```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

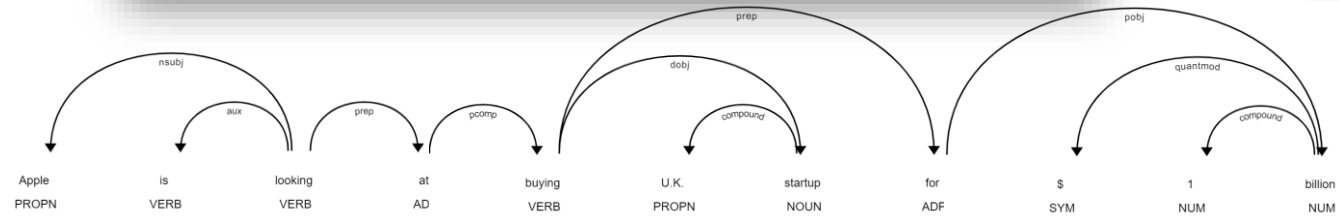
Hochschule Offenburg
offenburg.university

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

- 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
```

Exercise 2: spaCy and NER

Named Entities

The EIB ORG will finance around 40 CARDINAL projects in Barcelona GPE that aim to support climate change mitigation and adaptation in the city. To this end, the EU ORG bank will provide €95 million MONEY to promote urban regeneration, with a focus on the environment but also on social inclusion and job creation to boost the economic recovery in the wake of the COVID-19 crisis.

The Italian NORP healthcare system is also being reinforced to tackle the emergency situation caused by the COVID-19 pandemic. This is being conducted with the backing of the EU ORG bank, the EIB ORG , which is providing the Italian NORP government with a €2 billion MONEY loan covering around two-thirds CARDINAL of the resources needed for the operations contained in the Decree GPE for revival of the healthcare system

The EIB Group ORG and Banco Santander Consumer Portugal ORG (BSCP) are joining forces to support Portuguese NORP small and medium-sized enterprises (SMEs) and mid-caps affected by the COVID-19 crisis. The EU ORG bank and BSCP ORG have signed two CARDINAL agreements to provide EUR GPE 587 million CARDINAL to inject liquidity and finance investments at a critical time.

Amadou Hott PERSON , Senegalese NORP Minister of the Economy, Planning and Cooperation ORG , and Ambroise Fayolle PERSON , EIB ORG Vice-President responsible for Africa LOC , today DATE formally agreed a CFAF 49 billion CARDINAL concessional loan to the Republic of Senegal GPE . It comes on top of a CFAF 200 billion CARDINAL financing mechanism established by Macky Sall PERSON , President of the Republic of Senegal GPE , as part of the country's Economic and Social Resilience Programme.

The health, economic and social impact of COVID-19 across Africa LOC continues to evolve with more than 12,000 CARDINAL deaths, in excess of 540,000 CARDINAL confirmed cases and the livelihoods of millions CARDINAL of Africans NORP damaged.

This has been one of the Italian NORP government's top priorities for at least three years DATE . The government has decidedly sped up the process following the restrictions caused by the COVID-19 pandemic, and this has now become a key pillar to aid the country's recovery. Digitalisation of public administrations (PAs) and their relations with the country's citizens and businesses have become the focus of the digital innovation measures contained in the "Simplifications" Decree Law approved by the Prime Minister's Office ORG on 7 July DATE . It is against this backdrop that the four-year DATE investment plan of publicly-owned company PagoPA S.p.A. ORG is receiving strong backing from the European Union's ORG bank.

NER on the German Press releases

Die EIB-Gruppe ORG und die Banco Santander Consumer Portugal ORG (BSCP ORG) unterstützen gemeinsam von der Coronakrise betroffene kleine und mittlere Unternehmen (KMU) und Midcap-Unternehmen in Portugal LOC .

Die Bank der EU ORG und die BSCP ORG haben zwei Vereinbarungen unterzeichnet, um 587 Millionen Euro als Liquiditätshilfe und zur Finanzierung von Investitionen in einer kritischen Zeit MISC bereitzustellen.

Ganz Afrika LOC leidet zunehmend unter den gesundheitlichen, wirtschaftlichen und sozialen Folgen von Covid-19. ORG

Inzwischen sind auf dem Kontinent über 12 000 Menschen an dem Virus MISC gestorben, mehr als 540 000 haben sich infiziert, und für Millionen sind die Existenzgrundlagen gefährdet.

Seit mindestens drei Jahren steht die Digitalisierung der öffentlichen Verwaltung ORG ganz oben auf der Tagesordnung der italienischen Regierung ORG .

Aufgrund der Covid-19-Beschränkungen MISC wurde sie stark beschleunigt und bildet heute eine der Säuler für den Aufbau des Landes LOC .

Das Gesetzesdekret Nr. 76 MISC vom 16. Juli 2020 über dringende Maßnahmen zur Vereinfachung und zur digitalen Innovation zielt ebenfalls auf die Digitalisierung der öffentlichen Verwaltung und deren Beziehungen zu Bürgern und Unternehmen ab.

In diesem Rahmen ist die Tätigkeit des staatlichen Unternehmens PagoPA S.p LOC . A. PER zu sehen, dessen Vier-Jahres-Investitionsplan die EIB MISC unterstützt.

Die EIB MISC unterstützt in der Autonomen Provinz Trient LOC die Konjunkturerholung nach der Coronapandemie und nachhaltige Projekte des öffentlichen Sektors LOC .

Heute wurde die entsprechende Vereinbarung unterzeichnet – die EIB MISC stellt 300 Millionen Euro für die norditalienische Provinz bereit, wovon eine erste Tranche über 160 Millionen Euro bereits abgeschlossen wurde.

Die EIB MISC hat heute neue Finanzierungen im Umfang von 16,6 Milliarden Euro für Projekte in Europa LOC und anderen Regionen der Welt genehmigt, davon mehr als zehn Milliarden Euro für Investitionen im Zusammenhang mit Covid-19.

NER on job posts

Extract data from a variety of sources (databases, web, text files) and in a variety of formats (structured, unstructured, JSON, Parquet GPE , etc.) and prepare the data for analysis Describe data using qualitative and quantitative aggregation/summarization and build tools (static reports, interactive dashboards, web applications) that allow clients to easily view and understand important patterns and trends in the data Provide quantitative comparisons of data points using advanced statistical analysis including data reduction techniques (PCA ORG , cluste ring), resampling (permutation tests, Monte Carlo PERSON testing) as well as traditional statistical hypothesis testing Construct ORG and validate predictive models using a variety of techniques

Physical Actions ORG Physical Environment

Education Requirements

Master's degree in scientific or quantitative field and documented academic excellence. PhD in scientific field preferred

Experience Requirements

At least year DATE in a data analytics role (business intelligence analyst, data analyst, etc.)

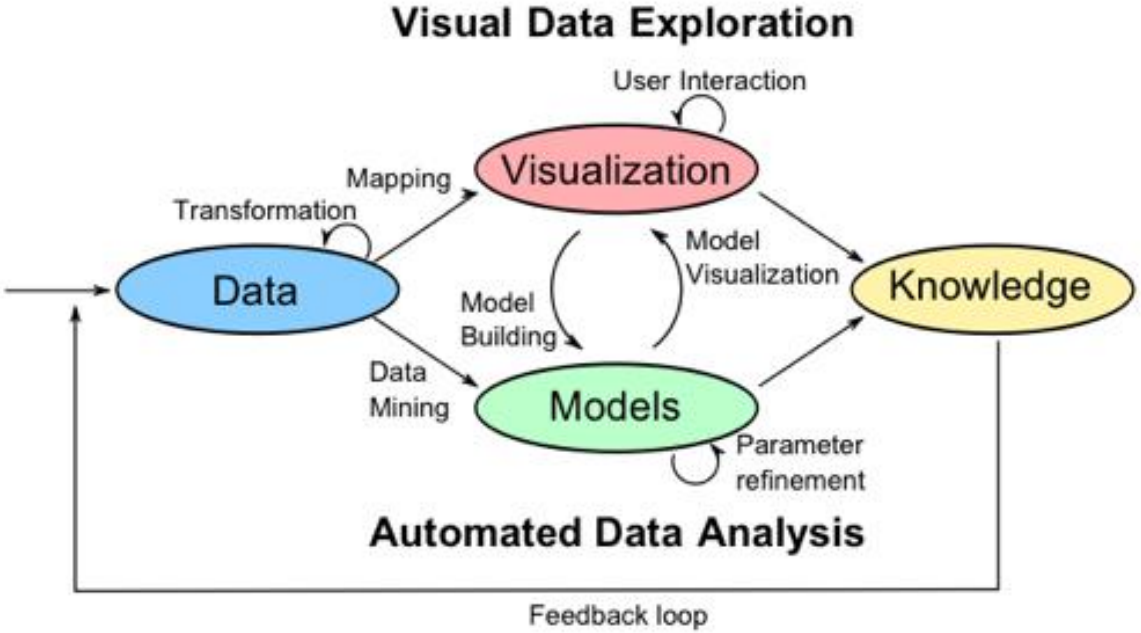
Special Skill Requirement

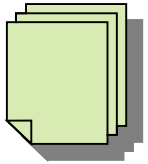
Proficiency in at least one CARDINAL of the following is absolutely required: R, Python Basic ORG proficiency in at least one CARDINAL of the following: PowerBI, Tableau ORG , QlikView ORG , D3.js Basic proficiency in SQL Able ORG to implement and utilize advanced statistical and machine learning techniques (GLMs, PCA ORG , cluster analyses, ARIMA ORG , ETS ORG , decision trees, SVM ORG , neural networks, ensemble methods, etc.) Basic familiarity with project management tools (Asana PERSON , Basecamp PERSON , LiquidPlanner ORG)

Farmers ORG is an equal opportunity employer, committed to the strength of a diverse workforce.

