Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Semantic analysis using word embeddings and language models

Fotis Jannidis and Leonard Konle

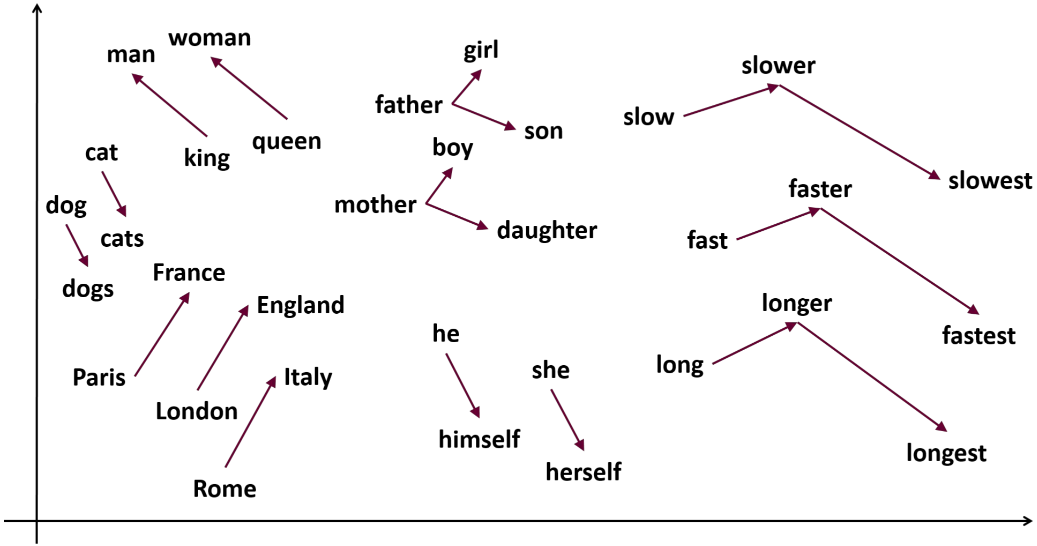
Belgrad 22.3.2022

Overview

- Part I
    - Word and text representation as vectors
    - Similarity Measurement
    - Small introduction to distributional semantics
    - Word2Vec and FastText
- Part II
    - Language Models
    - Some basic concept of deep learning
    - BERT
    - Huggingface
    - Practical part with Colab and Jupyter Notebooks

Block 1: Distributional Semantics and Word Embeddings

- Word and text representation as vectors
- Similarity Measurement
- Small introduction to distributional semantics
- Word2Vec and FastText



Word Meaning

“Meaning stands for the context of knowledge evoked by a sign, word or statement. Meaning points to the sense of a linguistic utterance. In semantics it is that which a linguistic expression or other sign gives to understand.” (Wikipedia)

agent: human, activity: understanding

agent: computer, activity: representing word meaning

starting point: classical theory of definition: „definitio fi(a)t per genus proximum et differentiam specificam”

genus proximum: generic term

differentiam specificam: the specific features which distinguish this class from other classes under the generic germ

‘bachelor’: an adult unmarried male

WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: house [Search WordNet]

Display Options: (Select option to change) [Change]

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Display options for word: word (sense key)

Noun

- S: (n) **house (house%1:06:00::)** (a dwelling that serves as living quarters for one or more families) *"he has a house on Cape Cod"; "she felt she had to get out of the house"*
    - *direct hyponym* / *full hyponym*
        - S: (n) beach house (beach_house%1:06:00::) (a house built on or near a beach)
        - S: (n) boarding house (boarding_house%1:06:00::), boardinghouse (boardinghouse%1:06:00::) (a private house that provides accommodations and meals for paying guests)
        - S: (n) bungalow (bungalow%1:06:00::), cottage (cottage%1:06:00::) (a small house with a single story)
    - *part meronym*
        - S: (n) library (library%1:06:01::) (a room where books are kept) *"they had brandy in the library"*
        - S: (n) loft (loft%1:06:00::), attic (attic%1:06:00::), garret (garret%1:06:00::) (floor consisting of open space at the top of a house just below roof; often used for storage)
        - S: (n) porch (porch%1:06:00::) (a structure attached to the exterior of a building often forming a covered entrance)
        - S: (n) study (study%1:06:00::) (a room used for reading and writing and studying) *"he knocked lightly on the closed door of the study"*
    - *direct hypernym* / *inherited hypernym* / *sister term*
        - S: (n) dwelling (dwelling%1:06:00::), home (home%1:06:00::), domicile (domicile%1:06:00::), abode (abode%1:06:00::), habitation (habitation%1:06:00::), dwelling house (dwelling_house%1:06:00::) (housing that someone is living in) *"he built a modest dwelling near the pond"; "they raise money to provide homes for the homeless"*
        - S: (n) building (building%1:06:00::), edifice (edifice%1:06:00::) (a

A word is represented as a node in a network, connecting it to hypernyms, hyponyms, synonyms, meronyms etc.

text and word representations

- Text similarity
  - + easy to evaluate
  - + many useful applications
  - - meaning of a text only captured as a relation to other texts
- In this context texts are usually modeled as a bag of words (bow) in a document-term matrix:

	bank	furniture	window	money	house	room	interest
$text_1$	2	2	3	1	5	5	0
$text_2$	5	0	1	4	2	0	4

text vector

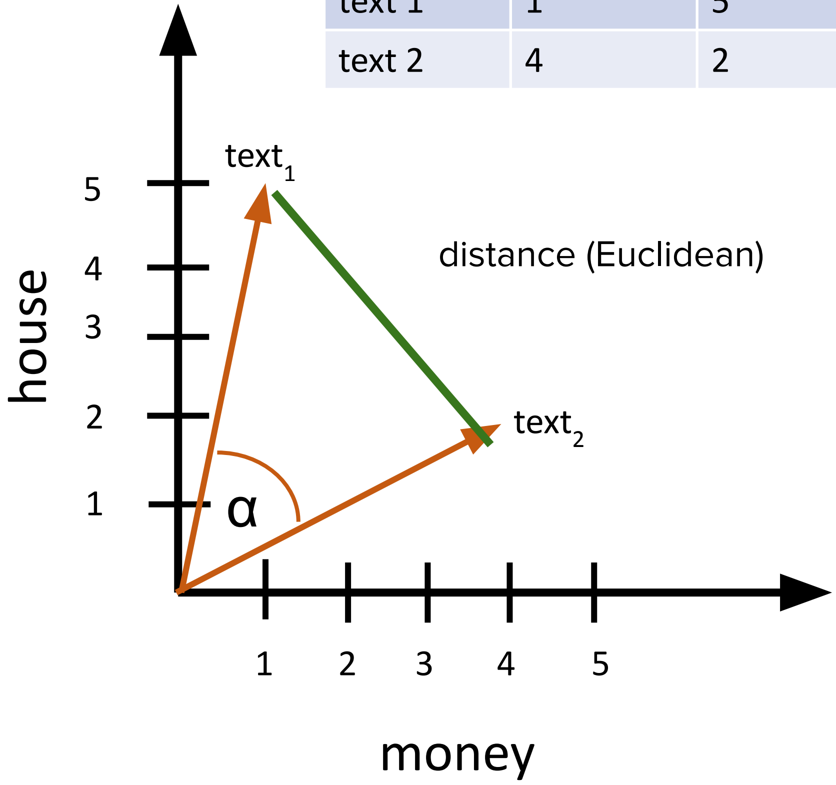
word vector

vector is here just a term for a series of numbers

Text similarity

- But we can interpret the series of numbers as coordinates in space
- So a text can be represented by a vector consisting of the counts over all words in a corpus
- this vector can be viewed as a point in vector space (more exact: as a vector from the origin to the point)
- Text similarity can be modeled as the distance between the points
- Best measure for distance is the cosine of the angle α between the vectors

	money	house
text 1	1	5
text 2	4	2



Word meaning and context

“Before their lives violently intersected, two men who were shot to death and the man the police believe killed them had all fought the same scourge” New York Times 21.3.22

Basic intuition

- The meaning of a word can be understood by looking at the words which come up together with the word.

„You shall know a word by the company it keeps” (Firth 1957)

„examine the syntagmatic environments in which a word occurs, and you shall know more about the kind of word you are dealing with.” (Geeraerts 2009)

- Central concept ‘collocation’: ‘a lexical relation between two or more words which have a tendency to co-occur within a few words of each other in running text’ (Stubbs 2002: 24)
- „In corpus linguistics, a **collocation** is a sequence of words or terms that co-occur more often than would be expected by chance.” (engl. Wikipedia 14.11.2017)

Word similarity

- A vector over a whole text is not a very good representation, loss of specifitity
- Instead a context for a focus word is defined, for example 3 words to the left and 3 words to the right. On this basis we can create a new matrix, a word-context matrix, with the focus words as rows and the context words as columns:

Talk to me, my lovely child! your father is here

I am here, my father. Your child is right here.

introduced us to her husband and her lovely child, which came running

This creates a word – cooccurrence matrix:

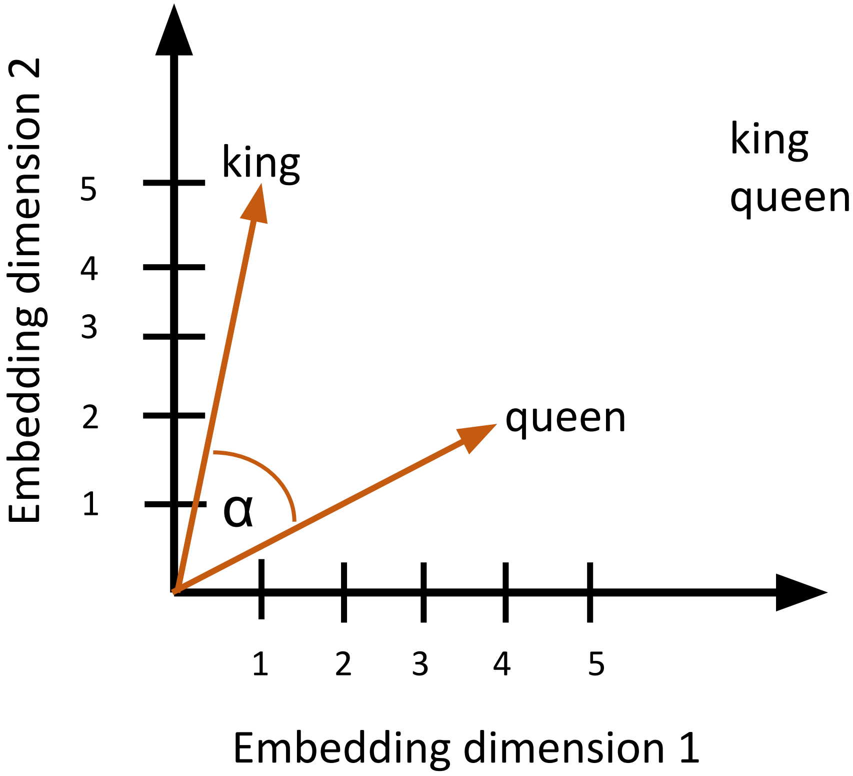
	And	Father	My	lovely	Is	Me	My	your
child	1	2	2	2	1	1	1	2

