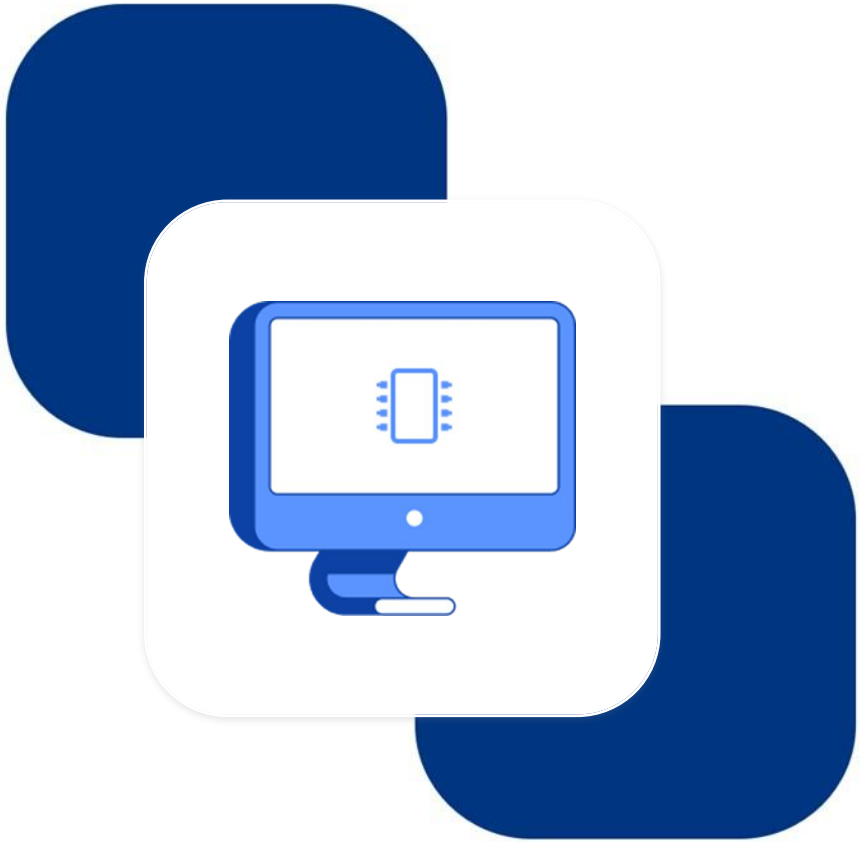
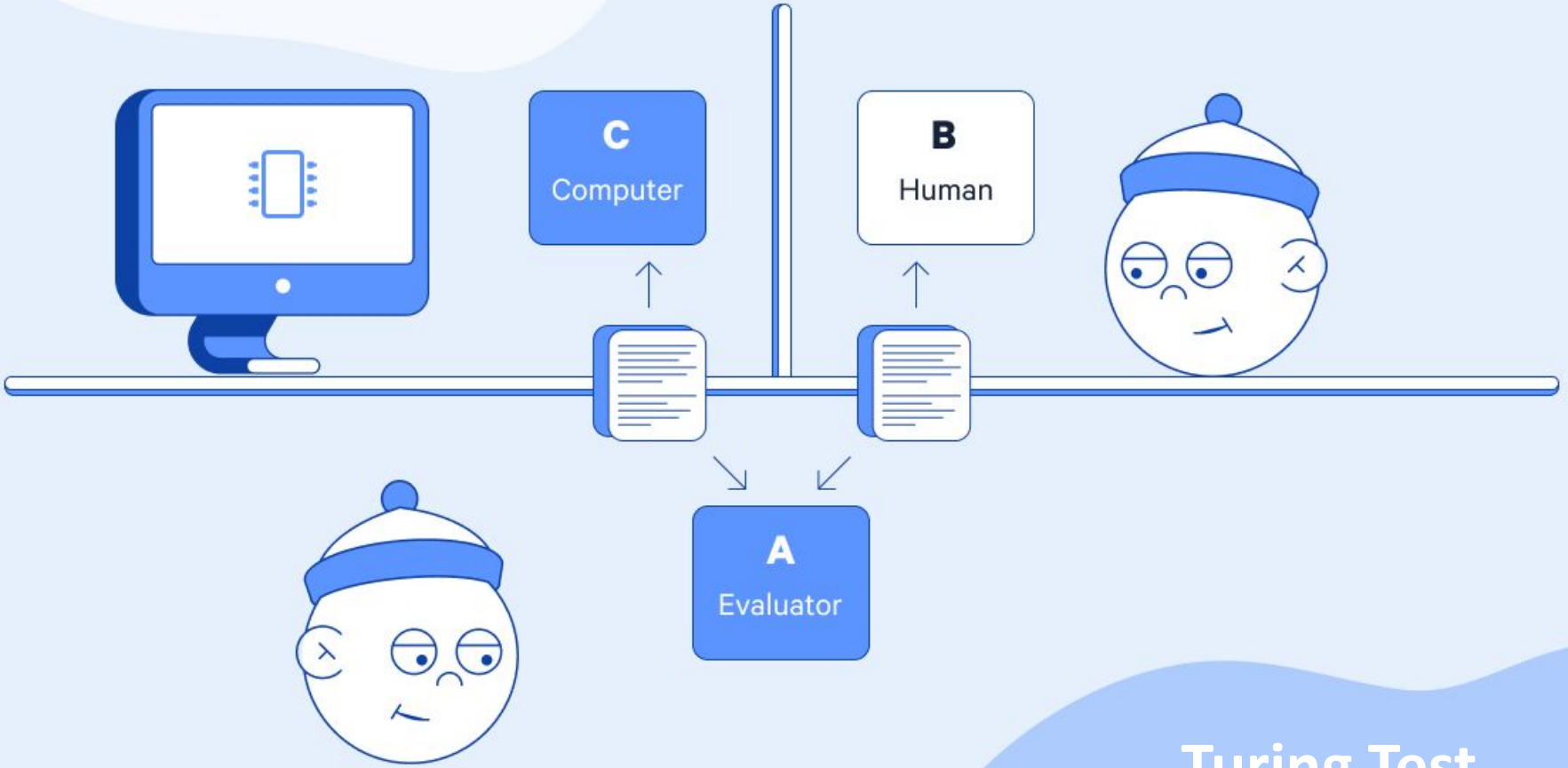
Developers