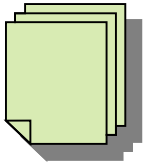
Visual Analytics

Visual Analytics





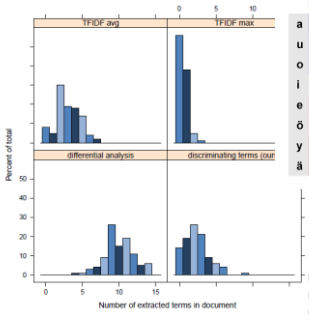
Documents

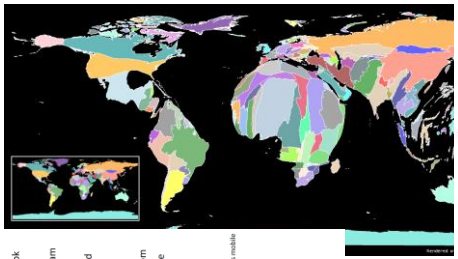


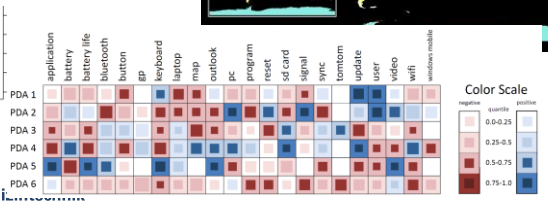
automatic
(pre)processing

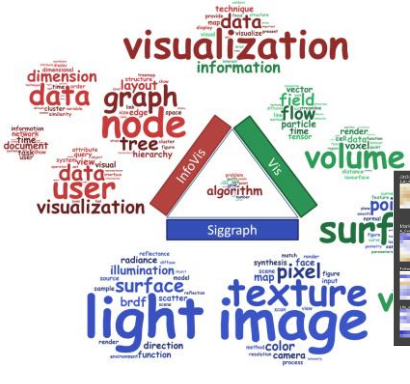
completely structured

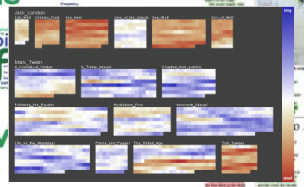
text properties are important











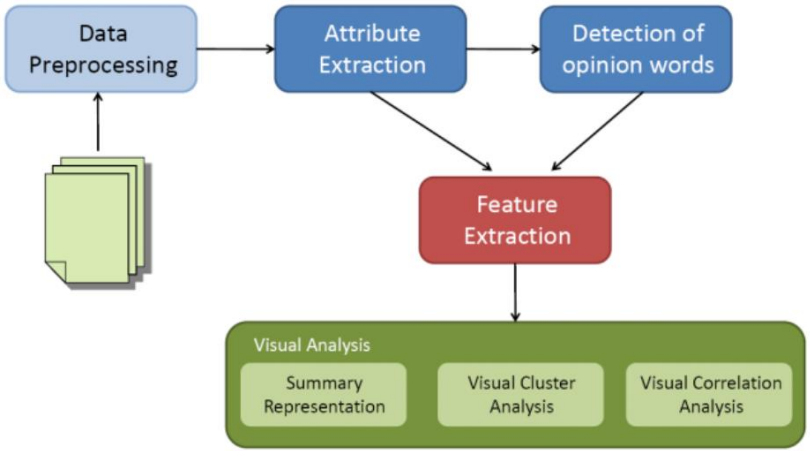
Analysis of customer reviews

Hochschule Offenburg
offenburg.university

Example review:

„I feel obligated to counter the bad reviews. This printer is just fine. I don‘t know what people are complaining about regarding the software, but it installed seamlessly and is intuitive in its operation. Even though the paper tray jams sometimes altogether I am happy that I bought this wonderful printer.“

Preprocessing of reviews



Step 1: Identification of attributes and sentiment

I feel obligated to counter the bad reviews. This **printer** is just fine. I don't know what people are complaining about regarding the **software**, but it installed seamlessly and is intuitive in its operation. Even though the **paper tray** jams sometimes altogether I am happy that I bought this wonderful **printer**.

Step 2: Mapping between attributes and sentiment

[…] Even though the **paper tray** jams sometimes altogether

I am happy that I bought this wonderful **printer**.

Step 3: Determining the overall sentiment of an attribute

[…] This **printer** is just fine. […] Even though the paper tray jams sometimes altogether I am happy that I bought this wonderful **printer**.

➔ Overall sentiment for *printer*: positive

Resulting Feature Vector

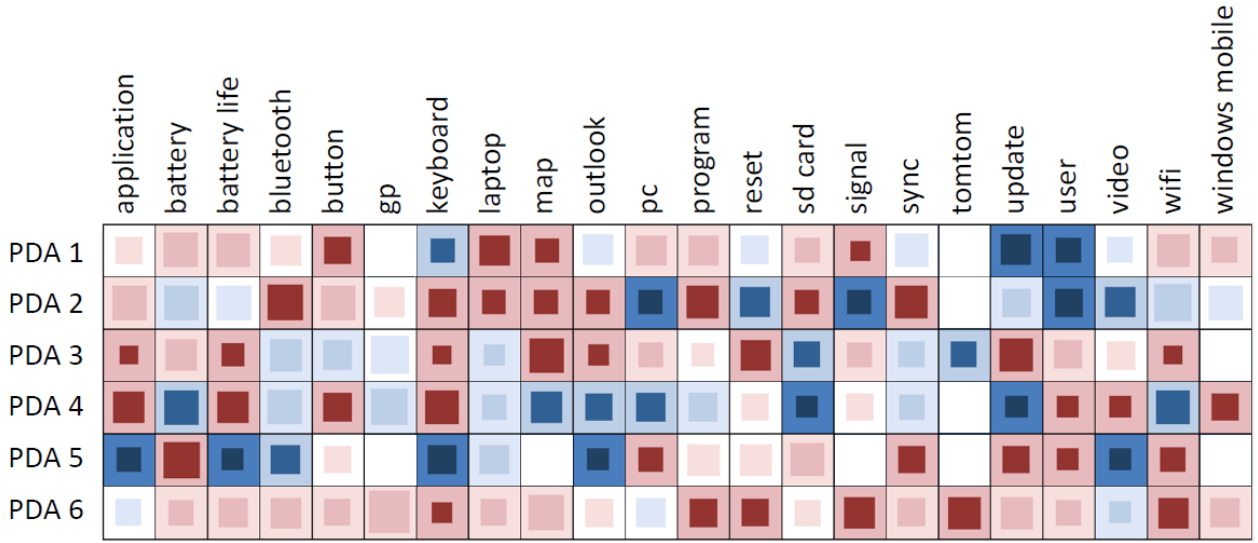
Printer	Ink	Software	Paper tray	Price
1	0	1	-1	0

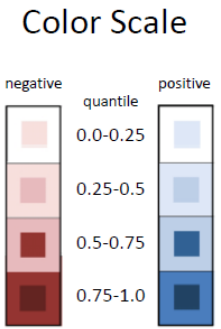
Evaluation of the algorithm

- Overall accuracy: 72 %
  (only subjective sentences, the attributes were given)

Main Error Sources	In % of all Errors
1. Domain-dependency of opinion words	23.5%
2. Fixed expressions and phrases	21.6%
3. Errors in opinion word list	19.6%
4. Opinion word combinations	13.7%
5. But-clauses	7.8%
6. Attribute-dependency of opinion words	5.9%
7. Implicit attributes	2.0%

Review Analysis (example: printer reviews)