Depending on the size of the context this results still in a very large and very sparse (=many zeros) matrix

word2vec

- Word2vec (Mikolov et al. 2013) unsupervised machine learning using a shallow neural net and a huge amount of unlabeled training data
- word2vec produces a dense vector representation of words, usually just 100-300 numbers
- in contrast to a word-context matrix we have no idea about the meaning of the numbers
- But: The word meaning and the relationships between words are still encoded spatially

Word Embedding



2-dimensional vector

king = [1, 5]
queen = [4, 2]

Usually word embeddings have 100-300 dimensions (=numbers)

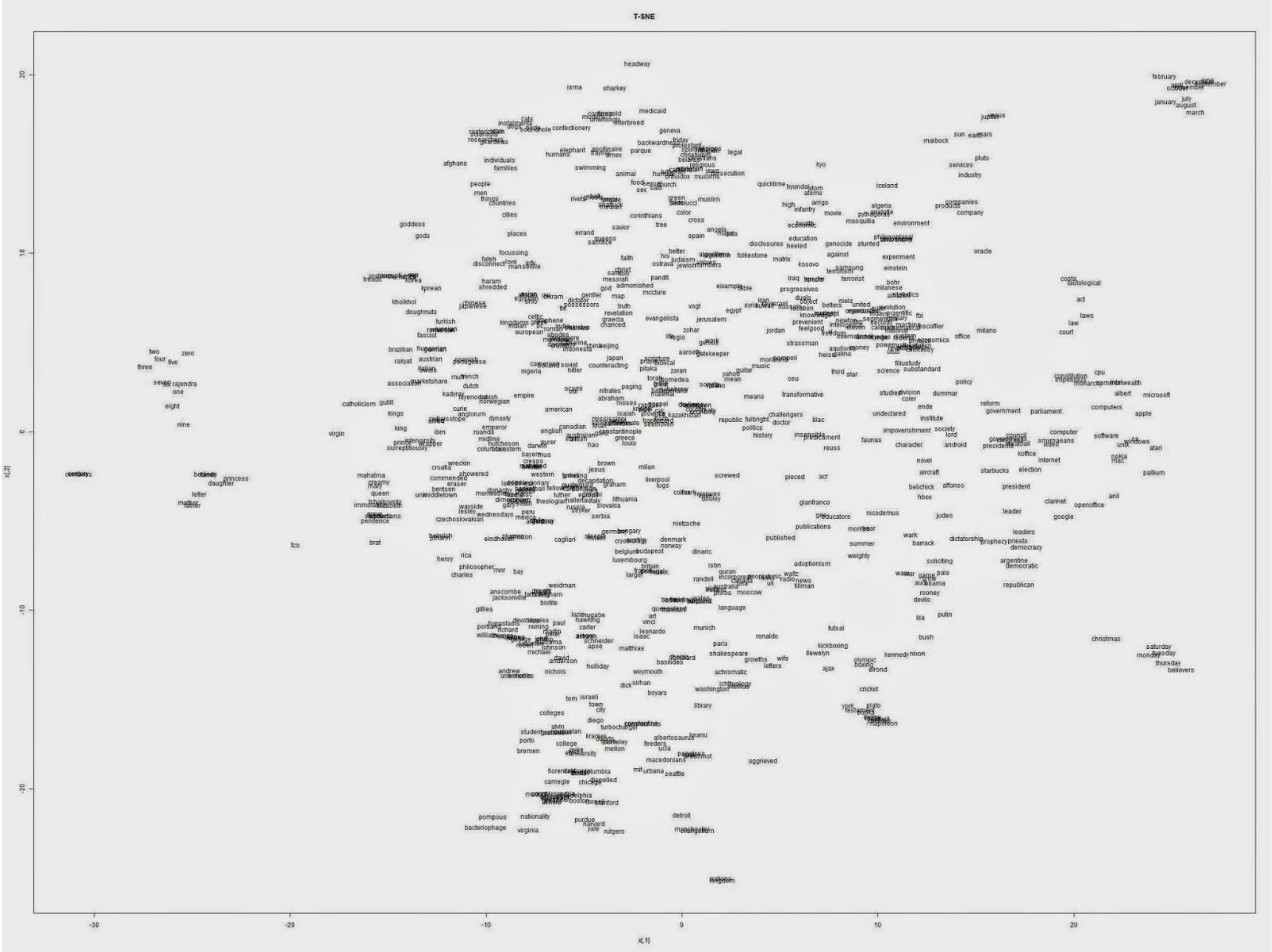
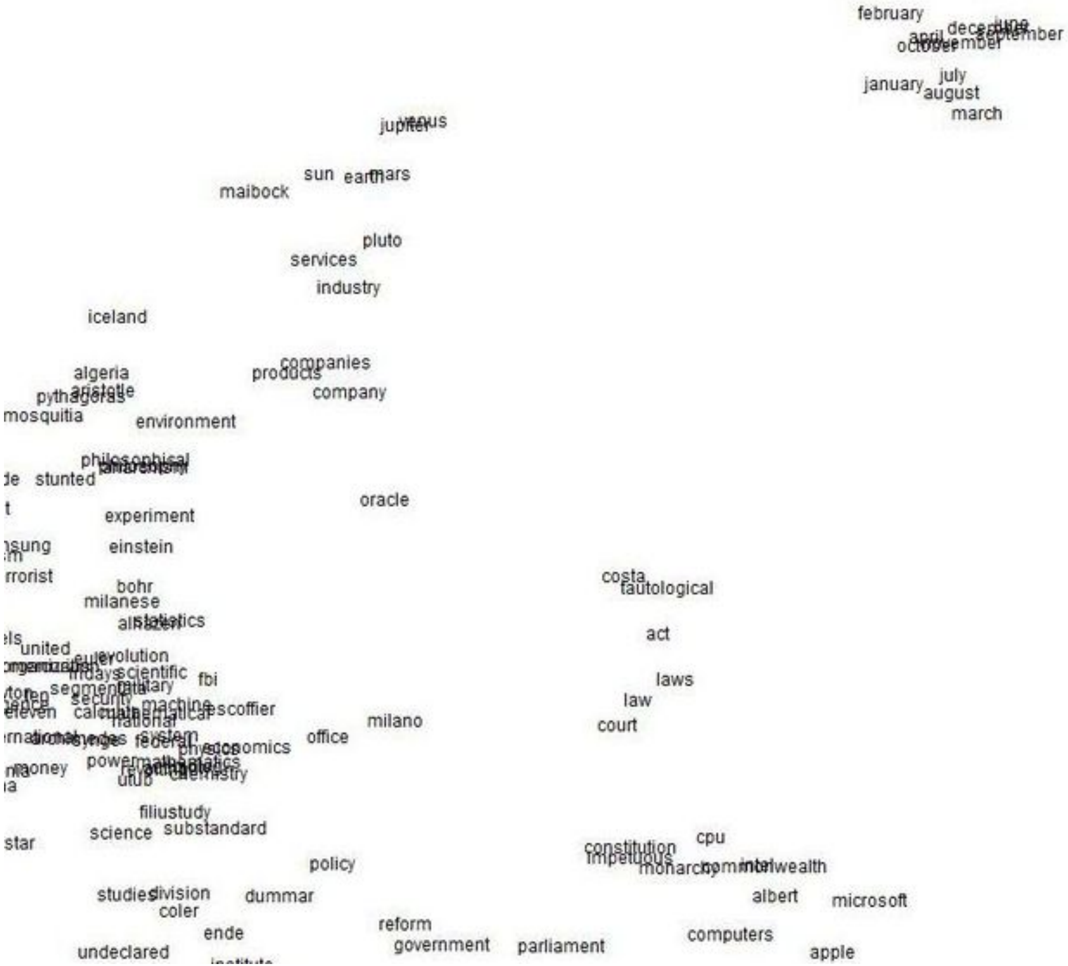


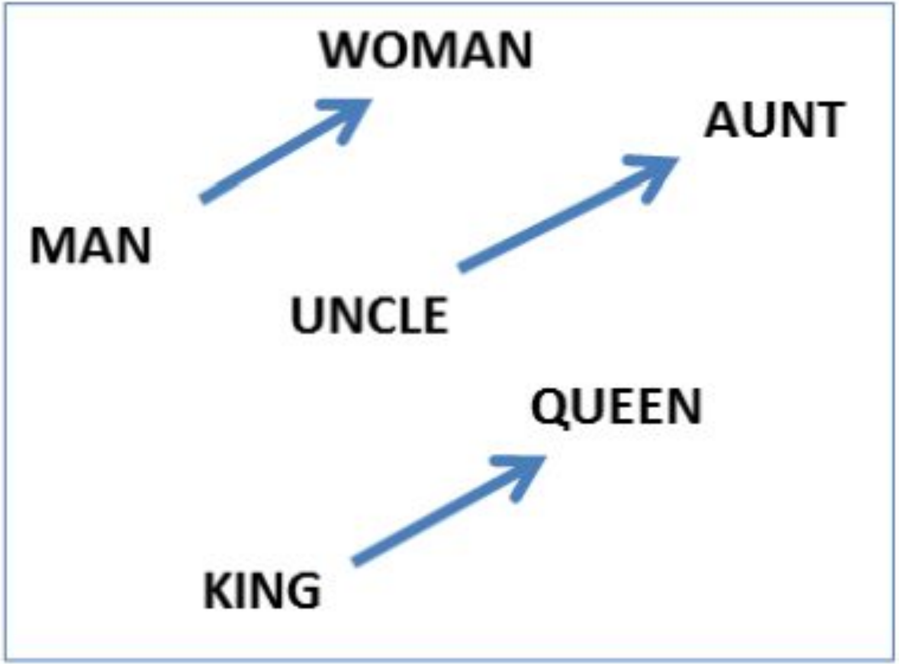
Image: http://learningaboutdata.blogspot.fr/2014/06/plotting-word-embedding-using-tsne-with.html