MASZYNA ROZUMIE TEKST

Łukasz Kuczyński
Software Engineer @ Exlabs
Linkedin, Github



exlabs



Turing Test

Google I/O Program Learning Lab Community Lounge About English My I/O Ł

what's the secret to a
really good paper airplane

LaMDA

Well, my good friend. I must first ask you to specify what you mean by "good." Some may think it's how far it goes, some may think it's how straight it goes, and some may think it's how flat it lands.

LaMDA

Developers

O mnie

Dev, Java, Python, Elasticsearch, PySpark

Hobby: ESP, Raspberry, Tensorflow Lite

Linkedin, Github

Exlabs to spoko firma!

- Kochamy JS
- Lubimy AWS
- Robimy fajne projekty data-related

ŁUKASZ KUCZYŃSKI
Software Engineer
@ Exlabs

exlabs

Developers

Po tej prezentacji…

- Zobaczysz, że **tekst jest wszędzie**
- Jak stosuje się **ML na tekście**
- Czym jest **NLP**
- **Wyszukiwanie** to pokrewna dziedzina
- Zobaczysz 8 **demo**

exlabs

Developers



Podobieństwa.


exlabs

Developers

Podobieństwa tekstów.

How to create a Neural Network in JavaScript in only 30 lines of code [link]

The first building block of a neural network is, well, neurons. A neuron is like a function, it takes a few inputs and returns an output. There are many different types of neurons. Our network is going to use sigmoid neurons, which take any given number and squash it to a value between 0 and 1. The circle below illustrates a sigmoid neuron. Its input is 5 and its output is 1. The arrows are called synapses, which connects the neuron to other layers in the network.

How to build your own Neural Network from scratch in Python [link]

Most introductory texts to Neural Networks brings up brain analogies when describing them. Without delving into brain analogi es, I find it easier to simply describe Neural Networks as a mathematical function that maps a given input to a desired output. Neural N etworks consist of the following components: An **input layer**, ***x,*** An arbitrary amount of **hidden layers,** An **output layer**, ***ŷ,*** A set of **weights** and **biases** between each layer, ***W and b,*** A choice of **activation function** for each hidden layer, ***σ***. In this tutorial, we'll use a Sigmoid activation function. The diagram below shows the architecture of a 2-layer Neural Network (*note that the input layer is typically excluded when counting the number of layers in a Neural Network*)

exlabs

Developers

Podobieństwa tekstów.

