WTUM: Przewidywanie "śmieszności" żartu

Maciej Łukasik, Kajetan Larwa, Anton Yushkevich, Maksim Makaranka

April 2022

1 Pod-modele określające podobeństwo dwóch dowcipów

Skonstruowano następujące pod-modele mające na celu określenie podobieństwa między dwoma żartami:

- Jaccard similarity model
  W tym modelu do wyznaczenia podobeństwa żartów został wykorzystany indeks Jaccarda, który jest zdefiniowany jako iloraz mocy części wspólnej zbiorów i mocy sumy tych zbiorów.

```
Explained variance score: 0.2432117374221553
Mean absolute error: 3.3927190175797635
Mean squared error: 20.911941112971295
Median absolute error: 2.514470106167693
R2 coefficient: 0.2367212295390374
```

- Cosine Similarity model
  W celu zdefiniowania tej metody ciągi są postrzegane jako wektory w przestrzeni unitarnej, a podobieństwo jest definiowane jako cosinus kąta między nimi, czyli iloczyn skalarny wektorów podzielony przez iloczyn ich długości.

W metodzie Cosine Similarity konieczna była wcześniejsza wektoryzacja tekstu, do czego zostały użyte następujące metody:

- Term Frequency - Inverse Document Frequency
  TF-IDF to miara statystyczna używana do oceny ważności słowa w kontekście dokumentu. Waga słowa jest proporcjonalna do częstotliwości występowania tego słowa w dokumencie i odwrotnie proporcjonalna do częstotliwości występowania tego słowa we wszystkich dokumentach w kolekcji.

```
Explained variance score: 0.18371622140417865
Mean absolute error: 3.5097232397245515
Mean squared error: 22.409155959619465
Median absolute error: 2.5940000000000003
R2 coefficient: 0.17761894197234673
```

- SentenceBERT
  BERT to model językowy opracowany przez Google, który został wytrenowany za pomocą 3,3 mln angielskich słów. Jest to oparta na transformatorach technika uczenia maszynowego do przetwarzania języka naturalnego. Dużą zaletą tego modelu jest fakt, że BERT

„rozumie” kontekst, w którym zostało użyte dane słowo.

```
Explained variance score: 0.245060865531479
Mean absolute error: 3.401487091164569
Mean squared error: 20.783849113581297
Median absolute error: 2.5400696858800385
R2 coefficient: 0.2409516070506934
```

2 Modele wykorzystujące opracowane metryki

Zbudowano i przetestowane następujące modele:

- Popularity based filtering
  Model popularnościowy - model podstawowy oparty na popularności i ocenie dowcipu nadanej mu przez użytkownika.

```
Explained variance score: 0.18371622140417865
Mean absolute error: 3.5097232397245515
Mean squared error: 22.409155959619465
Median absolute error: 2.5940000000000003
R2 coefficient: 0.17761894197234673
```

- Collaborative filtering
  Jest to metoda, która wykorzystuje znane preferencje grupy użytkowników do przewidywania nieznanych ocen innego użytkownika. Główne założenie tej metody jest następujące: użytkownicy, którzy w przeszłości oceniali dane dowcipy w ten sam sposób, będą mieli tendencję do wystawiania podobnych ocen innym dowcipom w przyszłości.
  Wykorzystując tą metodę otrymaliśmy średni błąd równy 3.774.

- Autoencoder based
  Autoencoder to specjalny typ sieci neuronowej, która koduje dane, a następnie je dekoduje, przy czym dąży do jak największego podobieństwa dekodowanych danych do orginalnych. Z wykorzystaniem tej metody średni błąd wynosi 4.457.

- Category based
  Ta metoda polegała na pokategoryzowaniu żartów w grupy używając spectral clusteringu. Pozwoliło to uzyskać średni błąd równy 4.266.

- Content based
  Filtrowanie oparte na treści wykorzystuje wcześniejsze oceny danego użytkownika w celu stwierdzenia w jakim stopniu dany żart, którego jeszcze nie oceniał, może mu się spodobać. Używając tej metody udało się uzyskać średni błąd równy 3.509.

3 Wnioski

Jak możemy zobaczyć w statystykach dla poszczególnych modeli, mają one średni błąd około 3.5, co w ramach zadania nie jest najlepszym wynikiem. Wśród modeli warto wyznaczyć modeli BERT i Jaccard, które dają najlepsze wyniki.

Model BERT zgaduje ocenę na podstawie kontekstu, a model Jaccard'a bazuje się na (między innymi) mocy części wspólnej zbiorów, co też może być potraktowane jako miara kontekstu. Z tego faktu możemy wywnioskować, że temat żartu ma wpływ na średnią ocenę żartu.

Z drugiej strony błędu zgadywania pozostałych modeli są większe ale prawie takie same, z czego możemy wywnioskować, że temat żartu może mieć wpływ na ocenę żartu, ale nie jest to najważnejszym faktorem.