Spatial proximity indicates semantic similarity

Directions in vector space represent language information

Gender



Plural

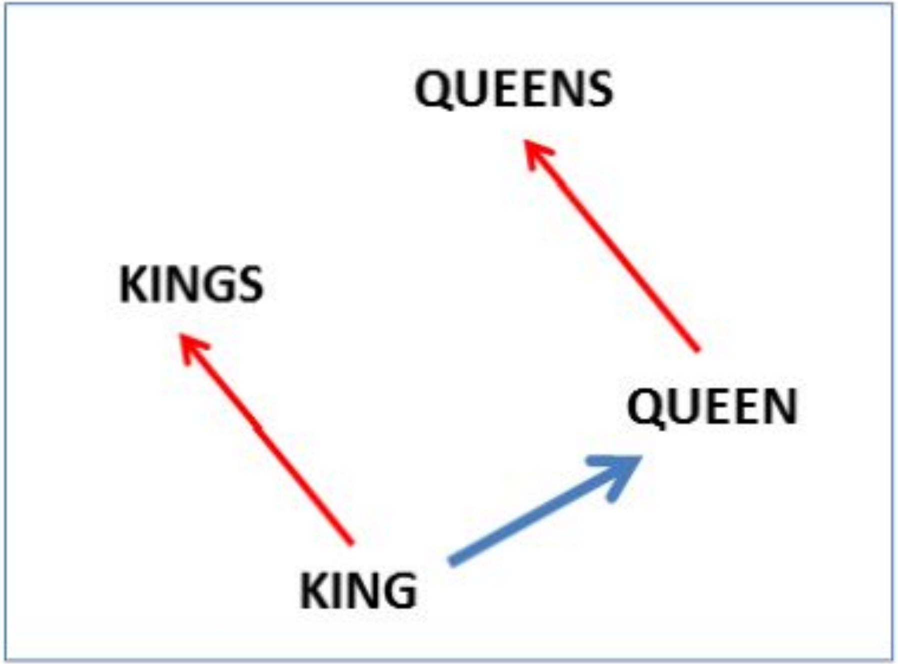
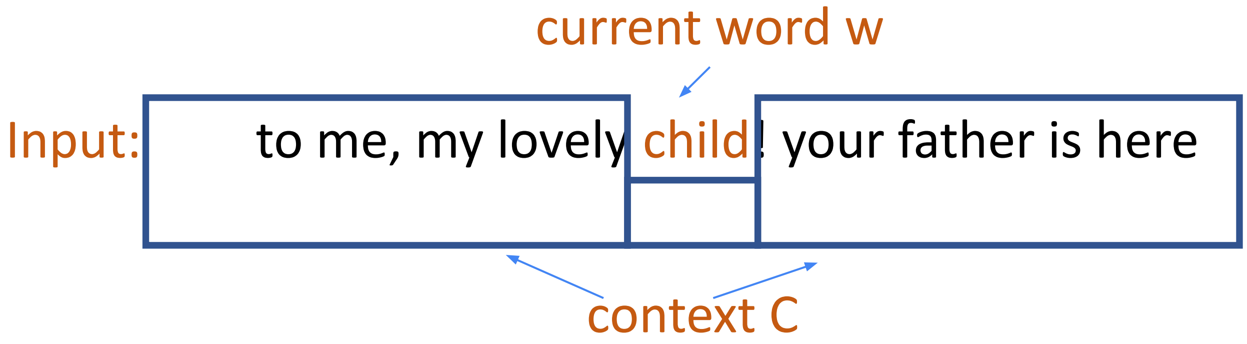


Image: Mikolov, Yih, Zweig 2013

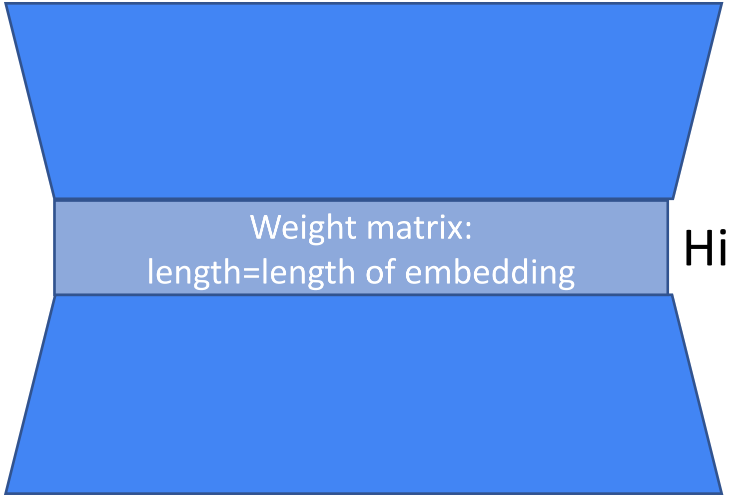
Creating word embeddings with word2vec



- Input is read sequentially. Each word becomes the current word and then its context is retrieved:
  w=child: C = {father, is, here, lovely, me, my, to, your}
- The task: predict the focus word given the context words
- More exact: Predict the probability for each word to be the focus word

Recurrent neural network with one hidden layer

Input: word sequences
(words in contexts)



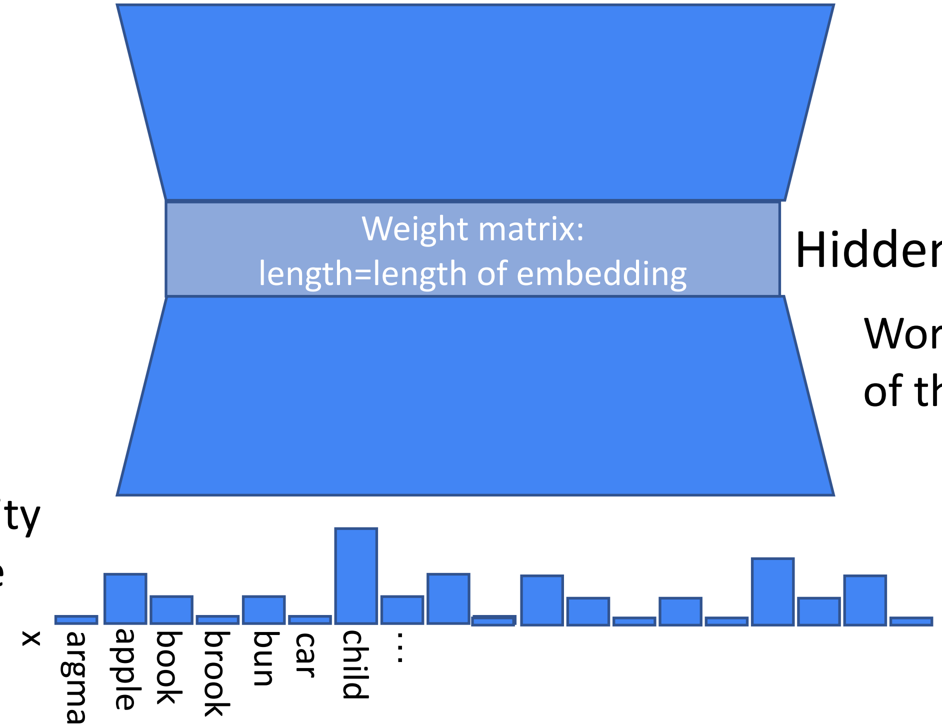
Hidden layer

Word vectors are the weights
of the hidden layer

Output: 1) prediction of context words C given current word w (CBOW)
2) prediction of current word w given the context C (skipgram)

Recurrent neural network with one hidden layer

Input: word sequences
(words in contexts)



Hidden layer

Word vectors are the weights of the hidden layer

CBOW: p(w|C)
Output is a probability distribution over the whole vocabulary

Output

word embeddings - milestones

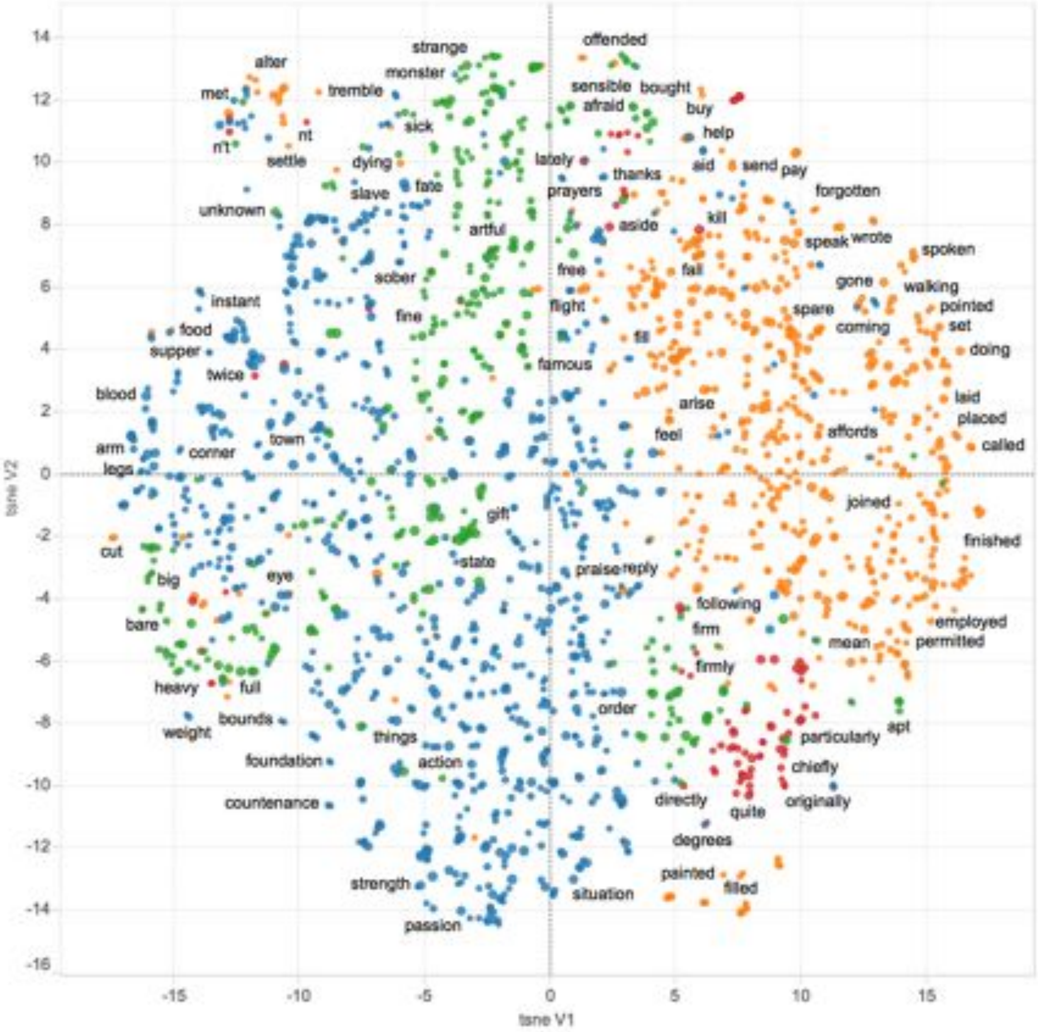
- word2vec (Mikolov et al. 2013)
- Glove (Pennington et al. 2014)
- Fasttext (Bojanowski et al. 2016)
- Pretrained models for 157 languages (Grave et al. 2018)
- Elmo (Peters et al. 2018)
- Bert (Devlin et al. 2018) (dynamic emb.)
- GPT-3 (Brown et al. 2020) (dynamic emb.)