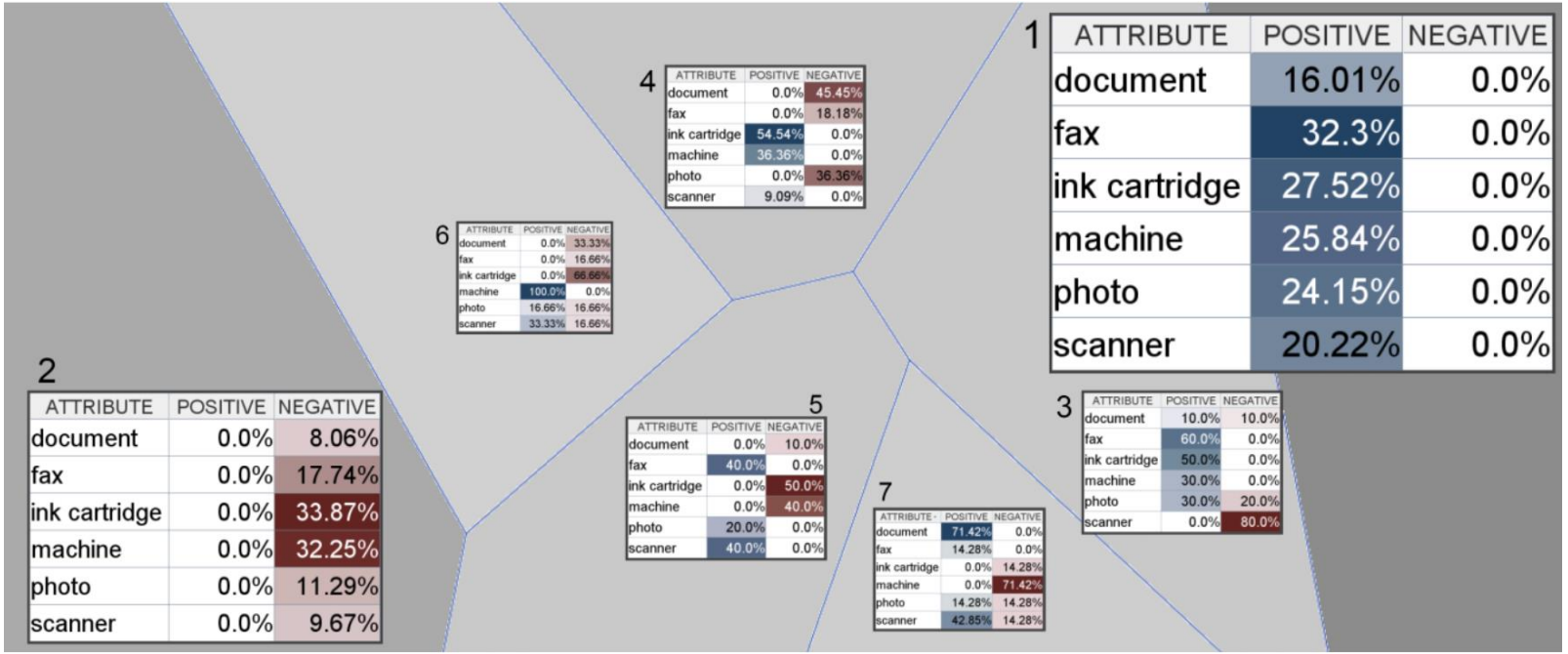




- **size inner rectangle**: amount of customers that commented on the attribute
- **blue color**: positive opinions
- **red color**: negative opinions
- **brightness of color**: degree of positiveness / negativeness of comments

D. Oelke, M. Hao, C. Rohrdantz, U. Dayal, L. Haug, H. Janetzko, "Visual opinion analysis of customer feedback data," *IEEE Symposium on Visual Analytics Science and Technology*, Atlantic City, NJ, 2009, pp. 187-194.

Opinion clusters



1

ATTRIBUTE	POSITIVE	NEGATIVE
document	16.01%	0.0%
fax	32.3%	0.0%
ink cartridge	27.52%	0.0%
machine	25.84%	0.0%
photo	24.15%	0.0%
scanner	20.22%	0.0%

2

ATTRIBUTE	POSITIVE	NEGATIVE
document	0.0%	8.06%
fax	0.0%	17.74%
ink cartridge	0.0%	33.87%
machine	0.0%	32.25%
photo	0.0%	11.29%
scanner	0.0%	9.67%

3

ATTRIBUTE	POSITIVE	NEGATIVE
document	10.0%	10.0%
fax	60.0%	0.0%
ink cartridge	50.0%	0.0%
machine	30.0%	0.0%
photo	30.0%	20.0%
scanner	0.0%	80.0%

4

ATTRIBUTE	POSITIVE	NEGATIVE
document	0.0%	45.45%
fax	0.0%	18.18%
ink cartridge	54.54%	0.0%
machine	36.36%	0.0%
photo	0.0%	36.36%
scanner	9.09%	0.0%

5

ATTRIBUTE	POSITIVE	NEGATIVE
document	0.0%	10.0%
fax	40.0%	0.0%
ink cartridge	0.0%	50.0%
machine	0.0%	40.0%
photo	20.0%	0.0%
scanner	40.0%	0.0%

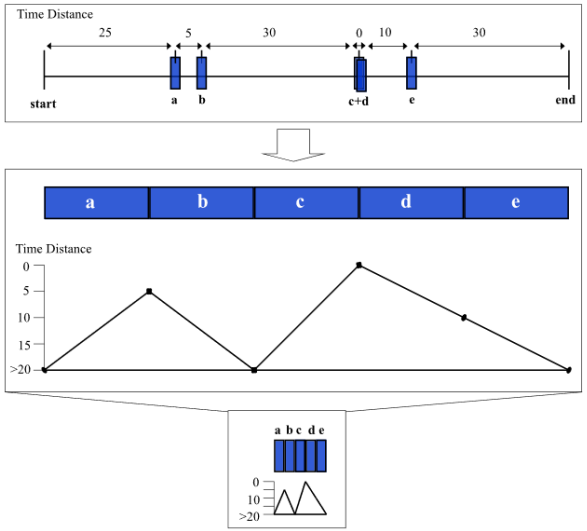
6

ATTRIBUTE	POSITIVE	NEGATIVE
document	0.0%	33.33%
fax	0.0%	16.66%
ink cartridge	0.0%	66.66%
machine	100.0%	0.0%
photo	16.66%	16.66%
scanner	33.33%	16.66%

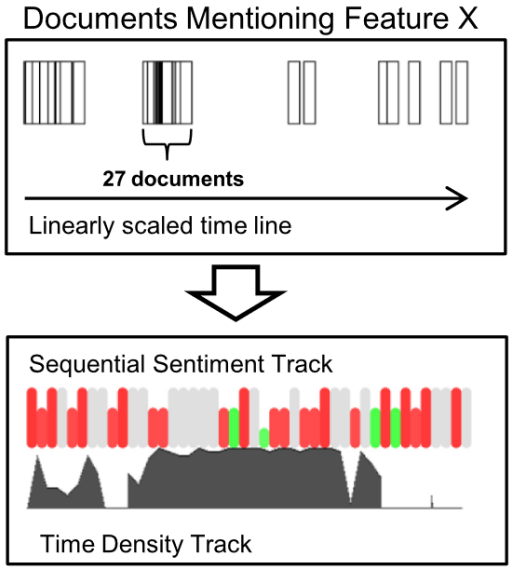
7

ATTRIBUTE	POSITIVE	NEGATIVE
document	71.42%	0.0%
fax	14.28%	0.0%
ink cartridge	0.0%	14.28%
machine	0.0%	71.42%
photo	14.28%	14.28%
scanner	42.85%	14.28%

Review Analysis over time (Time Density Plots)



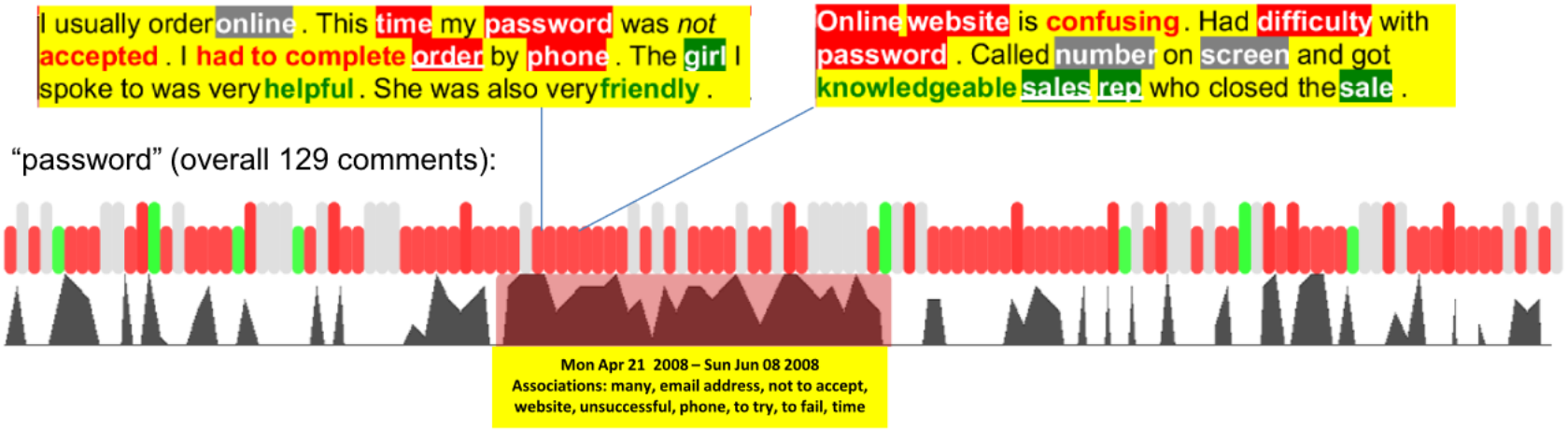
(a) details about the construction of time density curves



