SKAZANI NA PASTEBIN

NLP KONTRA ŚMIETNIK INTERNETU

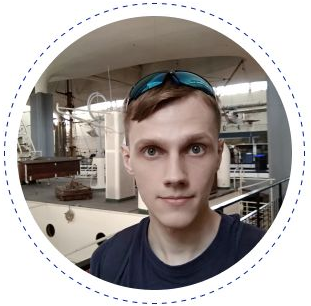
MARTA ZAJKOWSKA & MICHAŁ JADCZUK


HAIK
YE H

Zespół

Marta Zajkowska

UX/UI Designer



Maciej Sawicki

Full Stack Developer

Michał Jadczuk

Data Scientist

HACK YEAH

EXATEL

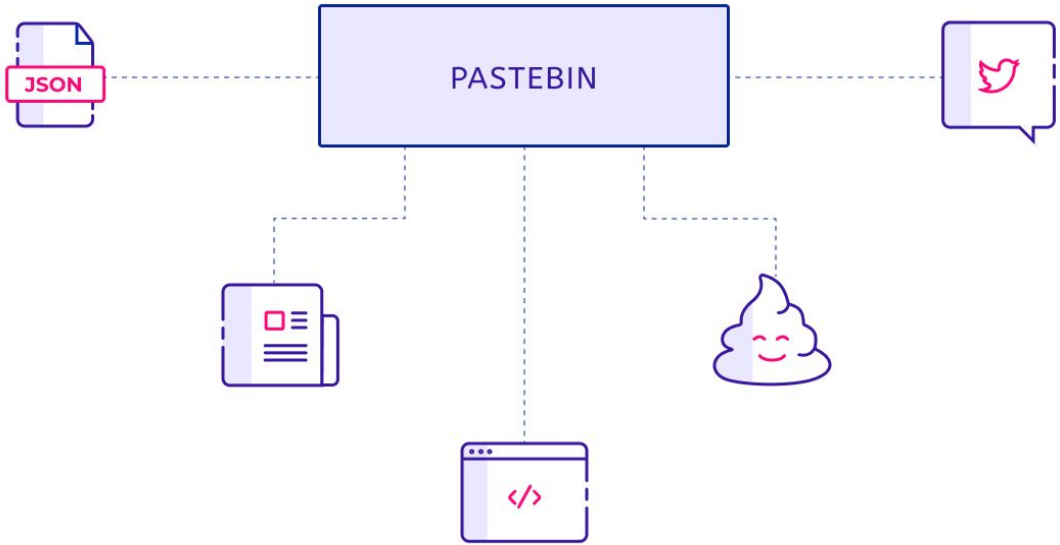
Amount of available data increases every day. Finding useful information – *information we can work with* – is possible, but quite tedious and difficult without spending lots of resources.

Can you write the software which, given a lot of uncategorized data (say – *tweets, blog posts, application logs, pcaps*), groups them by **the similarity of the discussed topic**?

The goal is that certain groups can be ignored as uninteresting and other browsed manually. Could you **score and sort** the information in a useful way?

If you've considered trying your skills with **unsupervised machine learning**, this challenge is for you!