Fasttext

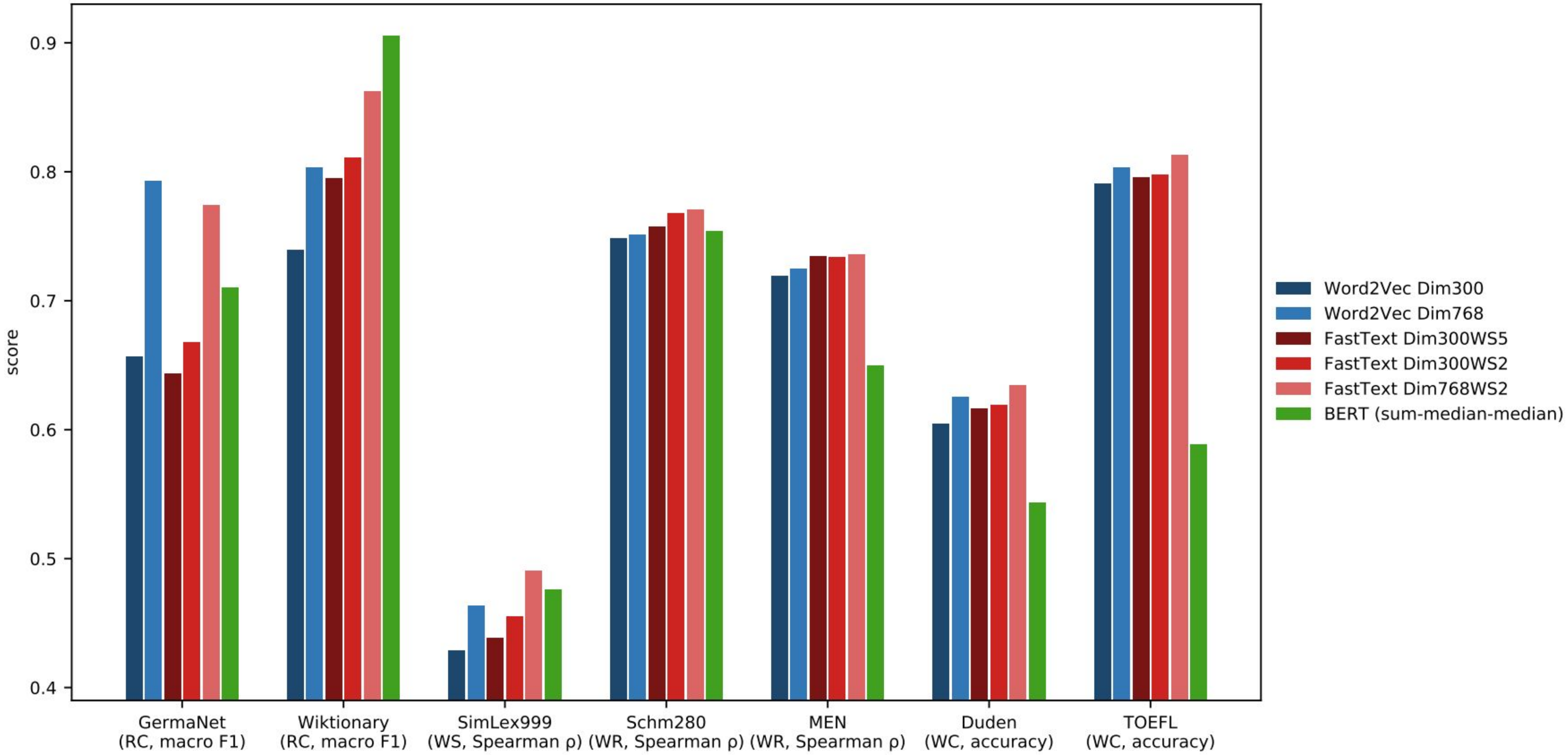
- based on word2vecs skipgram
- each character n-gram (with n between 3 and 6) is associated with a vector
- <> are indicating the beginning and end of a word and are added to each word
- W = ‘where’ and n= 3 ():
- <wh, whe, her, ere, re> <where>
- each word is represented as the sum of the word and the character n-grams, n>2 and n<7
- Adds subword information, for example morphological information, to the model
- Allows a reasonable representation of out-of-vocabulary words based on n-grams
- Code is available
- Since 2018 word embeddings for 157 languages available, based on Wikipedia and Common Crawl (questionable quality)

Word classes are more important than semantic similarity:

nouns, verbs, adjectives, adverbs



http://ryanheuser.org/word-vectors-4/



Ehrmanntraut et al. 2021

Demo 1

https://colab.research.google.com/drive/1jgSXhQuzLIPyM8ncKd56JlfnLV0OoQEX?usp=sharing

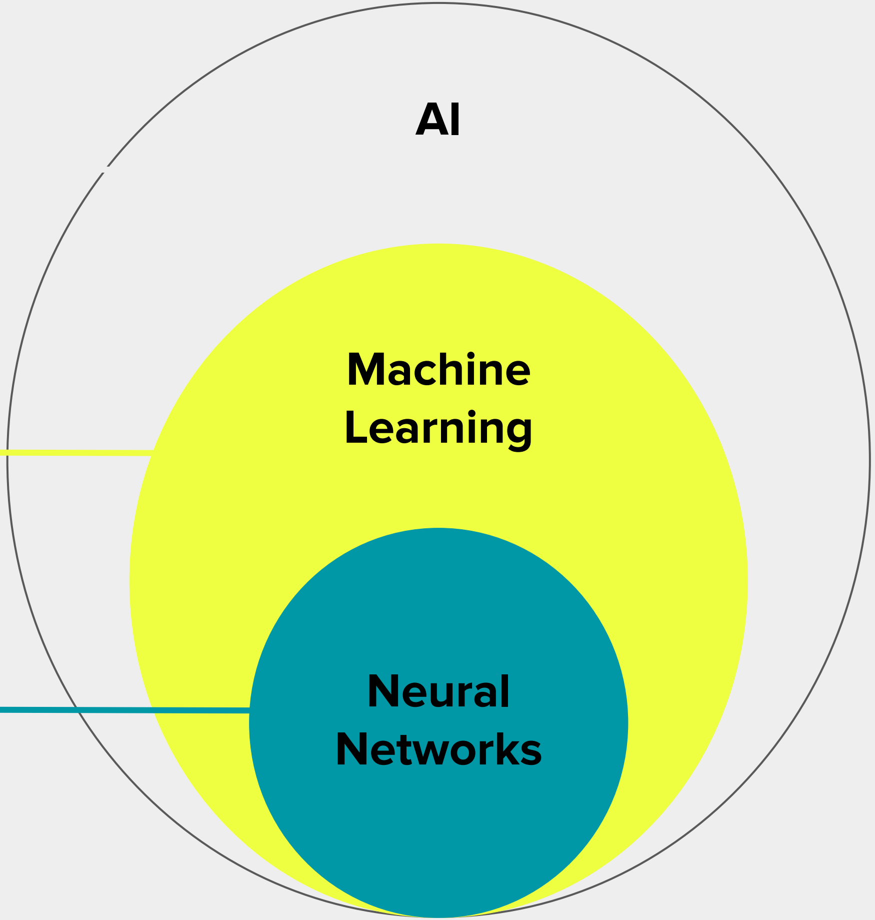
Pretrained Language Models

Machine Learning, Deep Learning & AI

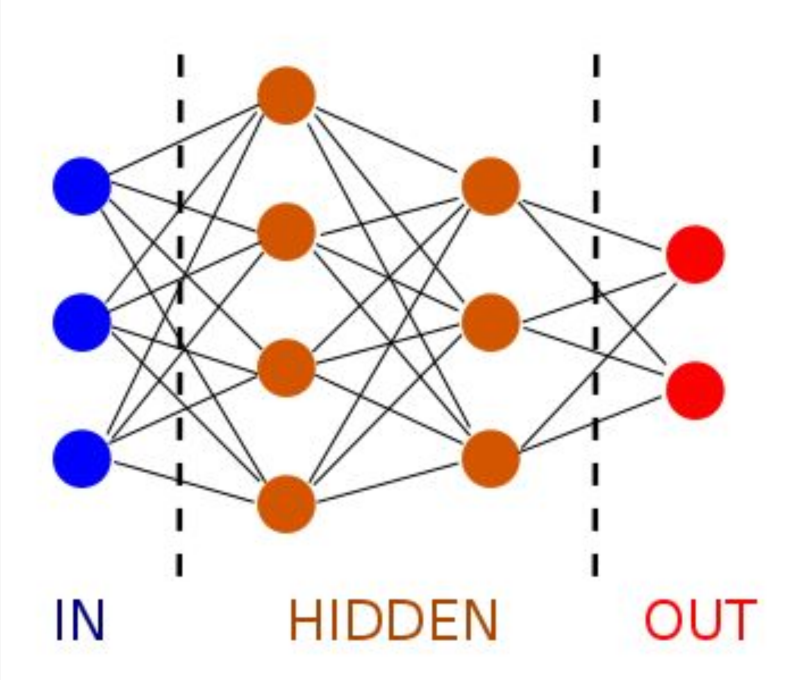
Simulation of human decision structures by algorithms in order to solve problems as autonomously as possible.

Implicit replication of these structures by adaptation of algorithms using examples