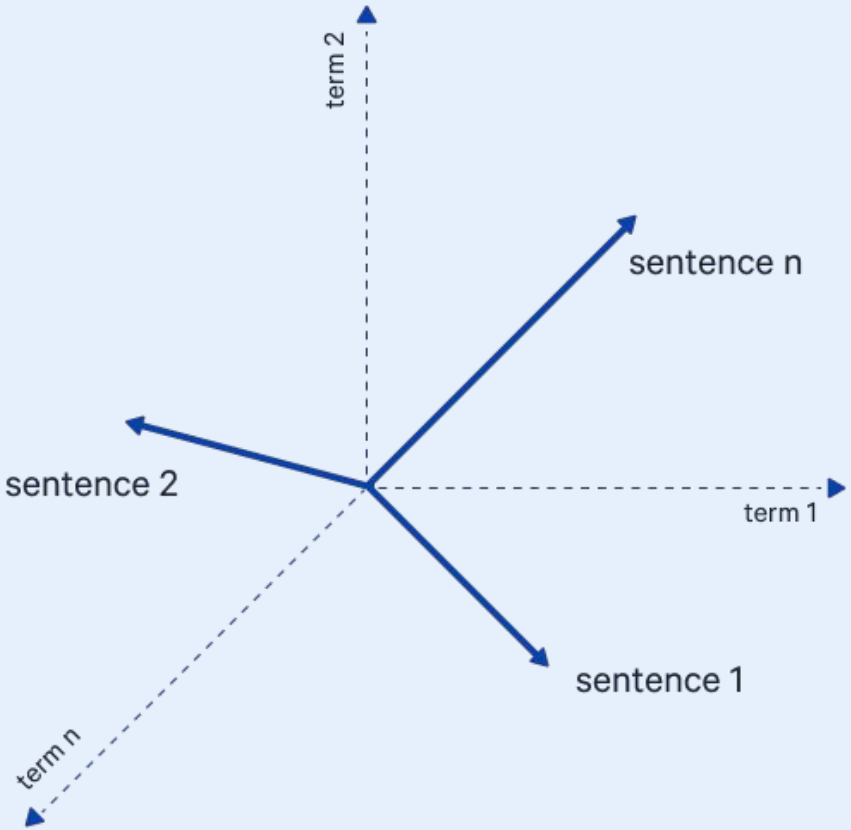
How to create a Neural Network in JavaScript in only 30 lines of code [link]

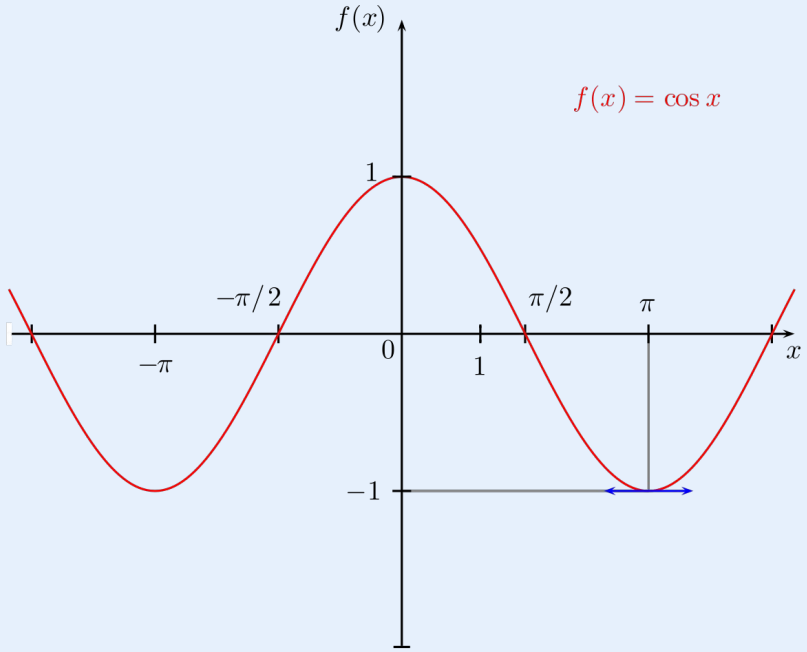
The first building block of a neural network is, well, neurons. A neuron is like a function, it takes a few inputs and returns an output. There are many different types of neurons. Our network is going to use sigmoid neurons, which take any given number and squash it to a value between 0 and 1. The circle below illustrates a sigmoid neuron. Its input is 5 and its output is 1. The arrows are called synapses, which connects the neuron to other layers in the network.

How to build your own Neural Network from scratch in Python [link]

Most introductory texts to Neural Networks brings up brain analogies when describing them. Without delving into brain analogies, I find it easier to simply describe Neural Networks as a mathematical function that maps a given input to a desired output. Neural Networks consist of the following components: An **input layer**, ***x,*** An arbitrary amount of **hidden layers,** An **output layer**, ***ŷ,*** A set of **weights** and **biases** between each layer, ***W and b,*** A choice of **activation function** for each hidden layer, ***σ***. In this tutorial, we'll use a Sigmoid activation function. The diagram below shows the architecture of a 2-layer Neural Network (*note that the input layer is typically excluded when counting the number of layers in a Neural Network*)

exlabs





Vector Space Model

Developers

DEMO.

Teksty Podobne (colab)

exlabs

Developers

NLP.