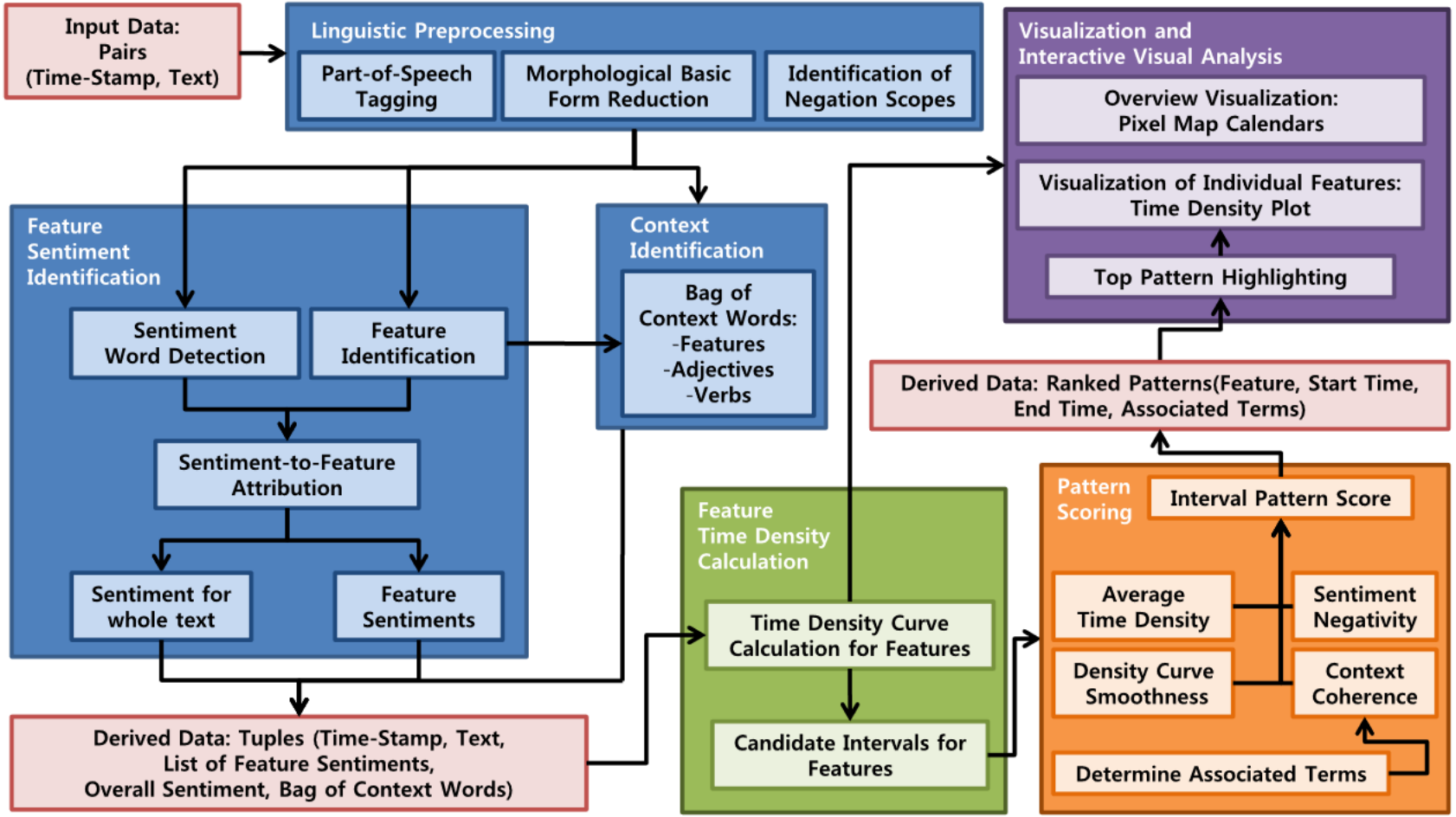
(b) example for a beneficial application of *time density plots*

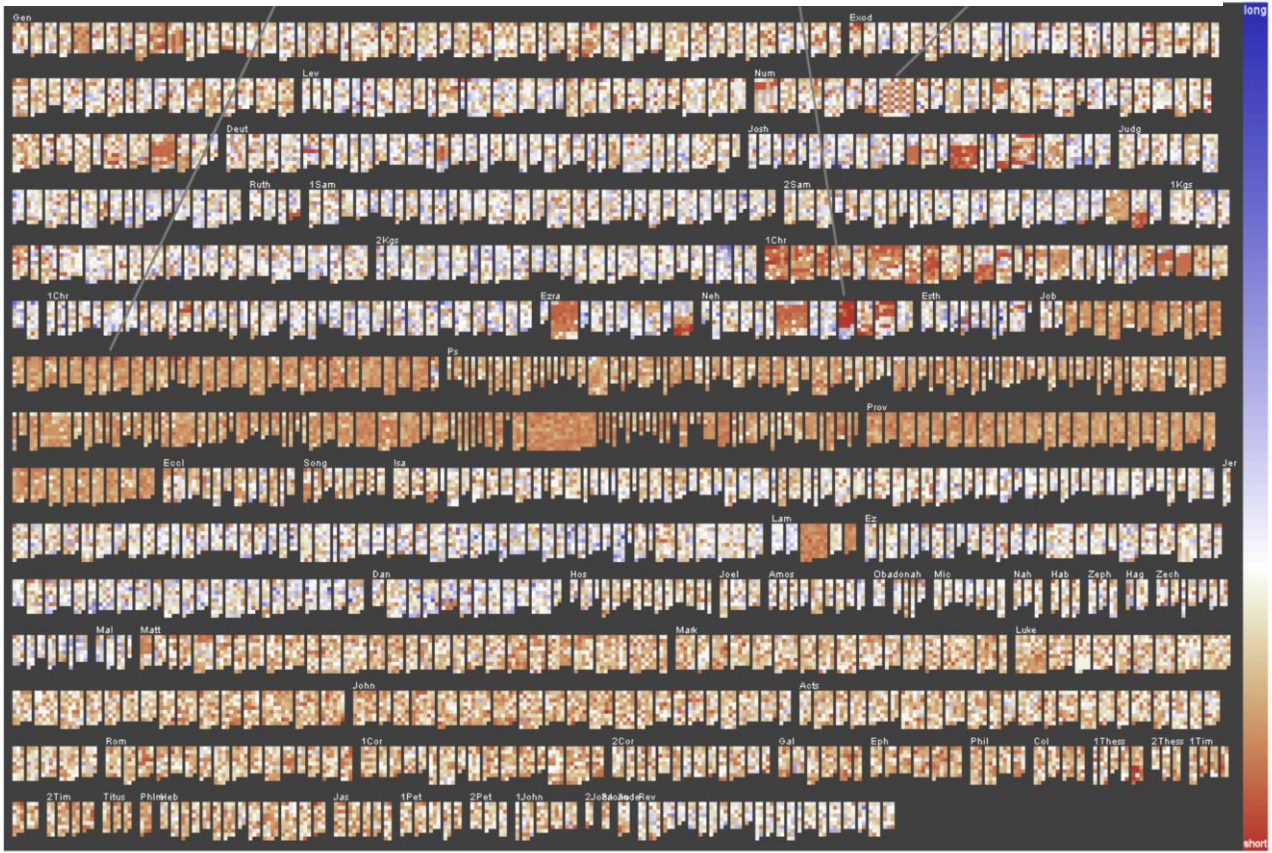
Christian Rohrdantz, Ming C. Hao, Umeshwar Dayal, Lars-Erik Haug, and Daniel A. Keim. *Feature-Based Visual Sentiment Analysis of Text Document Streams*. ACM Transactions on Intelligent Systems and Technology (TIST), vol. 3, no. 2, pp. 26:1-26:25, 2012.