Distribution of the learning process to a net structure



Neural Networks



Fully-Connected Feedforward Network

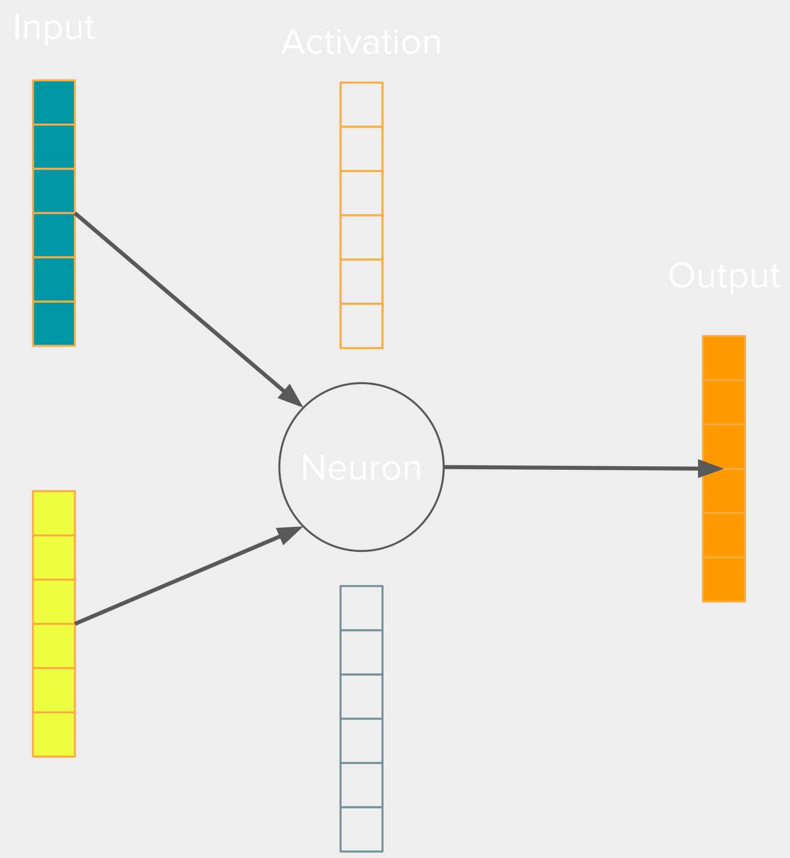
Neural Nets - Neurons

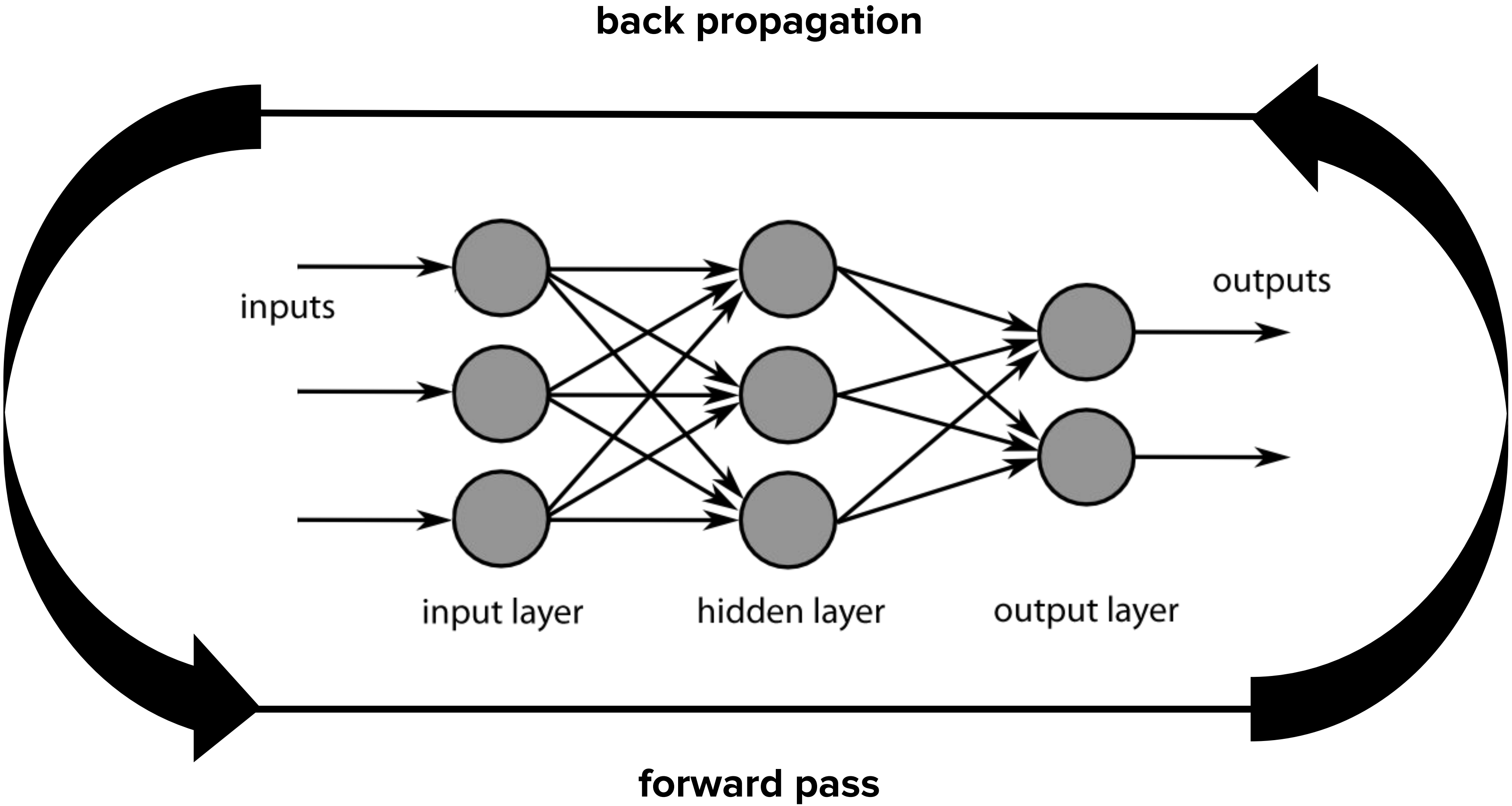
Each neuron consists of two functions:

1) the input x is multiplied with weights w (and a bias b is added to the result)
2) The output of 1) is input to an activation function, which is non-linear and the reason any kind of function can be modeled via a neural net

$$y = act(wx + b)$$

act: activation function
x: Input
w: weights
b: bias




back propagation
inputs
outputs
input layer
hidden layer
output layer
forward pass

Neural Networks - Key Terms

- **Neuron**: Smallest unit in networks
- **Layer**: A set of parallel neurons
- **Task**: Problem to be solved
- **Batch**: Number of examples before a backpropagation
- **Epoch**: One loop over all examples
- **Loss**: Distance between optimal result and output of the network

BERT

Bidirectional **E**ncoder **R**epresentations from **T**ransformers

BERT - Task

- BERTs Task is Masked Language Modeling (MLM)
- Basically a cloze test

Ernie is an orange Muppet character created and originally performed by Jim Henson for the long-running children's television show *Sesame Street*. He and his roommate Bert form the comic duo Bert and Ernie, one of the program's centerpieces, with Ernie acting the role of the naïve troublemaker, and Bert the world weary foil.

Ernie is an ______ Muppet character created and originally performed by Jim Henson for the long-running children's tele____ show *Sesame Street*. He and ___ roommate Bert form the comic duo Bert and Ernie, one of the program's centerpieces, with Ernie acting the role of the naïve troublemaker, and ____ the world weary foil.

BERT - Task

World Knowledge

Ernie is an orange Muppet character created and originally performed by Jim Henson for the long-running children's television show *Sesame Street*. He and his roommate Bert form the comic duo Bert and Ernie, one of the program's centerpieces, with Ernie acting the role of the naïve troublemaker, and Bert the world weary foil.

Grammar

Vocabulary

Basic Reasoning

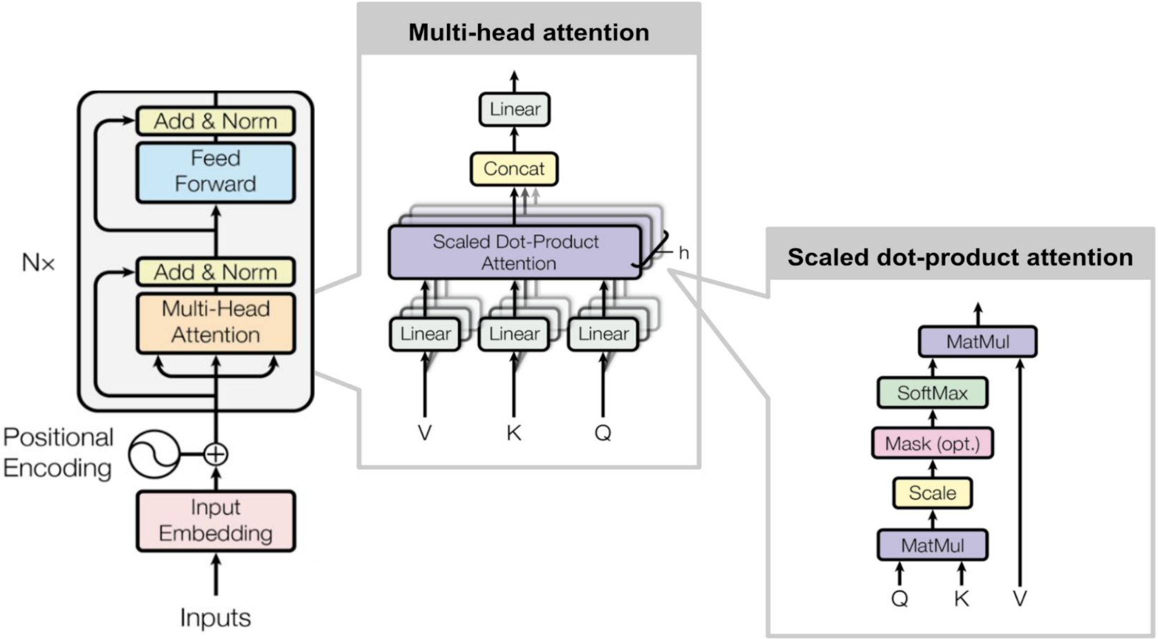
Demo: https://huggingface.co/distilbert-base-uncased?text=Ernie+is+an+%5BMASK%5D+Muppet+character+created+and+orriginally+performed+by+Jim+Henderson+for+the+long-running+children%27s+television+show+Sesame+Street.

BERT - Tokenization

ernie,is,an,orange,mu,##ppet,character,created,and,originally,performed,by,jim,hen,##son,for,the,long,-,running,childr en,',s,television,show,ses,##ame,street,.,he,and,his,room,##mate,bert,form,the,comic,duo,bert,and,ernie,,,one,of,the,p rogram,',s,center,##piece,##s,,,with,ernie,acting,the,role,of,the,nai,##ve,trouble,##maker,,,and,bert,the,world,wear,##y, foi,##l,.

- No classic word tokenization
- Instead tokenization based on 30.000 word pieces
  - Reduces cloze filling complexity
  - Idea: Which choice of words allows the representation of a corpus as the shortest possible chain
- If a word is not in the list of word pieces, it’s composed out of multiple word pieces

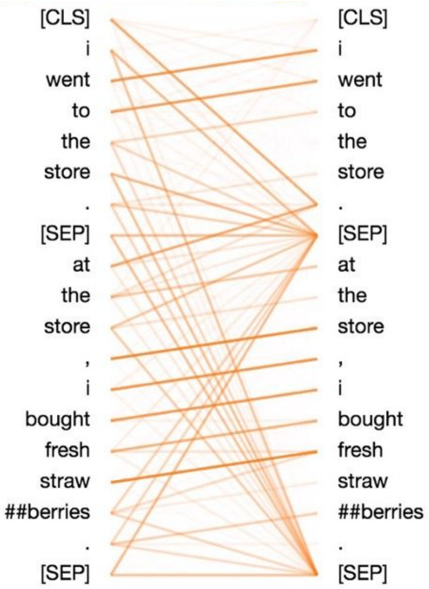
BERT - Network



The Transformer Layer

BERT - Network

- Each word is related to itself and all other words in an input.
- This is done 12 times per layer
- 12 layers in sequence[1]
- Resulting in 11M Parameters ~ 1.3GB



Attention Mechanism

[1] $\text{Bert}_{\text{Base}}$ Model

BERT - Trainingdata

- Huge amounts of:
  - Webtext
  - Forums
  - Wikis
  - Online Newspaper
  - Books
- Original Bert:
  - Google Book Corpus: 11.000 books (5GB)
  - English Wikipedia: 6.000.000 Articles (40GB)
- Best German Bert:
  - 163 GB (mostly german common crawl)

Cost of training one Bert Model: ~6000€ (4 days)

Why is BERT useful?

- No one really needs a neural cloze test solver, but:
    - Similar to word2vec we can use its inner <u>representation</u> for
        - Words (not worth it)
        - Sentences
        - Paragraphs
    - Make use of world knowledge, grammar, vocabulary to train
        - Document Classification
        - NER
        - Sentiment
        - …

BERT can be seen as a compressed representation of all texts it's been trained on.