Problem



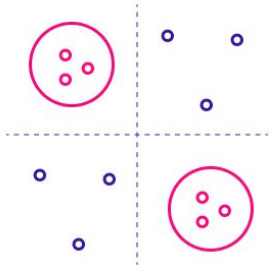
Przypadki użycia

Użytkownik nie zna zbioru danych i nie wie, czego szuka

Użytkownik nie zna zbioru danych, ale wie, czego szuka

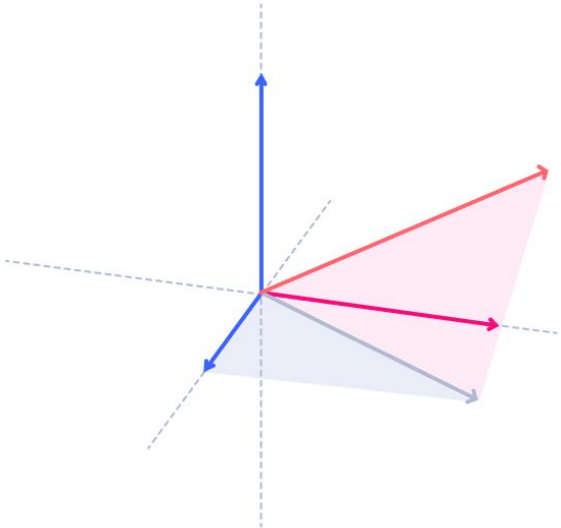
Funkcjonalności

Wyszukiwanie kontekstowe



Grupowanie danych

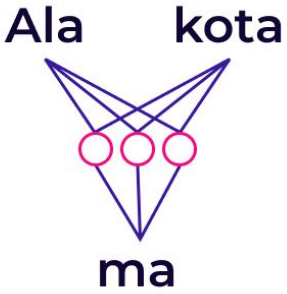
Wektoryzacja



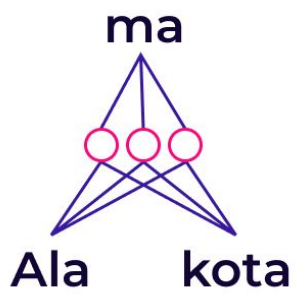
Przykład

FRAZA	ALA	MIEĆ	KOT	SZCZEŚCIE
Ala ma kota	1	1	1	0
Kot ma Alę	1	1	1	0
Mamy kota, mamy szczęście	0	2	1	1

