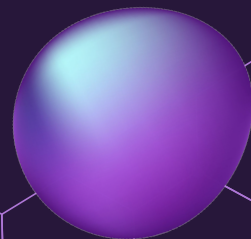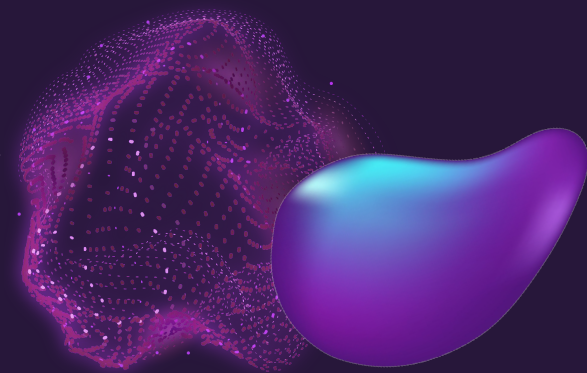# QUORA QUESTION PAIRS

Narzędzie do porównywania tekstów
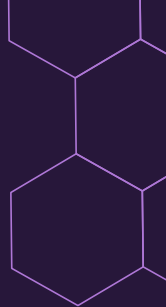
Lenart Piotr
Mizera Damian

Cel Projektu

- Stworzyć program do porównywania tekstów pod względem ich znaczenia
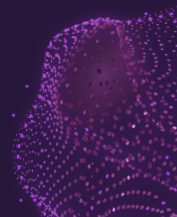
Baza danych wejściowych:

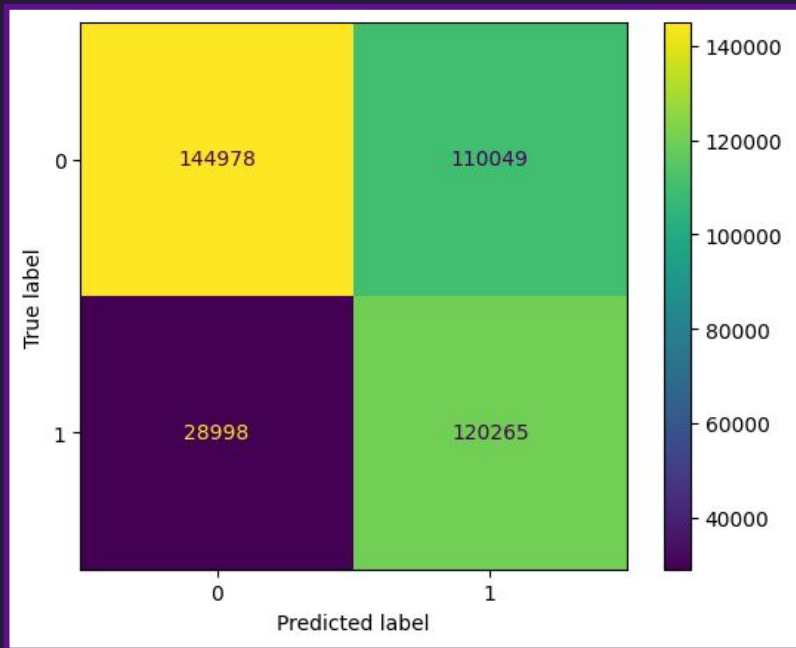|	id	qid1	qid2	question1	question2	is_duplicate
0	0	1	2	What is the step by step guide to invest in sh...	What is the step by step guide to invest in sh...	0
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0
2	2	5	6	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	0
3	3	7	8	Why am I mentally very lonely? How can I solve...	Find the remainder when [math]23^{24}[/math] i...	0
4	4	9	10	Which one dissolve in water quikly sugar, salt...	Which fish would survive in salt water?	0
...	...	...	...	...	...	...
404285	404285	433578	379845	How many keywords are there in the Racket prog...	How many keywords are there in PERL Programmin...	0
404286	404286	18840	155606	Do you believe there is life after death?	Is it true that there is life after death?	1
404287	404287	537928	537929	What is one coin?	What's this coin?	0
404288	404288	537930	537931	What is the approx annual cost of living while...	I am having little hairfall problem but I want...	0
404289	404289	537932	537933	What is like to have sex with cousin?	What is it like to have sex with your cousin?	0

https://www.kaggle.com/c/quora-question-pairs/overview

Przygotowanie danych

Podział słów w
pytaniach na tokeny

+ Lematyzacja - forma
podstawowa słów

Tokeny

Stemming

Korpus

Stop words

TFIDF

Przygotowanie korpusu
wszystkich pytań

Usuwanie zbędnych
słów

Przekształcenie wyrazów
na wektory liczb

"What is one coin?" vs "What's this coin?"