Review Analysis over time (Time Density Plots)



Christian Rohrdantz, Ming C. Hao, Umeshwar Dayal, Lars-Erik Haug, and Daniel A. Keim. *Feature-Based Visual Sentiment Analysis of Text Document Streams*. ACM Transactions on Intelligent Systems and Technology (TIST), vol. 3, no. 2, pp. 26:1-26:25, 2012.


Input Data:
Pairs
(Time-Stamp, Text)
Linguistic Preprocessing
Part-of-Speech Tagging
Morphological Basic Form Reduction
Identification of Negation Scopes
Visualization and Interactive Visual Analysis
Overview Visualization: Pixel Map Calendars
Visualization of Individual Features: Time Density Plot
Top Pattern Highlighting
Feature Sentiment Identification
Sentiment Word Detection
Feature Identification
Sentiment-to-Feature Attribution
Sentiment for whole text
Feature Sentiments
Context Identification
Bag of Context Words:
-Features
-Adjectives
-Verbs
Derived Data: Ranked Patterns(Feature, Start Time, End Time, Associated Terms)
Derived Data: Tuples (Time-Stamp, Text, List of Feature Sentiments, Overall Sentiment, Bag of Context Words)
Feature Time Density Calculation
Time Density Curve Calculation for Features
Candidate Intervals for Features
Pattern Scoring
Interval Pattern Score
Average Time Density
Sentiment Negativity
Density Curve Smoothness
Context Coherence
Determine Associated Terms



D. A. Keim and D. Oelke, "Literature Fingerprinting: A New Method for Visual Literary Analysis," *2007 IEEE Symposium on Visual Analytics Science and Technology*, Sacramento, CA, 2007, pp. 115-122, doi: 10.1109/VAST.2007.4389004.

Visual Analytics of Twitter News for Crisis Management

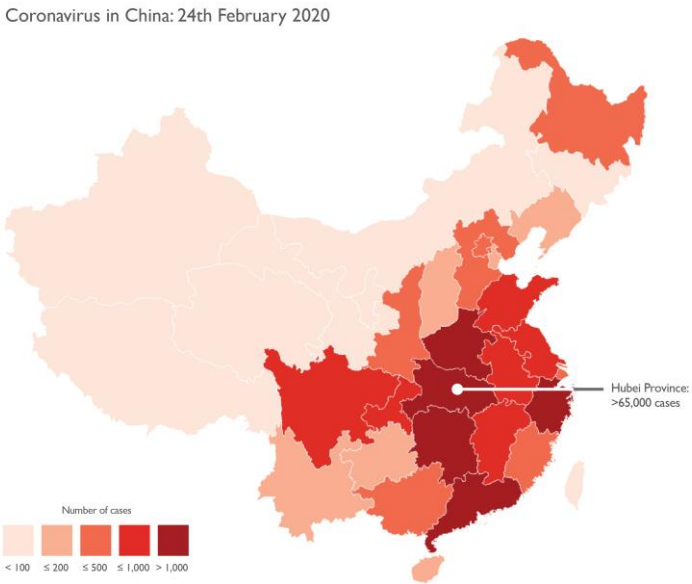
Hochschule Offenburg
offenburg.university

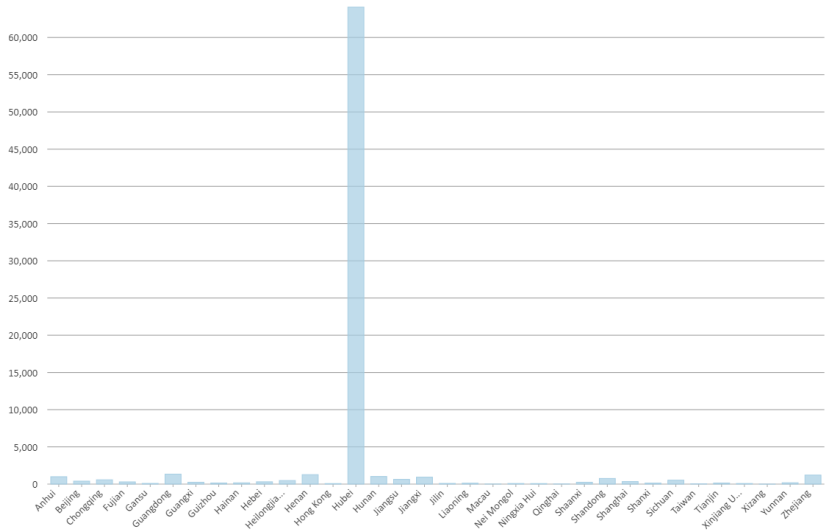
VIDEO

(VAST Challenge 2011)

Learning from Covid-19 Visualizations

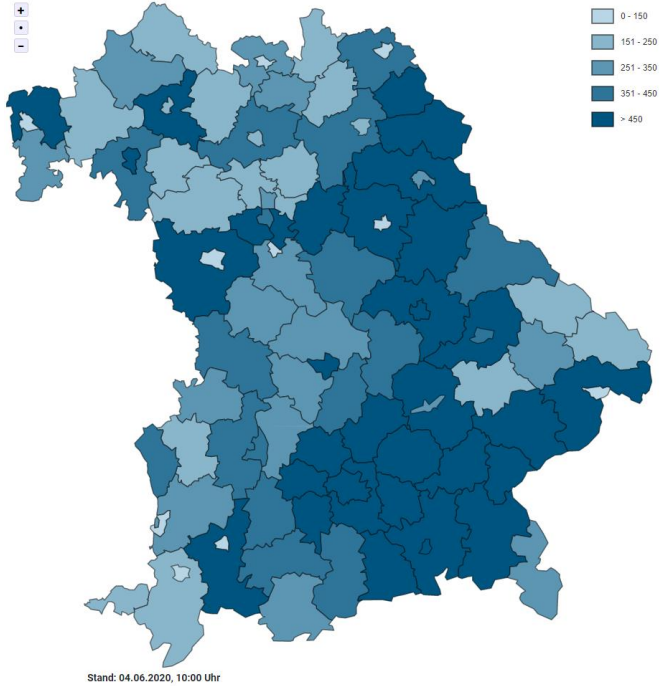
Wie unterschiedlich stark waren die einzelnen chinesischen Provinzen am 24.2.20 vom Corona-Virus betroffen?

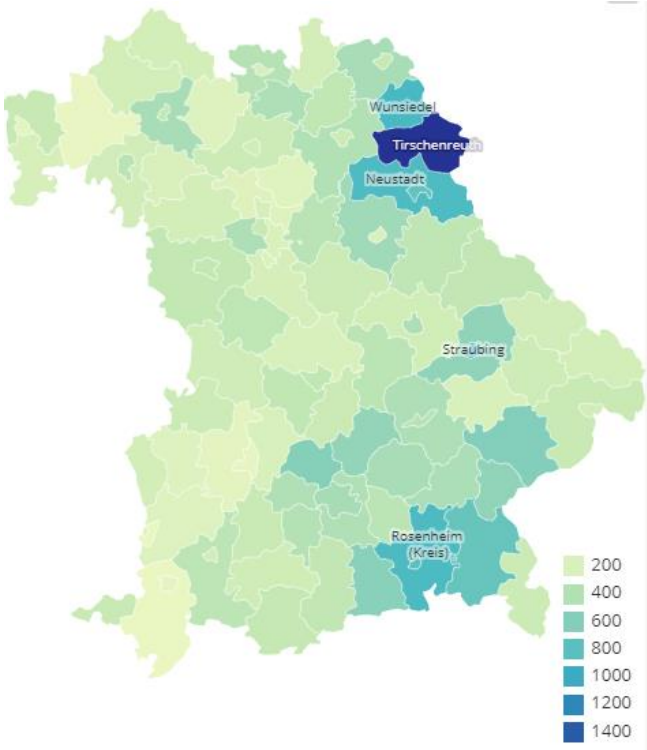




https://www.esri.com/arcgis-blog/products/product/mapping/mapping-coronavirus-responsibly/

Welche bayrischen Landkreise waren am 4.6.20 besonders stark von Corona betroffen?