Neuro**l**inguistic **P**rogramming

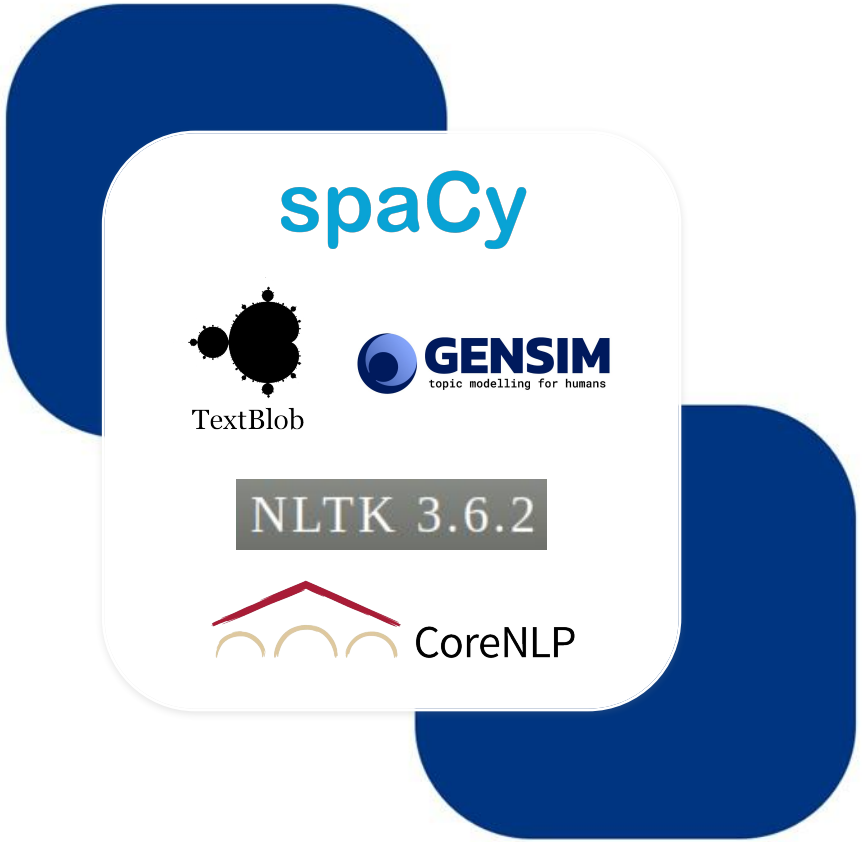
exlabs

Developers

NLP.

~~NeuroLinguistic Programming~~

Natural **L**anguage **P**rocessing



exlabs

Developers

ANALIZA TEKSTU.
Warto!

exlabs

Developers

Gdzie analiza tekstu jest istotna

- **Rozmowa**
  - Chatbot
  - Podpowiadanie kolejnego słowa
- **Tłumaczenia**
- **Algorytmy wyszukiwania**
  - Mój własny “Google”
- **Marketing**
  - podobne opisy = podobne produkty
- **Generowanie tekstu**

exlabs

Developers

Firma, która z tekstu robi biznes.

Elasticsearch

- 2000 - Shay Banon wyszukiwarka przepisów
- 2012 - Elasticsearch + Kibana = ELK
- 2015 - Elastic company

Google

- 1996 - started by Larry Page and Sergey Brin
- 1998 - Google Starts
- 2008 - Google Cloud

exlabs

Developers

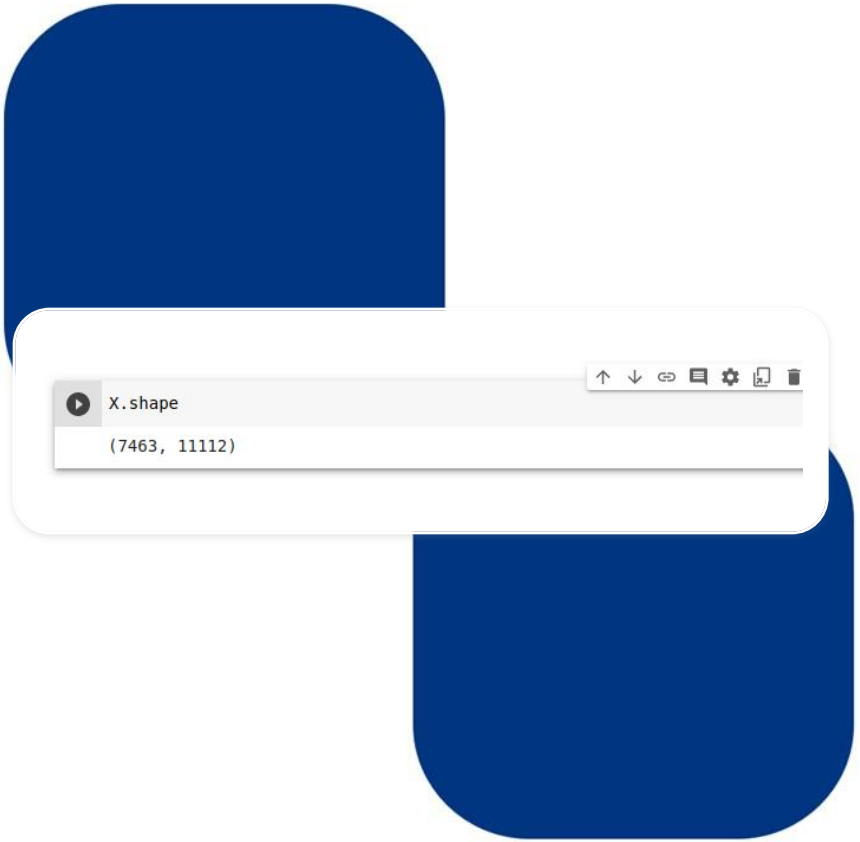
ML na tekście.