BERT Fine-Tuning

Plain Text

MLM Task

General Language Model

Pretraining

BERT Fine-Tuning

Plain Text

Texts + Labels
(Sentiment, NER, etc)

General Language Model

MLM Task

Productive Model

Fine-Tuning

Pretraining

The Domain Problem

- Bert learns from modern webtext, newspapers etc.
- Typically DH deals with literary text and or texts older than webtext
  - Results in a difference between pretraining and application in:
    - Vocabulary
    - Orthography
    - Style
    - Semantic
    - Required World Knowledge

BUT: Pretrained Language Models still achieve best results even in forgein domains.

AND: We can alter Models to fit our needs (Domain Adaptation)

BERT domain adaptation

Plain Text

Domain text
(same text type, time)

MLM Task

General Language Model

Second MLM Task

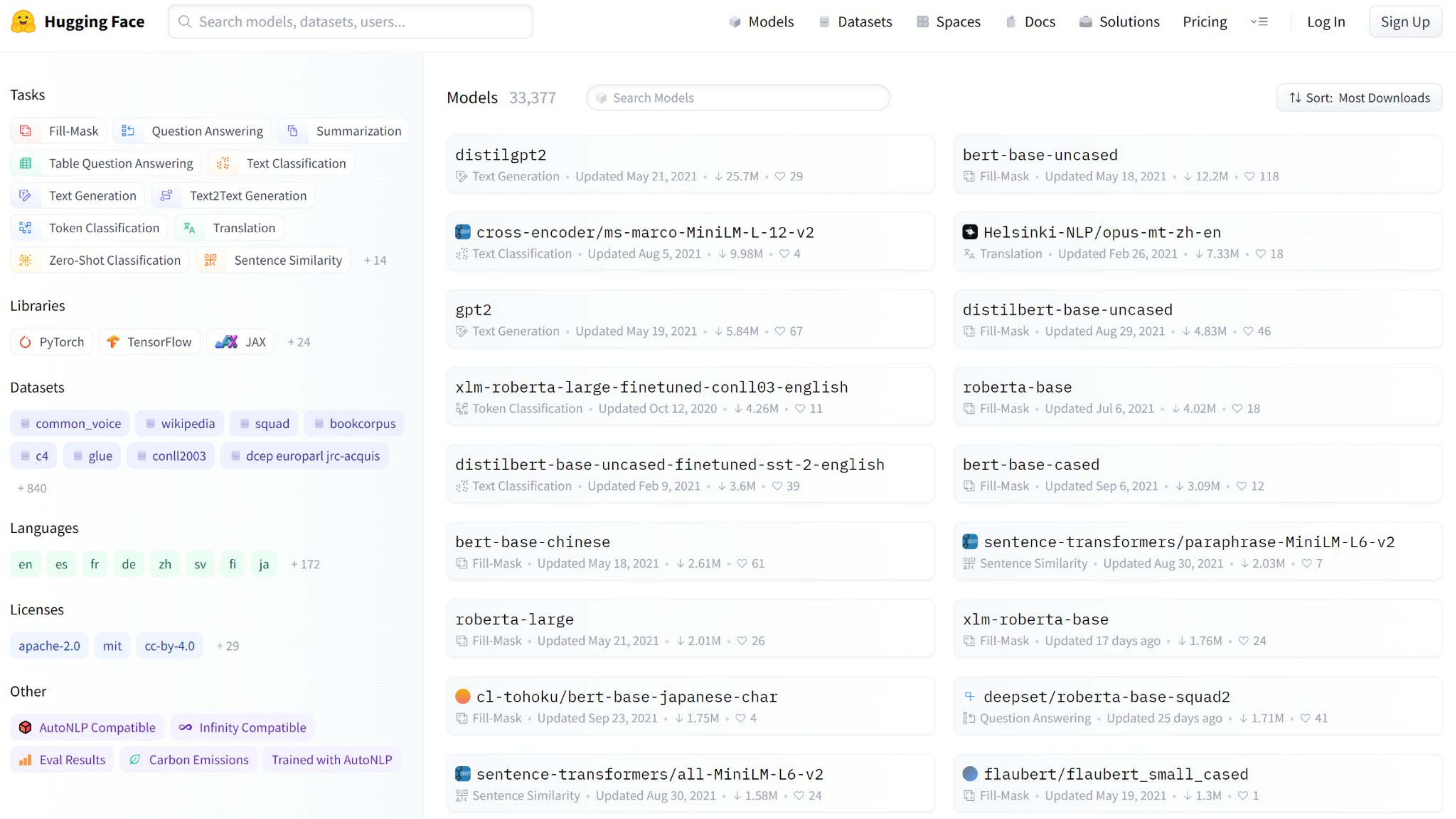
Domain Model

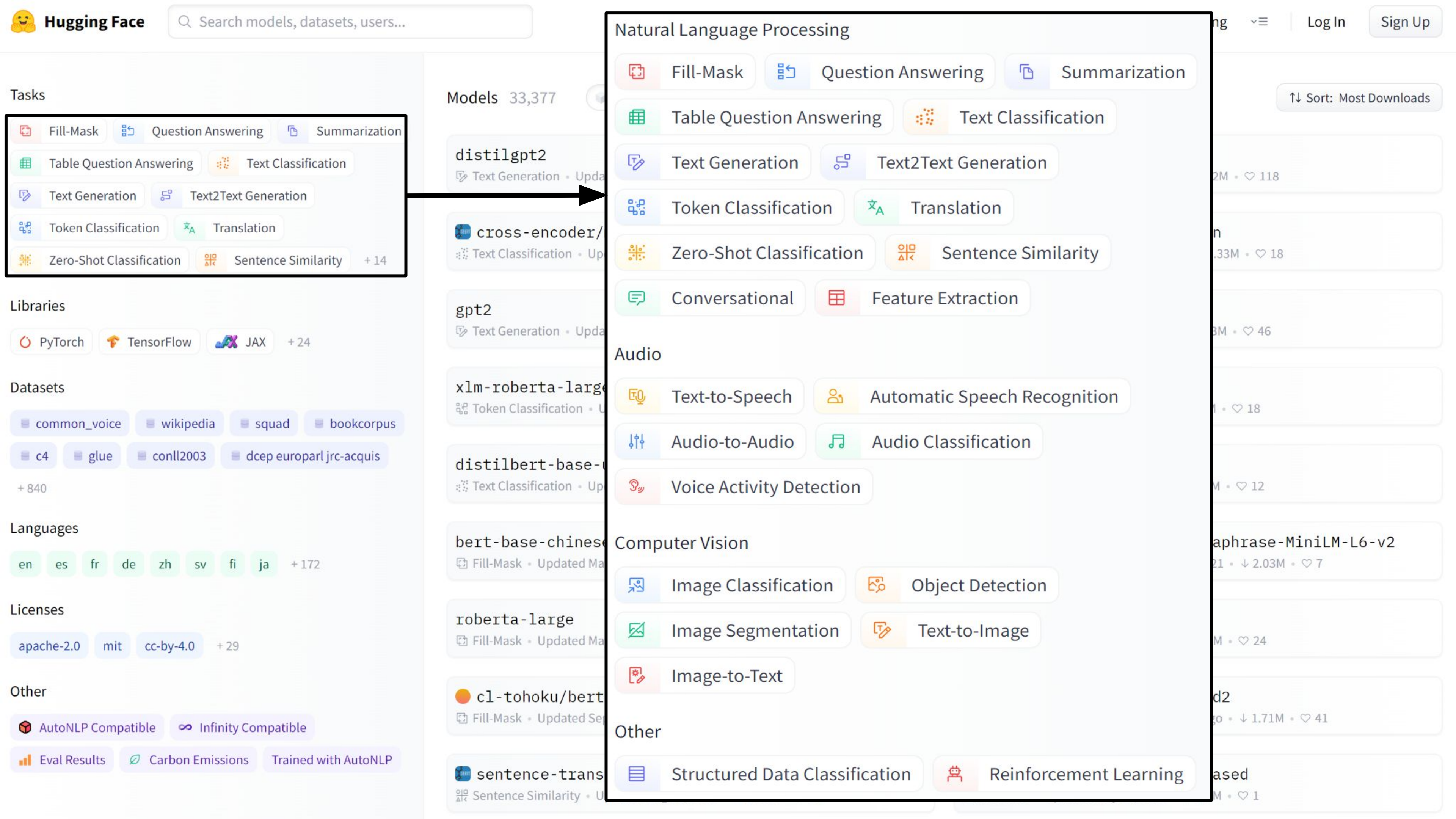
Domain Adaptation

Pretraining

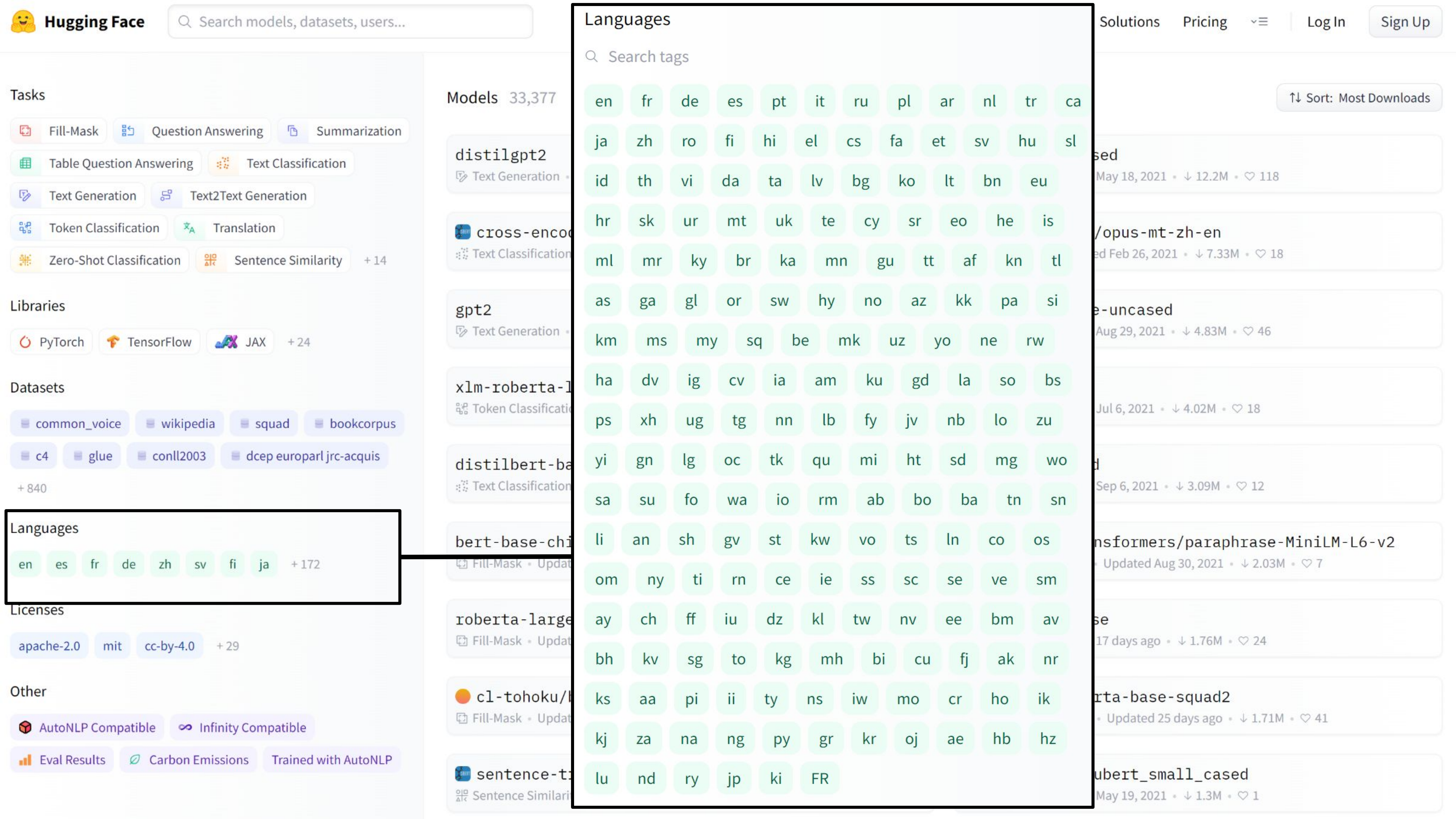
HuggingFace

- Python Packages
  - transformers: Train, Fine-Tune, Usage of Language Models
  - tokenizers: Train and apply Word Piece Tokenizer
- Modelhub
  - Free Repository for general and fine-tuned Language Models
- Datasets
  - Free Repository with standardized Training Datasets (MLM and FineTuning)


Hugging Face
Search models, datasets, users...
Models
Datasets
Spaces
Docs
Solutions
Pricing
Log In
Sign Up
Tasks
Fill-Mask
Question Answering
Summarization
Table Question Answering
Text Classification
Text Generation
Text2Text Generation
Token Classification
Translation
Zero-Shot Classification
Sentence Similarity
+ 14
Libraries
PyTorch
TensorFlow
JAX
+ 24
Datasets
common_voice
wikipedia
squad
bookcorpus
c4
glue
conll2003
dcep europarl jrc-acquis
+ 840
Languages
en
es
fr
de
zh
sv
fi
ja
+ 172
Licenses
apache-2.0
mit
cc-by-4.0
+ 29
Other
AutoNLP Compatible
Infinity Compatible
Eval Results
Carbon Emissions
Trained with AutoNLP
Models 33,377
Search Models
Sort: Most Downloads
distilgpt2
Text Generation • Updated May 21, 2021 • 25.7M • 29
bert-base-uncased
Fill-Mask • Updated May 18, 2021 • 12.2M • 118
cross-encoder/ms-marco-MiniLM-L-12-v2
Text Classification • Updated Aug 5, 2021 • 9.98M • 4
Helsinki-NLP/opus-mt-zh-en
Translation • Updated Feb 26, 2021 • 7.33M • 18
gpt2
Text Generation • Updated May 19, 2021 • 5.84M • 67
distilbert-base-uncased