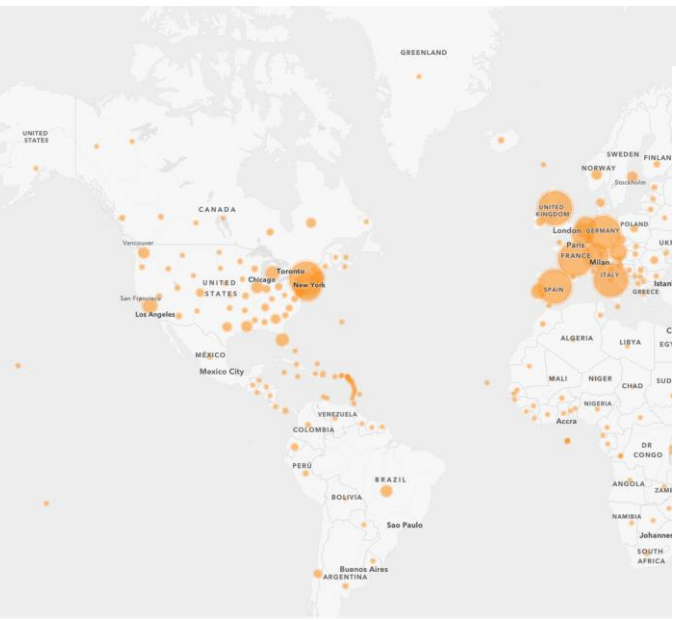




https://www.lgl.bayern.de/gesundheit/infektionsschutz/infektionskrankheiten_a_z/coronavirus/karte_coronavirus/index.htm
https://www.merkur.de/bayern/corona-bayern-karte-zahlen-faelle-covid-19-landkreise-tote-taeglich-infizierte-sars-cov-2-13610678.html

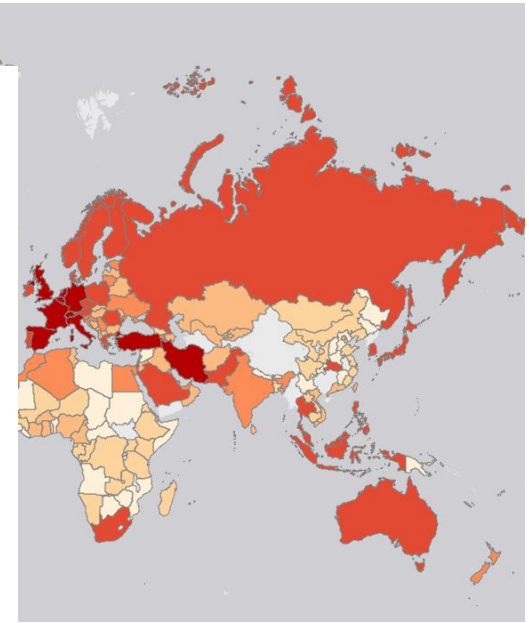
Welche Regionen der Erde waren zum Zeitpunkt des Screenshots am stärksten von Covid-19 betroffen?

Hochschule Offenburg
offenburg.university



Bundesland	Bestätigte Fälle	Fälle je 100.000	Todes-fälle	Verdopp-lung	Trend
Nordrhein-Westfalen	12.017	67	88	7 Tage	●●○
Bayern	10.180	78	59	4 Tage	●○○
Baden-Württemberg	9.729	88	101	4 Tage	●○○
Niedersachsen	3.175	40	11	6 Tage	●○○
Hessen	2.392	38	7	6 Tage	●○○
Rheinland-Pfalz	2.101	51	11	6 Tage	●○○
Berlin	1.955	54	8	5 Tage	●●○
Hamburg	1.759	96	2	5 Tage	●○○
Sachsen	1.505	37	9	5 Tage	●○○
Schleswig-Holstein	821	28	5	6 Tage	●○○
Brandenburg	632	25	3	6 Tage	●○○
Saarland	614	62	6	7 Tage	●○○
Thüringen	574	27	4	4 Tage	●○○
Sachsen-Anhalt	516	23	2	5 Tage	●●○
Mecklenburg-Vorpommern	261	16	0	12 Tage	●○○
Bremen	244	36	1	15 Tage	●○○

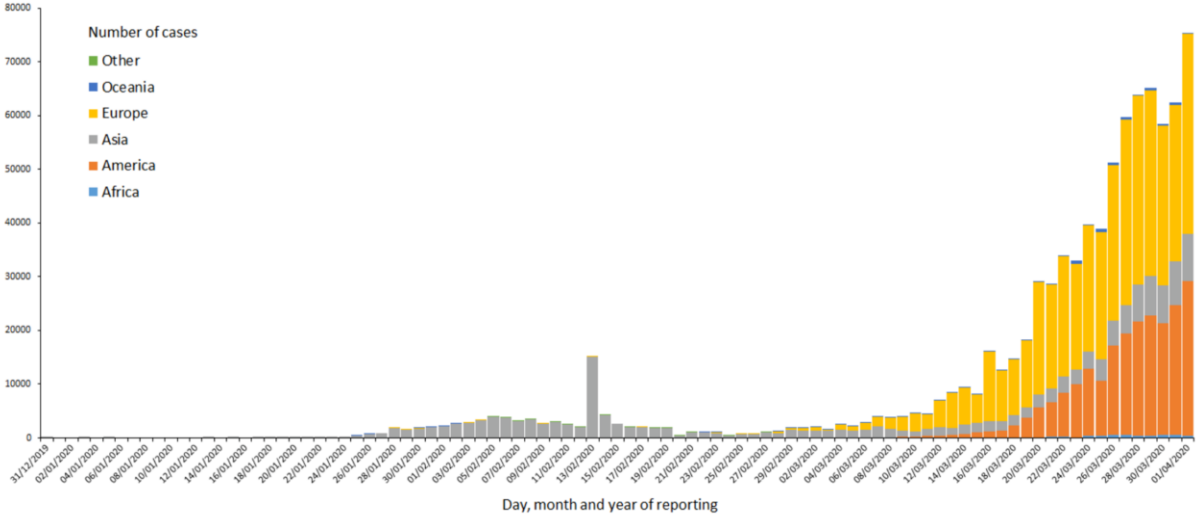
Die Verdopplungszeit gibt an, wie schnell sich die Epidemie ausbreitet. Der Trend zeigt an, wie sich dieses Tempo verändert: wird langsamer ●○○, bleibt gleich ●●○, wird schneller ●●●. Letzter Stand der Daten: 27.03.2020 20:30 Uhr

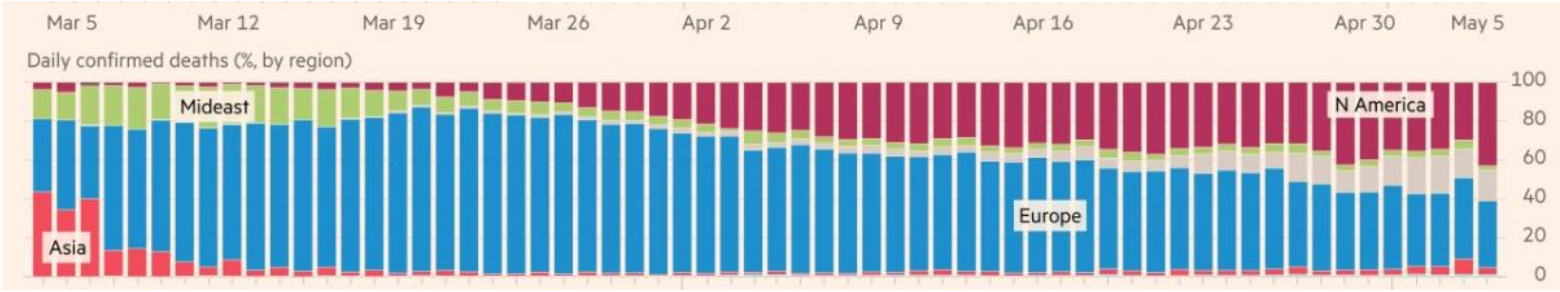


https://storymaps.arcgis.com/stories/4fdc0d03d3a34aa485de1fb0
https://www.quelle.de/stories/wissen/coronavirus-zahlen-aktuell-
https://nssac.bii.virginia.edu/covid-19/dashboard/

In welcher Region der Erde ist im betrachteten Zeitraum die stärkste Zunahme zu verzeichnen?