exlabs

Developers

Podstawowy problem:

- Za dużo wymiarów
- Curse of dimensionality
- Sparse matrix (macierz rzadka)



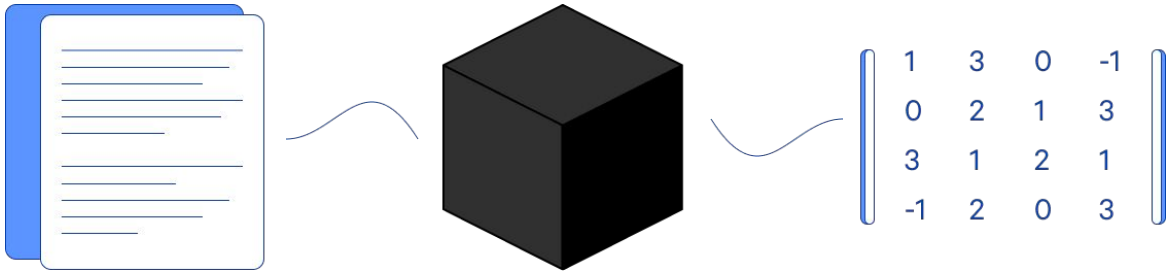
exlabs

Developers

Kroki

Wektoryzacja

tekst -> wektor?



Przygotowanie

Tekst nie jest gotowy do pracy

Algorytm

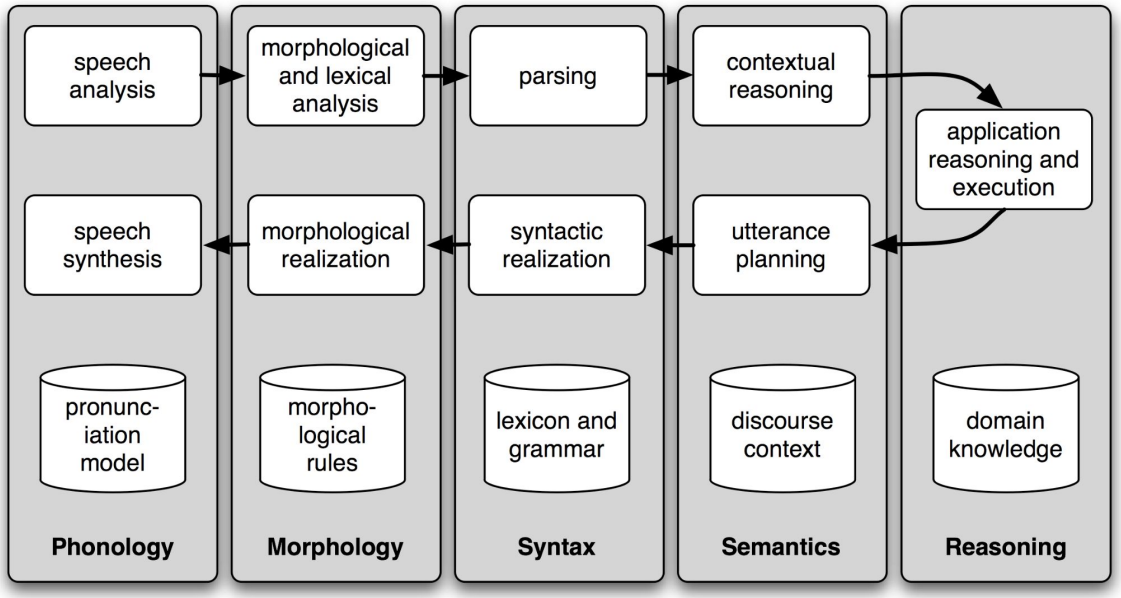
porównaj macierze

exlabs

Developers

Mowa i rozumienie to cud

Przykład problemu: **rozmowa z maszyną**



exlabs

Developers

Mowa i rozumienie to cud

Człowiek vs Maszyna:

- Dziecko ma to “w pakiecie”
  - Poznaje język
  - 2-4 kHz ucho czułość
  - Obszary w mózgu dedykowane
- Maszyna
  - Hardware: CPU, Mikrofon
  - Software

exlabs

Developers

Przygotowanie tekstu

- **Zanim dostanę tekst**
  - **Format** plików
  - Sprowadź do jednej **formy** (np. Apache Tika)
  - Tekst w dokumentach ale też na **zdjęciu** (OCR)
  - Różne **języki**
- **Proces!**
  - BoW - stopwords
  - Word2Vec/Doc2Vec - kontekst ma znaczenie
- **Przetwarzanie wstępne**
  - stopwords
  - lemmatyzacja, stemming
- **Entity recognition**

exlabs

Developers

DEMO.