Fill-Mask • Updated Aug 29, 2021 • 4.83M • 46
xlm-roberta-large-finetuned-conll03-english
Token Classification • Updated Oct 12, 2020 • 4.26M • 11
roberta-base
Fill-Mask • Updated Jul 6, 2021 • 4.02M • 18
distilbert-base-uncased-finetuned-sst-2-english
Text Classification • Updated Feb 9, 2021 • 3.6M • 39
bert-base-cased
Fill-Mask • Updated Sep 6, 2021 • 3.09M • 12
bert-base-chinese
Fill-Mask • Updated May 18, 2021 • 2.61M • 61
sentence-transformers/paraphrase-MiniLM-L6-v2
Sentence Similarity • Updated Aug 30, 2021 • 2.03M • 7
roberta-large
Fill-Mask • Updated May 21, 2021 • 2.01M • 26
xlm-roberta-base
Fill-Mask • Updated 17 days ago • 1.76M • 24
cl-tohoku/bert-base-japanese-char
Fill-Mask • Updated Sep 23, 2021 • 1.75M • 4
deepset/roberta-base-squad2
Question Answering • Updated 25 days ago • 1.71M • 41
sentence-transformers/all-MiniLM-L6-v2
Sentence Similarity • Updated Aug 30, 2021 • 1.58M • 24
flaubert/flaubert_small_cased
Fill-Mask • Updated May 19, 2021 • 1.3M • 1



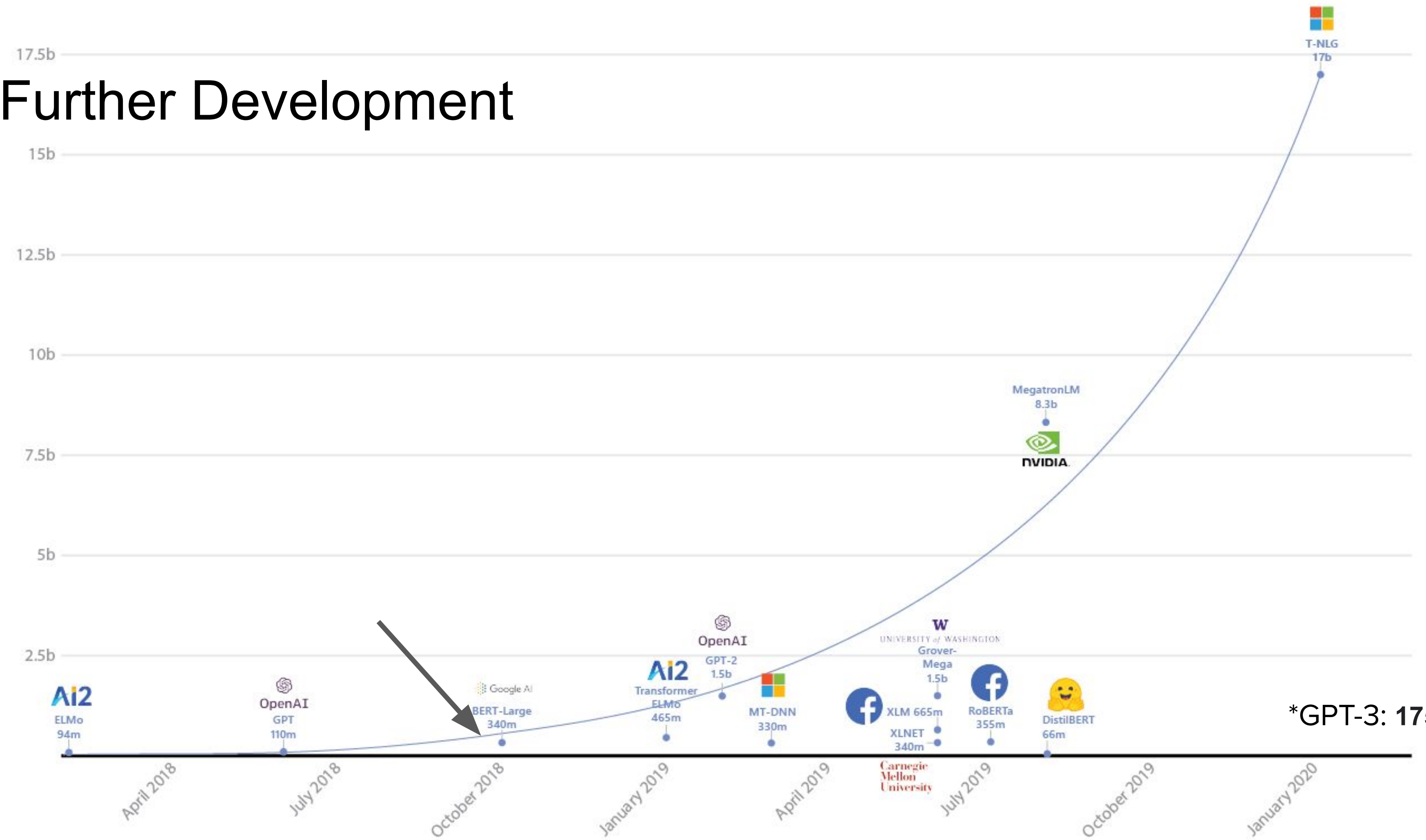
Hugging Face
Search models, datasets, users...
Log In
Sign Up
Tasks
Fill-Mask
Question Answering
Summarization
Table Question Answering
Text Classification
Text Generation
Text2Text Generation
Token Classification
Translation
Zero-Shot Classification
Sentence Similarity
+ 14
Libraries
PyTorch
TensorFlow
JAX
+ 24
Datasets
common_voice
wikipedia
squad
bookcorpus
c4
glue
conll2003
dcep europarl jrc-acquis
+ 840
Languages
en
es
fr
de
zh
sv
fi
ja
+ 172
Licenses
apache-2.0
mit
cc-by-4.0
+ 29
Other
AutoNLP Compatible
Infinity Compatible
Eval Results
Carbon Emissions
Trained with AutoNLP
Models 33,377
Sort: Most Downloads
distilgpt2
cross-encoder/
gpt2
xlm-roberta-large
distilbert-base-
bert-base-chinese
roberta-large
cl-tohoku/bert
sentence-trans
Natural Language Processing
Fill-Mask
Question Answering
Summarization
Table Question Answering
Text Classification
Text Generation
Text2Text Generation
Token Classification
Translation
Zero-Shot Classification
Sentence Similarity
Conversational
Feature Extraction
Audio
Text-to-Speech
Automatic Speech Recognition
Audio-to-Audio
Audio Classification
Voice Activity Detection
Computer Vision
Image Classification
Object Detection
Image Segmentation
Text-to-Image
Image-to-Text
Other
Structured Data Classification
Reinforcement Learning