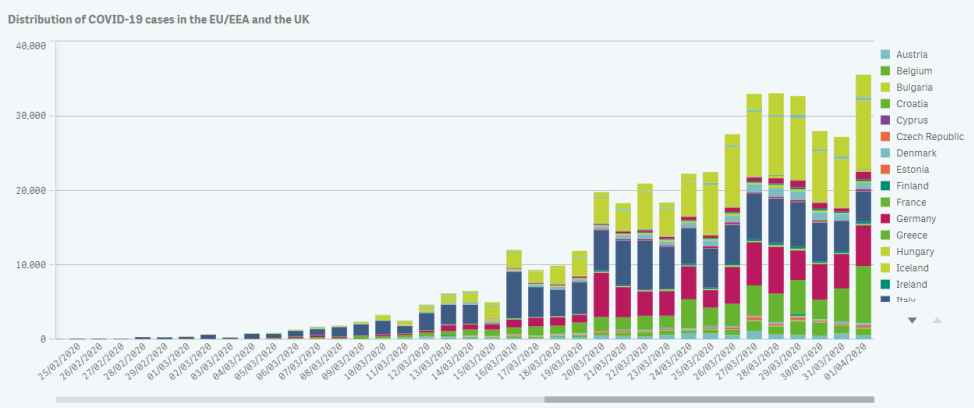
Hochschule Offenburg
offenburg.university

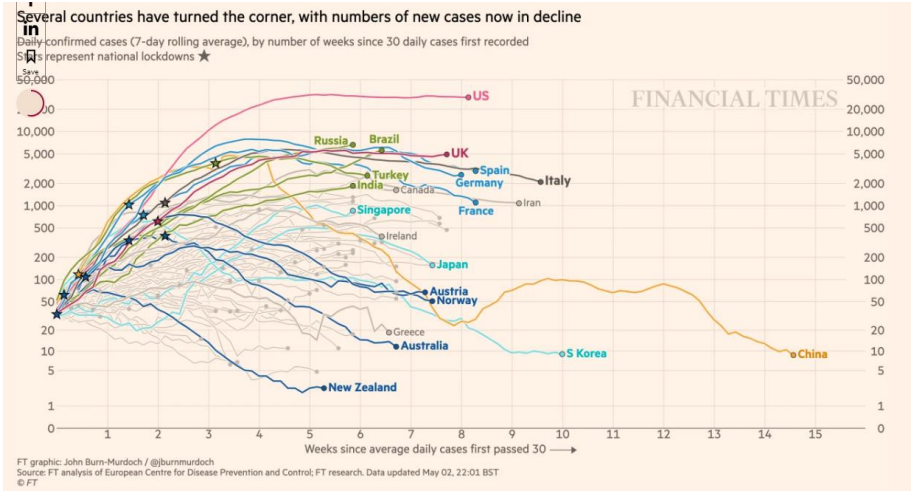




https://www.ft.com/content/a26fbf7e-48f8-11ea-aeb3-955839e06441, https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases

Gab es in Griechenland im betrachteten Zeitraum eine konstante Zunahme der Infektionszahlen oder zwischendrin Einbrüche?



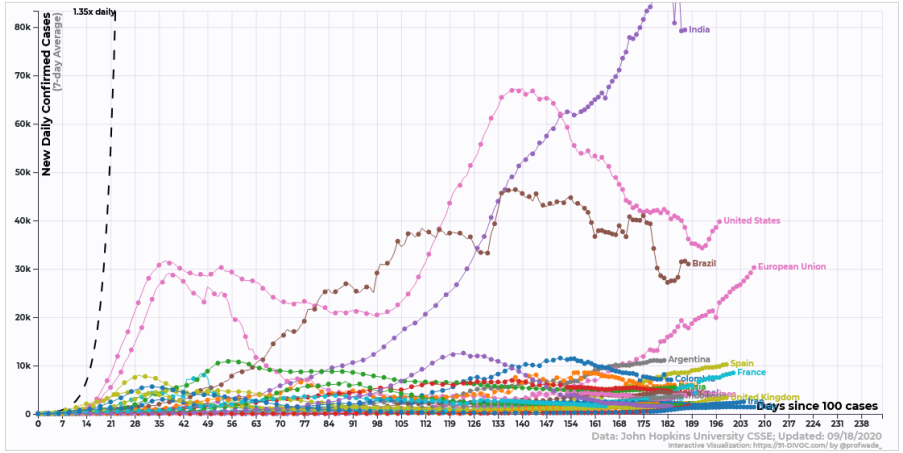


https://qap.ecdc.europa.eu/public/extensions/COVID-19/COVID-19.html, https://www.ft.com/content/a26fbf7e-48f8-11ea-aeb3-955839e06441

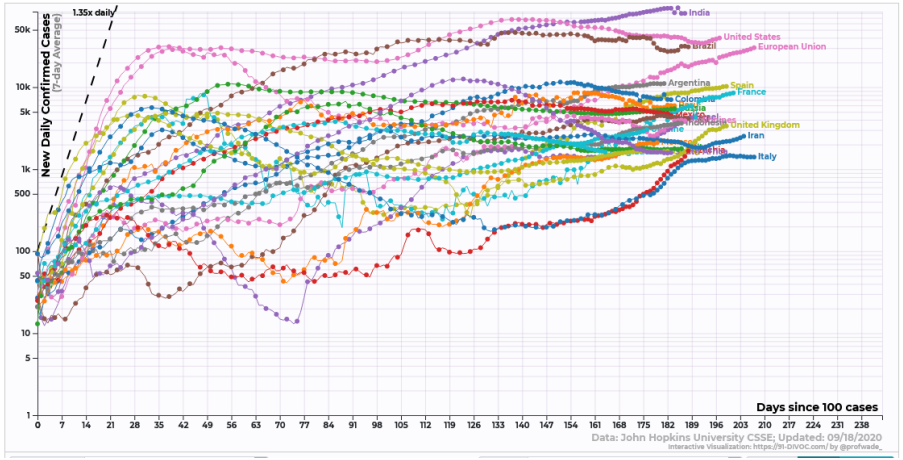
Um welchen Faktor sind die Fallzahlen in der EU zum Zeitpunkt des Erstellung der Graphik kleiner als in den USA?