Podobieństwo cosinusowe

Dokładność: 66%

Testowanie modeli

MultinomialNB

SGDClassifier

LinearSVC

LGBMClassifier

Tabela dokładności

| Model name	TfiDf - Lemmatization	TfiDf - Stemming	TfiDf - L + S
MultinomialNB	74,13%	73,76%	73,75%
SGDClassifier	74,32%	74,19%	74,26%
LinearSVC	75,54%	75,38%	75,43%
LGBMClassifier	75,31%	75,38%	75,44%

Własny model AI

Struktura modelu

```python
model = Sequential()
model.add(Dense(units=1024, input_dim=10000, activation='relu'))
model.add(Dense(units=1024, activation='relu'))
model.add(Dense(units=1024, activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(units=1024, activation='relu'))
model.add(Dense(units=1024, activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(units=1024, activation='relu'))
model.add(Dense(units=1, activation='sigmoid'))
```

Dokładność: 80%

Tabela prawdy i fałszu
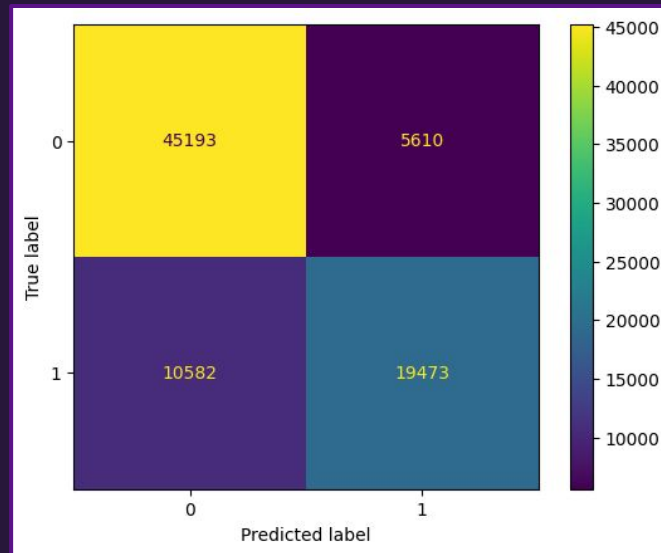
Model wielokanałowy

Wejście 1   Wejście 2

✖

Dokładność: 81%

+ word2vec

Dokładność: 75%

✖

Warstwy ukryte   Warstwy ukryte

Warstwa łączenia

Warstwy ukryte

Wyjście

```python
input_1 = Input(shape=(5000,))
input_2 = Input(shape=(5000,))

branch_1 = Dense(512, activation='relu')(input_1)
branch_1 = Dense(256, activation='relu')(branch_1)
branch_1 = Dense(128, activation='relu')(branch_1)

branch_2 = Dense(512, activation='relu')(input_2)
branch_2 = Dense(256, activation='relu')(branch_2)
branch_2 = Dense(128, activation='relu')(branch_2)

merged = Concatenate()([branch_1, branch_2])

merged = Dense(256, activation='relu')(merged)
merged = BatchNormalization()(merged)
merged = Dropout(0.5)(merged)
merged = Dense(128, activation='relu')(merged)
merged = BatchNormalization()(merged)
merged = Dropout(0.5)(merged)
merged = Dense(128, activation='relu')(merged)
merged = BatchNormalization()(merged)
merged = Dropout(0.5)(merged)
merged = Dense(64, activation='relu')(merged)

output = Dense(1, activation='sigmoid')(merged)

model = Model(inputs=[input_1, input_2], outputs=output)
```

Rozmowa z modelem

```python
while(True):
    q1 = input()
    q2 = input()
    if len(q1) == 0 or len(q2) == 0:
        break
    print(q1)
    print(q2)
    inp = hstack([vectorizer.transform([q1]), vectorizer.transform([q2])])
    ans = model.predict(vstack([inp]))[0]
    print('> Yes' if ans else '> No')
```

Pytanie 1	Pytanie 2	Czy podobne
how are you feel?	How are you?	Yes
What's this coin?	What is color of this coin?	No
Are you good with math?	Are you a mathematician?	No
Are you a mathematician?	Are you a it programmer?	No
Have you a computer mouse?	Have you a pet mouse at house?	No
Have you computer mouse?	Do you have a mouse to your computer?	Yes

Dziękujemy

DO YOU HAVE ANY QUESTIONS?
piotr.lenart@student.pk.edu.pl