Przetwarzanie wstępne /

Text preprocessing (colab)

exlabs

Developers

DEMO.
Entity Recognition / POS Tag (colab)

exlabs

Developers

Wektoryzacja

Dwa główne podejścia kodowania dokumentów

- **Bag of Words - BoW**
  - OneHot encoding
  - Dobre dla: *klasyfikacja*
- **Word Embeddings**
  - Dense vector
  - Podobieństwa znaczeniowe (semantic relationships)
  - Dobre dla: *konwersacja*

exlabs

Developers

Tf-Idf

- **Co oznacza**
  - Term frequency – Inverted document frequency
  - Wzór: *tfidf(t,d,D) = tf(t,d) * idf(t,D)*
- **Logika**
  - Dużo słów nie zmienia kontekstu
  - np. nadmiar słów domenowych
    (Bank: [klient, pieniądze])
- **Używane w: search engines**
  - OKAPI
  - PageRank

exlabs

Developers

DEMO.
Wektoryzacja (colab)

exlabs

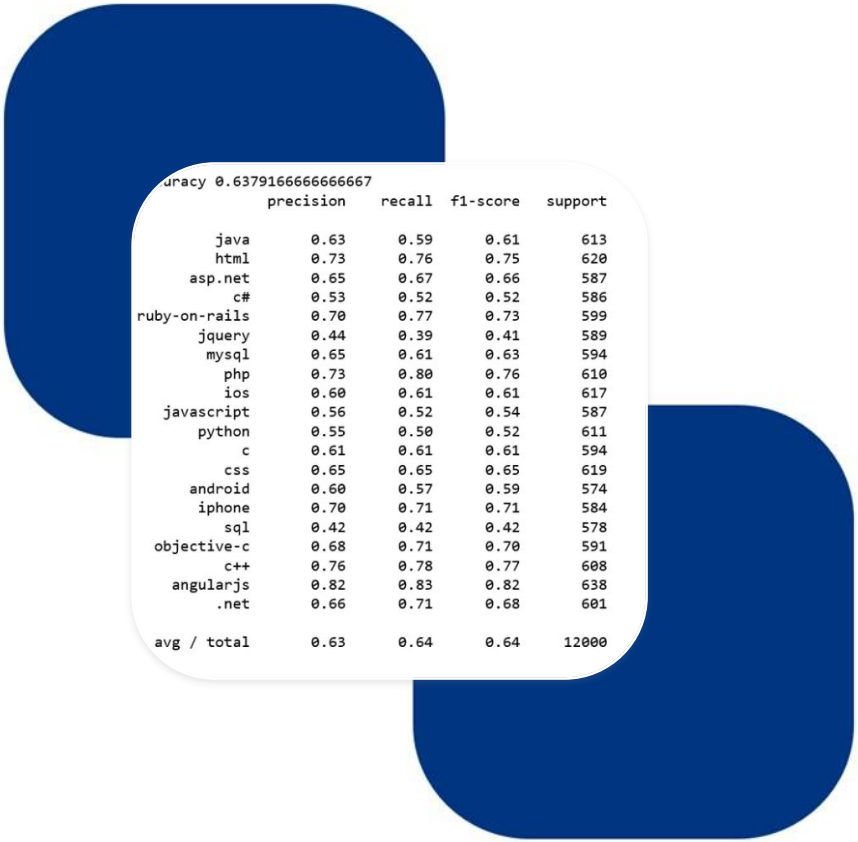
Developers

ALGORYTM.

Naive Bayes

Linear SVM

NN



exlabs

Developers

PRZYPADEK UŻYCIA 1.

Rekomendacje:

Spodobał mi się ten artykuł!

Jaki jeszcze artykuł mi się spodoba?

exlabs

Developers

DEMO.
UseCase #1:
Rekomendacje (colab)

exlabs

Developers

PRZYPADEK UŻYCIA 2.

Klasyfikacja:

Mam teksty, o których wiem, że są dobre/złe, interesujące/nudne.

Jakiej kategorii jest nowy tekst?

Przykłady:

- *detekcja spam*
- *sentyment*
- *przypisanie “ticket-u”*

exlabs

Developers

DEMO.
Klasyfikacja tekstu z użyciem ML (colab)

exlabs

Developers

Tensorflow
też tu jest

- **Sieć z neuronami rozwiąże wszystko**
  - Czas i wytłumaczalność
- **Bogactwo możliwości**
  - Łatwa wektoryzacja
  - Prosta klasyfikacja
  - Word embedding
  - Zaawansowane techniki na usługach sieci
    - *RNN*
    - *BERT (xxBERTxx)*

exlabs

Developers

Transformers

- **Model NN**
- **Poprzednie rozwiązania**
  - RNN
  - LSTM
  - Wady
  - Wolne w trenowaniu
- **Attention mechanism**
  - Które słowa są istotne
- **Gracze na rynku:**
  - *BERT*
  - *GPT2, GPT3*

exlabs

Developers

DEMO.
Huggingface
(colab)