word2vec

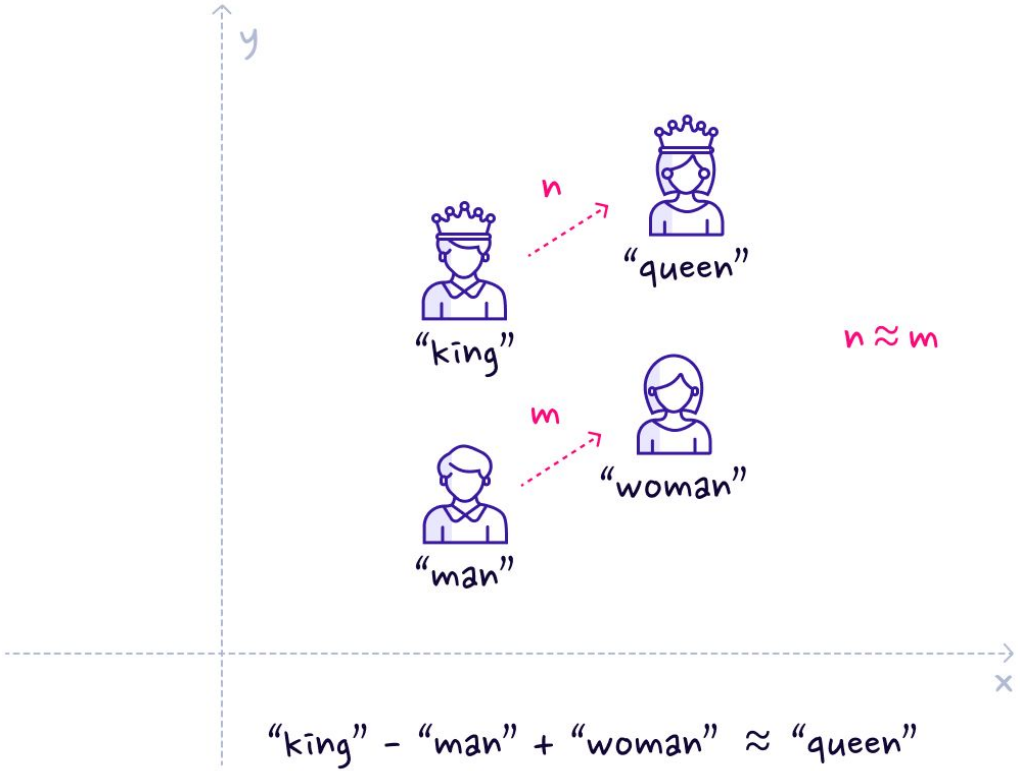


Continuous bag-of-words

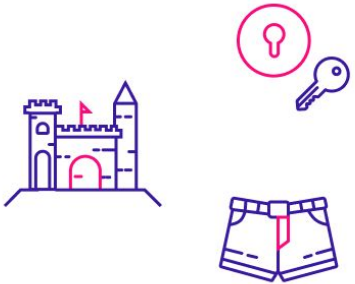


Skip-gram

word2vec



word2vec: wady

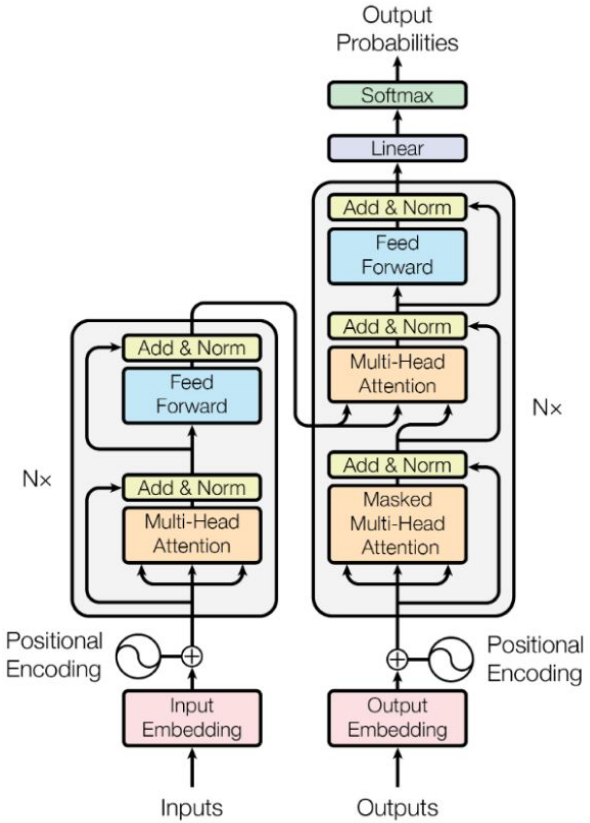


Wektory niezależne
od kontekstu

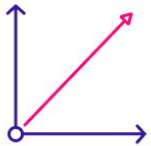
Wektory
na poziomie słów

ATTENTION IS ALL YOU NEED!