Hochschule Offenburg
offenburg.university

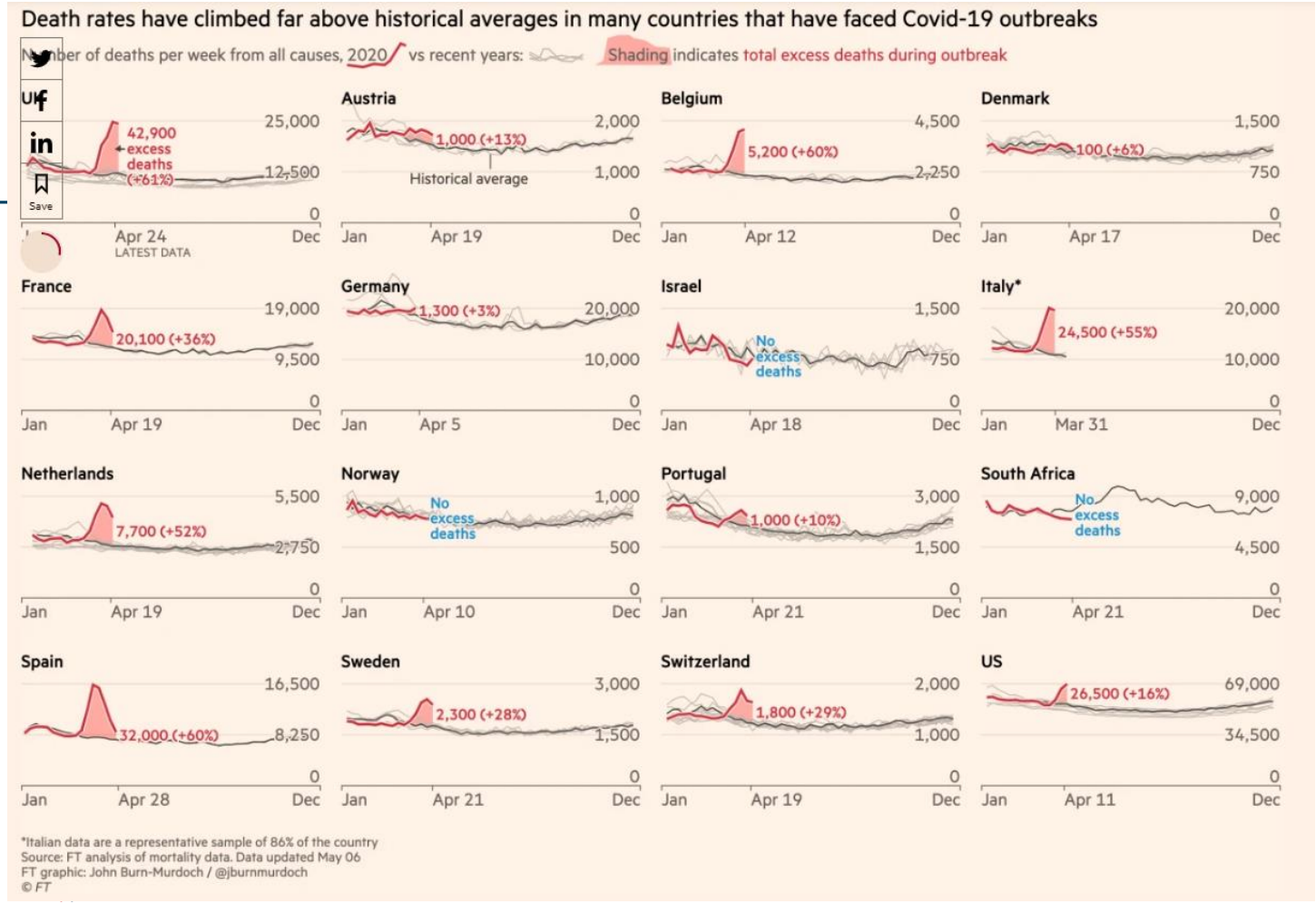
New Confirmed COVID-19 Cases per Day



New Confirmed COVID-19 Cases per Day

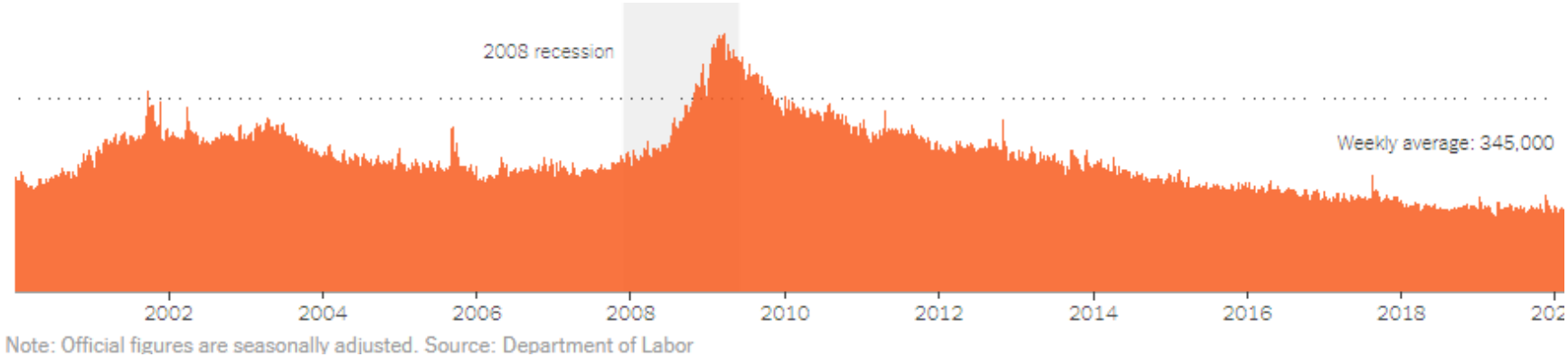


http://91-divoc.com/pages/covid-visualization/



https://www.ft.com/content/a26fbf7e-48f8-11ea-aeb3-955839e06441

Arbeitslosenzahlen USA

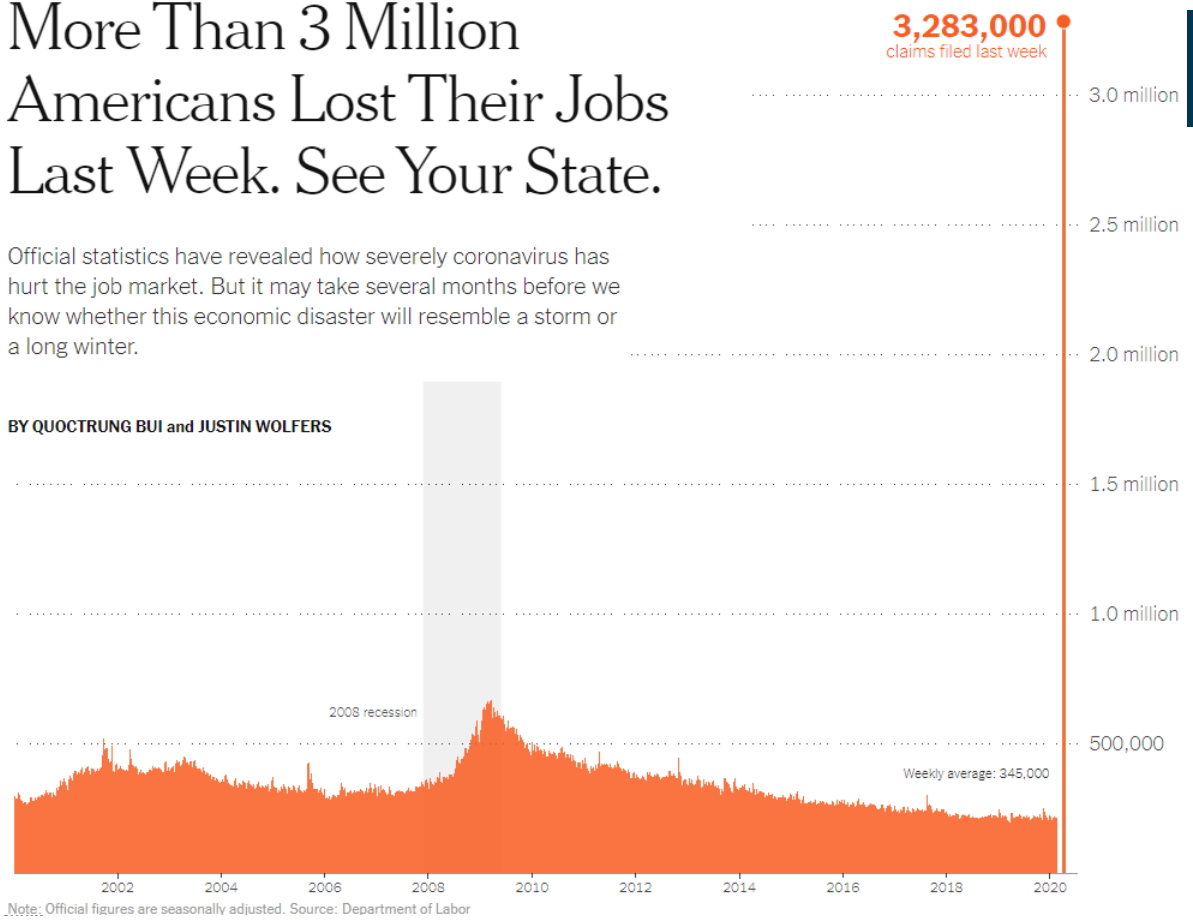


https://www.nytimes.com/interactive/2020/03/26/upshot/coronavirus-millions-unemployment-claims.html

More Than 3 Million Americans Lost Their Jobs Last Week. See Your State.

Official statistics have revealed how severely coronavirus has hurt the job market. But it may take several months before we know whether this economic disaster will resemble a storm or a long winter.

BY QUOCTRUNG BUI and JUSTIN WOLFERS



https://www.nytimes.com/interactive/2020/03/26/upshot/coronavirus-millions-unemployment-claims.html

Visualization Tools & Libraries

Visualization in Python

- **matplotlib**
- **seaborn**
- **plotly / plotly** express → interactivity
- **bokeh** → for building dashboards
- **dash** → for building dashboards

Interactive Dashboards

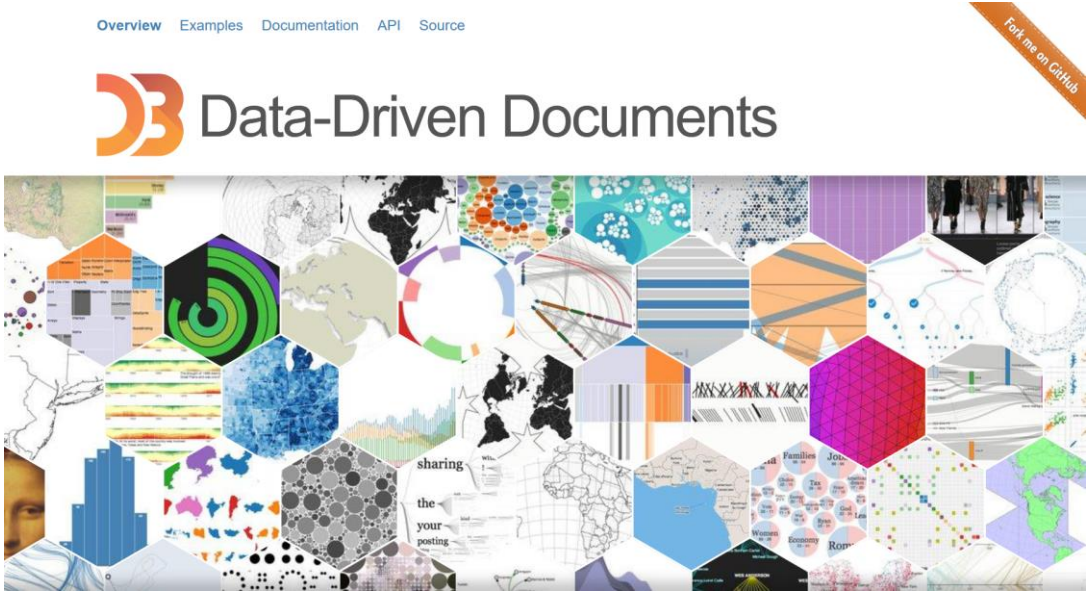
Hochschule Offenburg
offenburg.university

- Tableau
- Microsoft PowerBI
- QlikView
- SAS Visual Analytics
- …

Comparison: https://commercialtools.dbvis.de/

Visualizations in JavaScript

- Many libraries for „standard“ visualizations exist
- Use the library D3 if you need a maximum of flexibility



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