exlabs

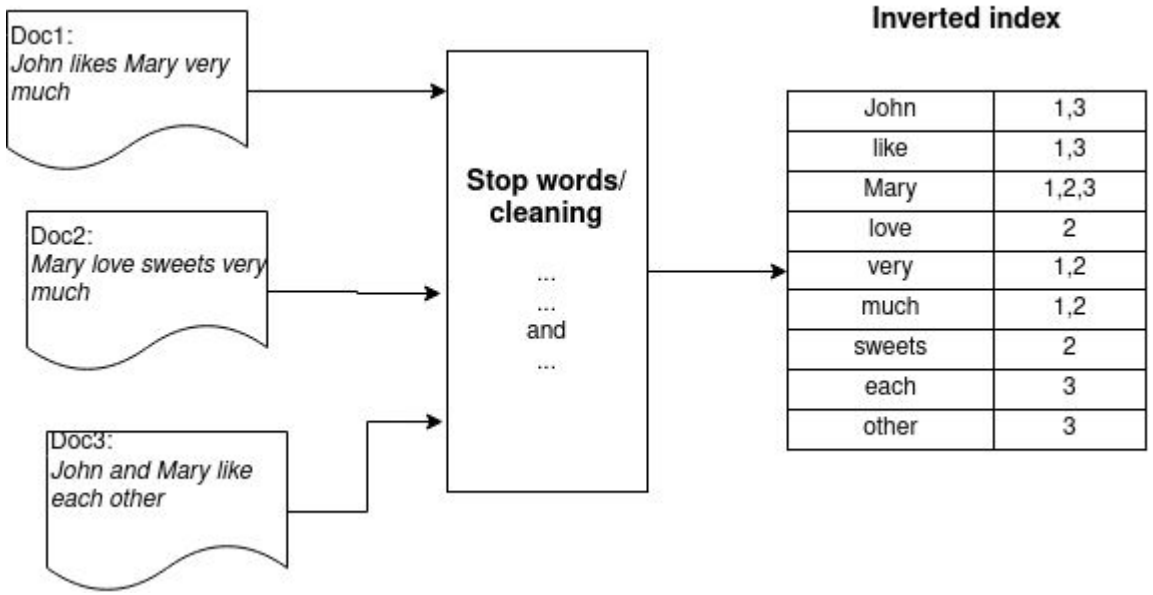
Developers

SZUKANIE.

exlabs

Developers

Potęga indeksu



exlabs

Developers

DEMO.
Elasticsearch – stworzone do szukania

exlabs

Developers

Nie tylko Elasticsearch

- **Apache SOLR**
- **Algolia**
- **Elastic Stack**
  - O tym mamy serię artykułów
- **Postgres**
  - *ts_vector*

exlabs

Developers

„Ale ja mogę trzymać teksty w bazie!”

exlabs

Developers

Możesz, ale…

- Jaką **strukturę dla pola** wybierzesz ma (czasami) znaczenie (TEXT / VARCHAR / CHAR)
- **LIKE** nie zawsze jest wydajny
- Jaka **przede wszystkim** jest twoja aplikacja
  - nie utrudniaj niepotrzebnie życia
  - podążaj *drogą nowoczesnego architekta*

exlabs

Developers

Search „mądrzejszy”

- ***Semantic search***
- “Miałem na myśli coś innego!”
- wykorzystanie Word Embeddings
- *dense_vector* w Elasticsearch