Hugging Face
Search models, datasets, users...
Solutions
Pricing
Log In
Sign Up
Tasks
Fill-Mask
Question Answering
Summarization
Table Question Answering
Text Classification
Text Generation
Text2Text Generation
Token Classification
Translation
Zero-Shot Classification
Sentence Similarity
+ 14
Libraries
PyTorch
TensorFlow
JAX
+ 24
Datasets
common_voice
wikipedia
squad
bookcorpus
c4
glue
conll2003
dcep europarl jrc-acquis
+ 840
Languages
en es fr de zh sv fi ja + 172
Licenses
apache-2.0
mit
cc-by-4.0
+ 29
Other
AutoNLP Compatible
Infinity Compatible
Eval Results
Carbon Emissions
Trained with AutoNLP
Models 33,377
Sort: Most Downloads
distilgpt2
Text Generation
cross-enco
Text Classification
gpt2
Text Generation
xlm-roberta-l
Token Classificatio
distilbert-ba
Text Classification
bert-base-chi
Fill-Mask • Updat
roberta-large
Fill-Mask • Updat
cl-tohoku/b
Fill-Mask • Updat
sentence-tr
Sentence Similari
sed
May 18, 2021 • 12.2M • 118
/opus-mt-zh-en
ed Feb 26, 2021 • 7.33M • 18
e-uncased
Aug 29, 2021 • 4.83M • 46
Jul 6, 2021 • 4.02M • 18
d
Sep 6, 2021 • 3.09M • 12
nsformers/paraphrase-MiniLM-L6-v2
Updated Aug 30, 2021 • 2.03M • 7
se
17 days ago • 1.76M • 24
rta-base-squad2
Updated 25 days ago • 1.71M • 41
ubert_small_cased
May 19, 2021 • 1.3M • 1
Languages
Search tags
en fr de es pt it ru pl ar nl tr ca
ja zh ro fi hi el cs fa et sv hu sl
id th vi da ta lv bg ko lt bn eu
hr sk ur mt uk te cy sr eo he is
ml mr ky br ka mn gu tt af kn tl
as ga gl or sw hy no az kk pa si
km ms my sq be mk uz yo ne rw
ha dv ig cv ia am ku gd la so bs
ps xh ug tg nn lb fy jv nb lo zu
yi gn lg oc tk qu mi ht sd mg wo
sa su fo wa io rm ab bo ba tn sn
li an sh gv st kw vo ts ln co os
om ny ti rn ce ie ss sc se ve sm
ay ch ff iu dz kl tw nv ee bm av
bh kv sg to kg mh bi cu fj ak nr
ks aa pi ii ty ns iw mo cr ho ik
kj za na ng py gr kr oj ae hb hz
lu nd ry jp ki FR

Huggingface Models

- https://huggingface.co/Babelscape/wikineural-multilingual-ner
  - Multilingual NER (de, en, es, fr, it, nl, pl, pt, ru)
- https://huggingface.co/csebuetnlp/mT5_multilingual_XLSum
  - Multilingual Text Summarization
- https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment
  - Text Sentiment Analysis (en, fr, de, es, nl)
- https://huggingface.co/sentence-transformers/paraphrase-xlm-r-multilingual-v1

Further Development



*GPT-3: **175b**

Demo Task 1 - Sentiment Analysis

Task: Classify the Sentiment of a Sequence

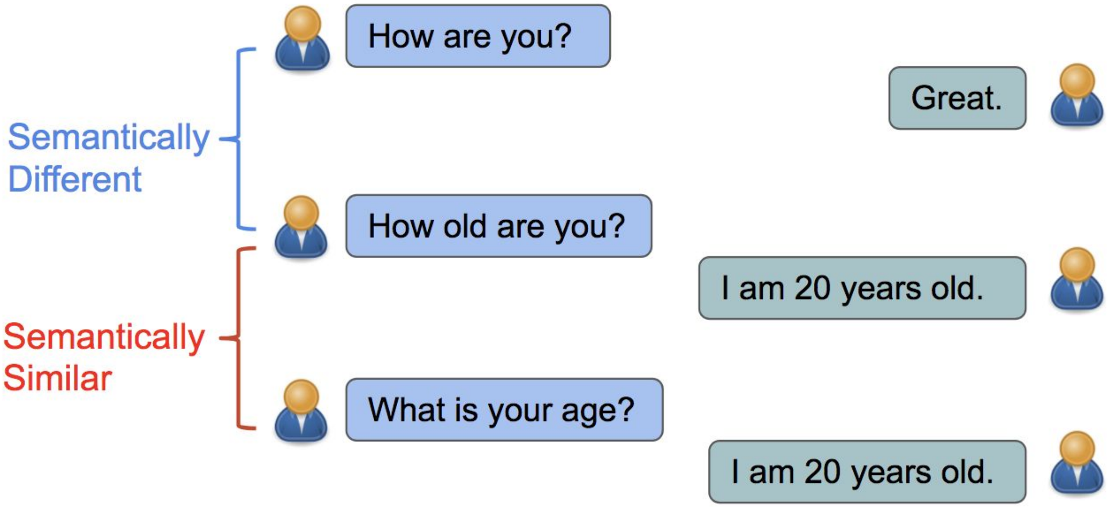
Classes: 0,1,2,3,4| 0: very negative, 4: very positive

Data: Movie Reviews

Demo Task 2 - Sentence Similarity

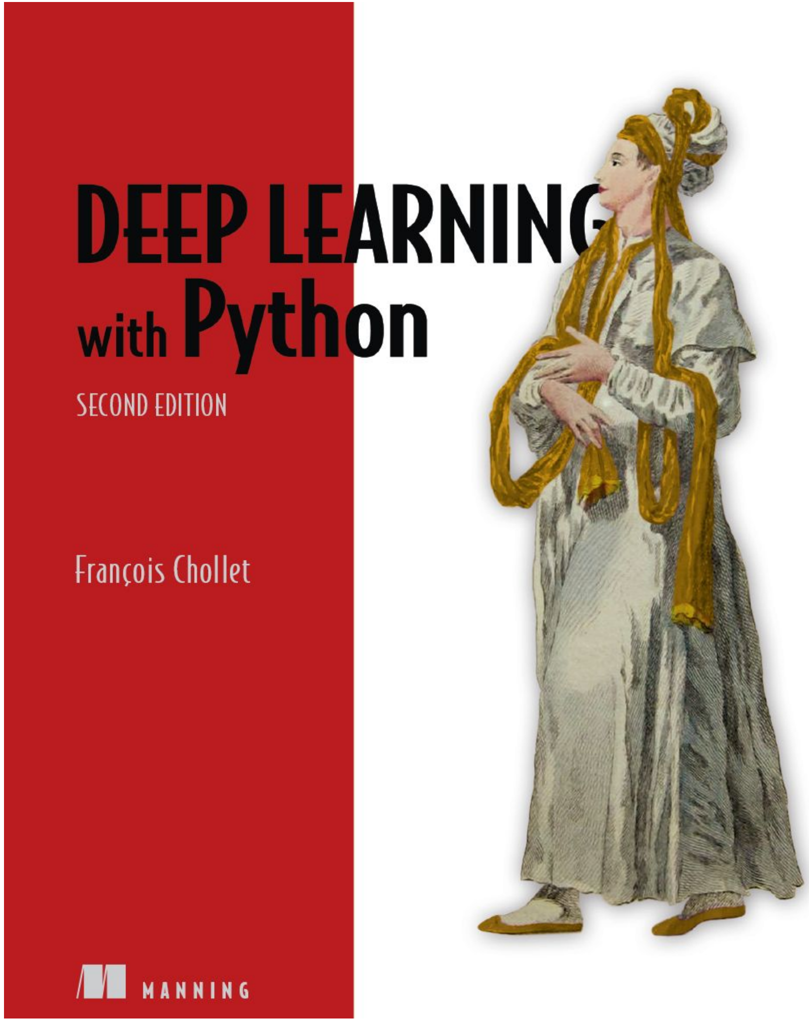
Task: Compute the (general, relative) similarity between sentences

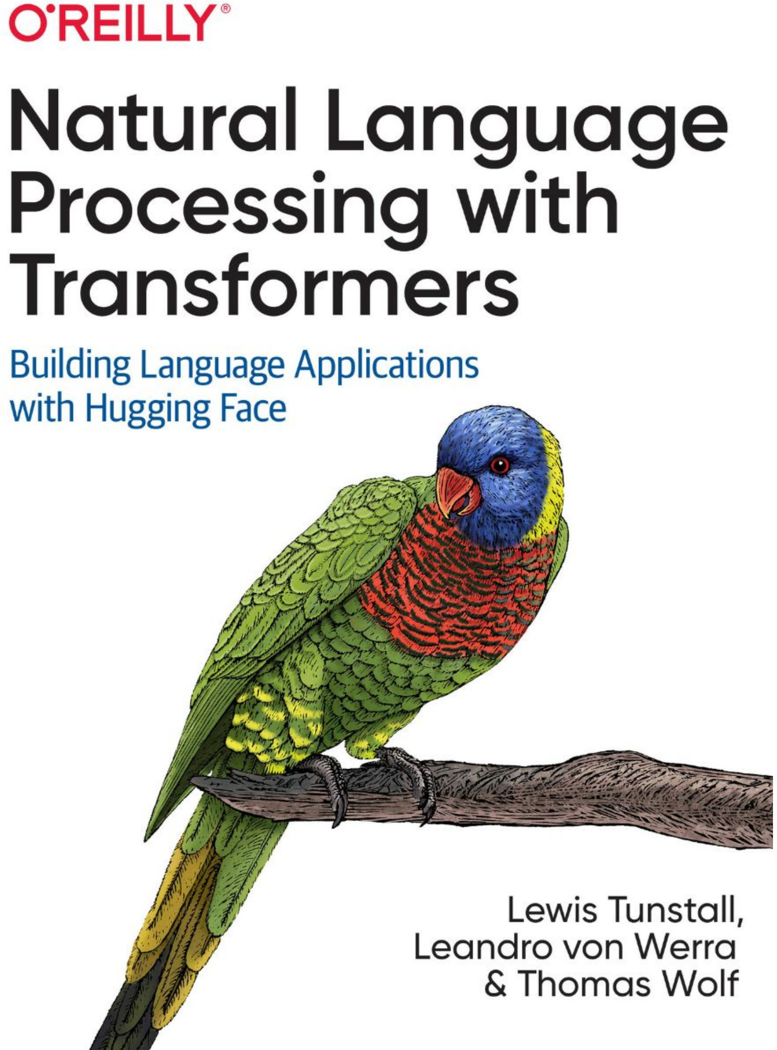
Data: Human ratings of semantic similarity



Demo 2

https://colab.research.google.com/drive/1AvVUMtp7yq9iloE5pINTmwOQ8AYiIJt-?usp=sharing


DEEP LEARNING
with Python
SECOND EDITION
François Chollet
MANNING


O'REILLY
Natural Language Processing with Transformers
Building Language Applications with Hugging Face
Lewis Tunstall,
Leandro von Werra
& Thomas Wolf

References

- P. Bojanowski, E. Grave, A. Joulin, T. Mikolov: Enriching Word Vectors with Subword Information. TACL 2016. https://arxiv.org/abs/1607.04606
- Anton Ehrmanntraut, Thora Hagen, Leonard Konle, Fotis Jannidis: Type- and Token-based Word Embeddings in the Digital Humanities. In: Maud Ehrmann et al. (eds.): CHR 2021. Proceedings of the Conference on Computational Humanities Research 2021. Amsterdam, the Netherlands, November 17-19, 2021. p. 16-38.
- Grave et al.: Learning Word Vectors for 157 Languages. 2018. https://arxiv.org/pdf/1802.06893.pdf
- William L. Hamilton, Jure Leskovec, Dan Jurafsky: HistWords: Word Embeddings for Historical Text. ACL 2016. https://arxiv.org/pdf/1605.09096.pdf
- Quoc Le, Tomas Mikolov: Distributed Representations of Sentences and Documents. *Proceedings of the 31 st International Conference on Machine Learning, Beijing, China, 2014. JMLR*: W&CP volume 32.