Architektura



Transformer: zalety



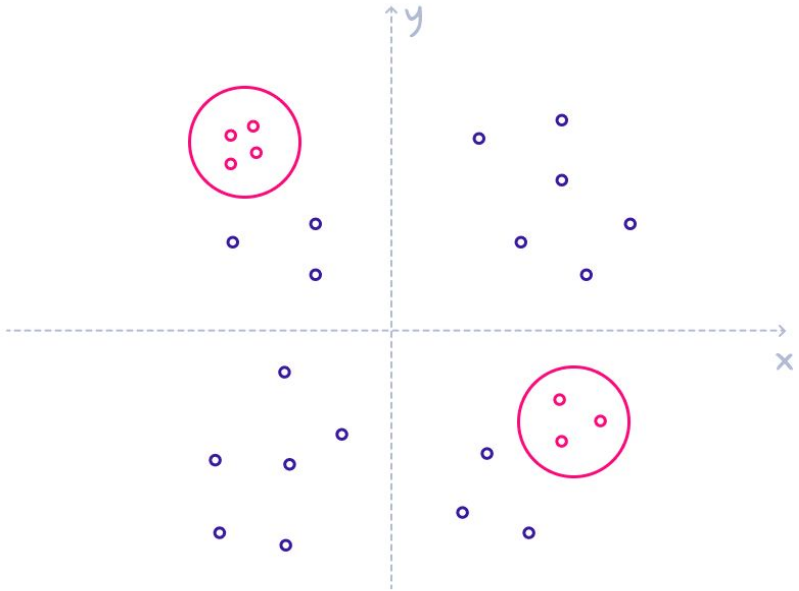
Skalowalność

Analiza w jednym kroku

Warstwy atencji

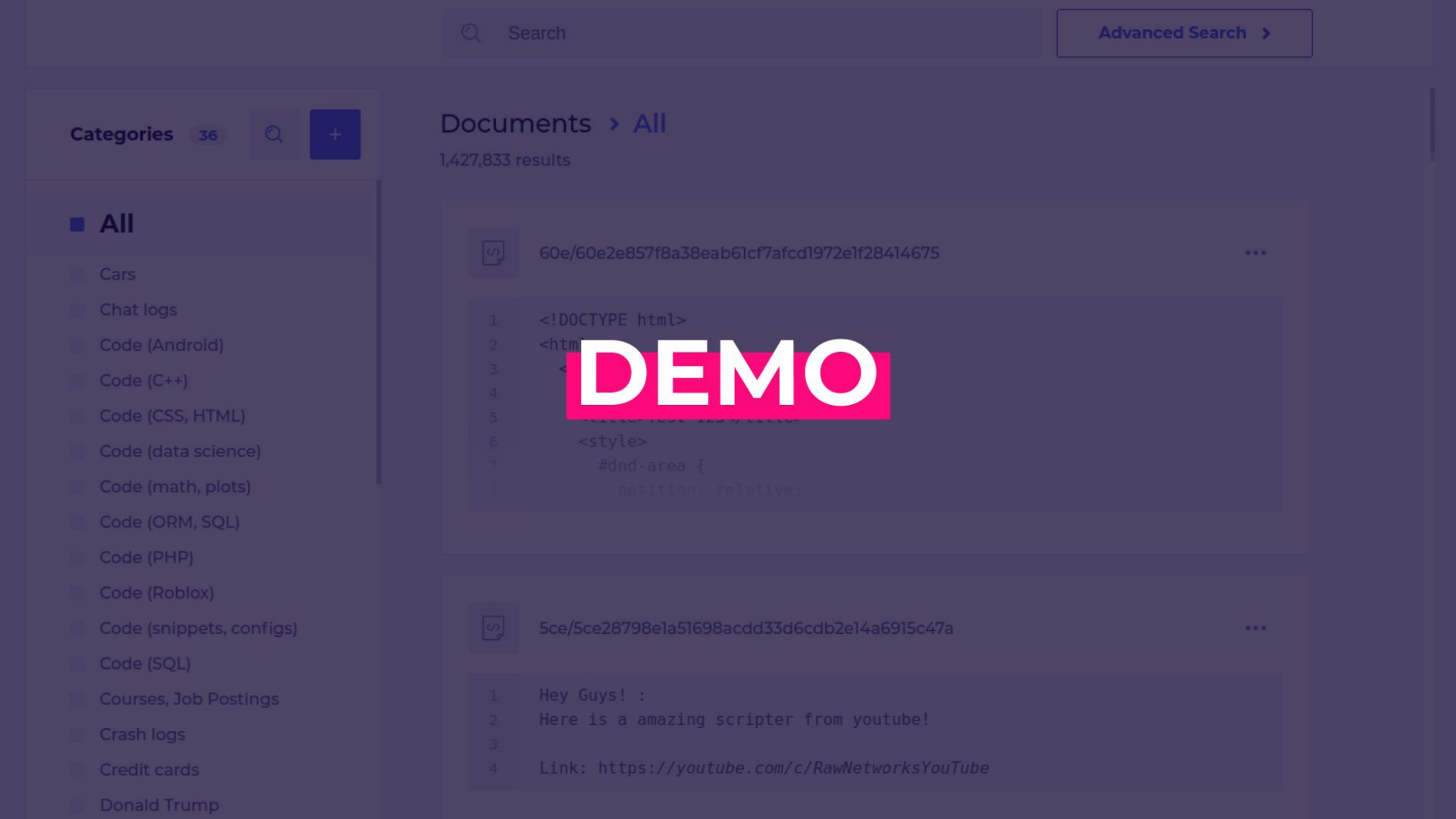

michael
had
a
beautiful
dog
.
he
walked
his
dog
every
day
.
michael
had
a
beautiful
dog
.
he
walked
his
dog
every
day
.

Grupowanie



Grupowanie


50 000
40 000
30 000
20 000
10 000
0
ERROR
25
50
75
100
125
CLUSTERS AMOUNT


Search
Advanced Search
Categories 36
All
Cars
Chat logs
Code (Android)
Code (C++)
Code (CSS, HTML)
Code (data science)
Code (math, plots)
Code (ORM, SQL)
Code (PHP)
Code (Roblox)
Code (snippets, configs)
Code (SQL)
Courses, Job Postings
Crash logs
Credit cards
Donald Trump
Documents › All
1,427,833 results
60e/60e2e857f8a38eab61cf7afcd1972e1f28414675
<!DOCTYPE html>
<style>
#dnd-area {
position: relative;
5ce/5ce28798e1a51698acdd33d6cdb2e14a6915c47a
Hey Guys! :
Here is a amazing scripter from youtube!
Link: https://youtube.com/c/RawNetworksYouTube
DEMO

Jak żyć na hackathonie?

Interdyscyplinarny zespół

Gotowy produkt

Dobra prezentacja

Współpraca z mentorami

#hackathons

DZIĘKUJEMY ZA UWAGĘ