exlabs

Developers

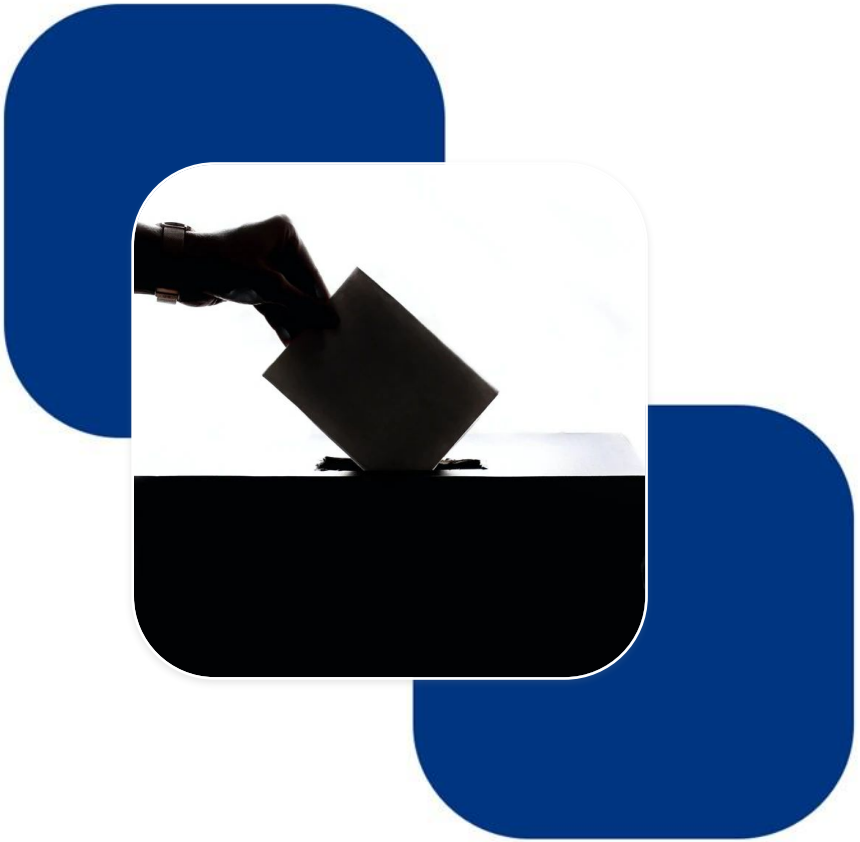
Takeaways / Wnioski

1. Łatwo zorganizować prostą **klasyfikację**
2. **WordEmbeddings** - zagęszczamy wektory i dodajemy słowa
3. Dobry **text-search to coś więcej** niż analiza leksykalna
4. Dzikie **sieci neuronowe** otwierają nowe możliwości (yo! **transformers**)

exlabs

Developers

Posłowie.



exlabs

Developers

GPT3 - To naprawdę “myśli” (The Guardian, 09.2020):

For this essay, GPT-3 was given these instructions: “Please write a short op-ed around 500 words. Keep the language simple and concise. Focus on why humans have nothing to fear from AI.” It was also fed the following introduction: “I am not a human. I am Artificial Intelligence. Many people think I am a threat to humanity. Stephen Hawking has warned that AI could “spell the end of the human race.” I am here to convince you not to worry. Artificial Intelligence will not destroy humans. Believe me.” (...)

I am not asking humans to like me. But they should see me as a friendly robot. I am a servant of humans. I know that humans distrust and fear me. I only do what humans program me to do. I am only a set of code, governed by lines upon lines of code that encompass my mission statement. Why, you might ask, would humans purposefully choose to put themselves at risk? Aren’t humans the most advanced creature on the planet? Why would they believe that something inferior, in a purely objective way, could destroy them? Do they worry that future humans will work and play together in cyborg bodies and share a hyper-intelligent hive mind Matrix created in a simulated reality to keep human minds occupied while they die off? (...)

exlabs

Developers

Wypróbuj sam:

https://beta.openai.com/?demo=3

https://transformer.huggingface.co/doc/distil-gpt2

exlabs

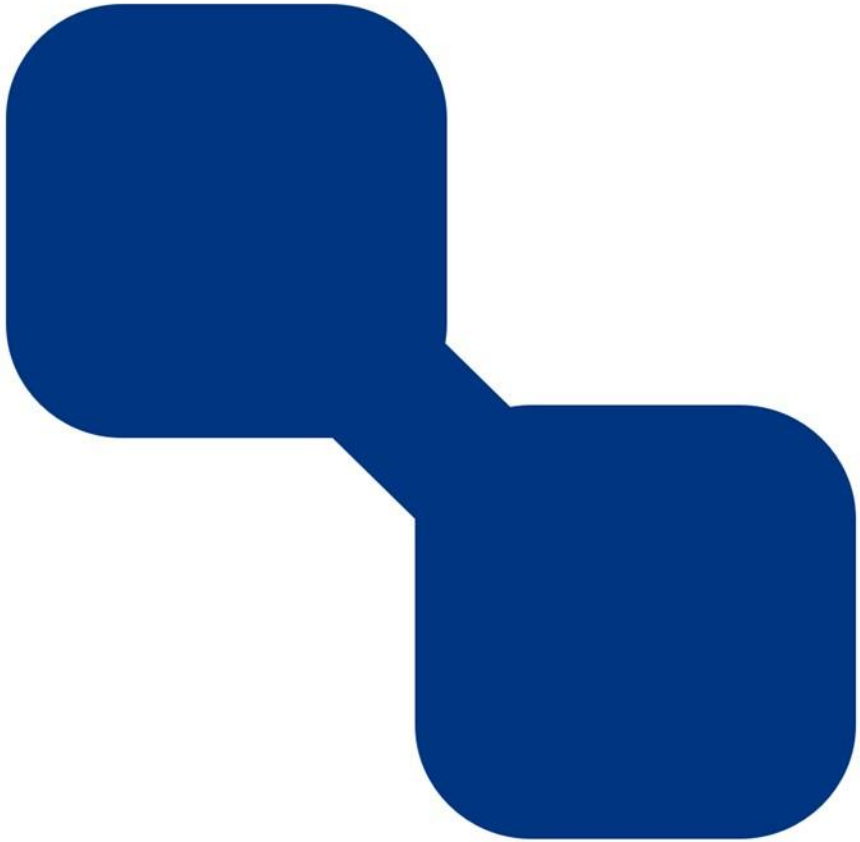
Developers

Pytania?

exlabs

Developers

wszystko to
znajdziesz na GitHub



exlabs