

DYSKUSJA UŻYTKOWNIKÓW W SERWISIE YOUTUBE PORUSZANA TEMATYKA W FILMIE A KOMENTARZE I SŁOWA KLUCZOWE

SŁOWA KLUCZOWE

Słowa które pomagają w wyszukiwaniu filmu przez wyszukiwarkę.
Tworzone mogą być przez twórców, aby zmaksymalizować zasięg i nie zawsze muszą przekładać się na odwzorowanie w nich tematyki filmu. Dodatkowo sekcja komentarzy może zupełnie odbiegać tematem od prezentowanego materiału bądź skupić się na jego części.
Jak można określić tematykę tych zbiorów bez konieczności przeglądania ich wszystkich?

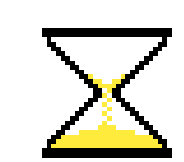
DANE

Dane zebrane z platformy YouTube zawierają: komentarze, transkrypcję filmu oraz słowa kluczowe. Kanały traktują odpowiednio o motoryzacji oraz modzie.

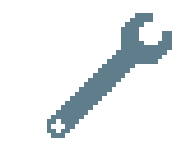
	Najczęstsze słowa		
	Kluczowe	W komentarzach	W transkrypcji
Kanał o modzie	haul clothing try affordable "imię autorki kanału"	love look like makeup hair beautiful	i'm like really going, gonna guys look
Kanał o motoryzacji	belt serpentine diy tensioner	like car one get would good	really like one going know car

Figure 1: Tabela najczęściej występujących słów w zebranych zbiorze danych

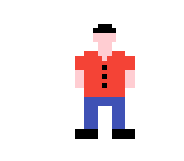
WYMAGANIA



Czas obliczeń



Dobór parametrów



Interpretacja wyników

ROZWIĄZANIE

Do określenia tematów wykorzystano LDA (Latent Dirichlet Allocation) z pakietu gensim.

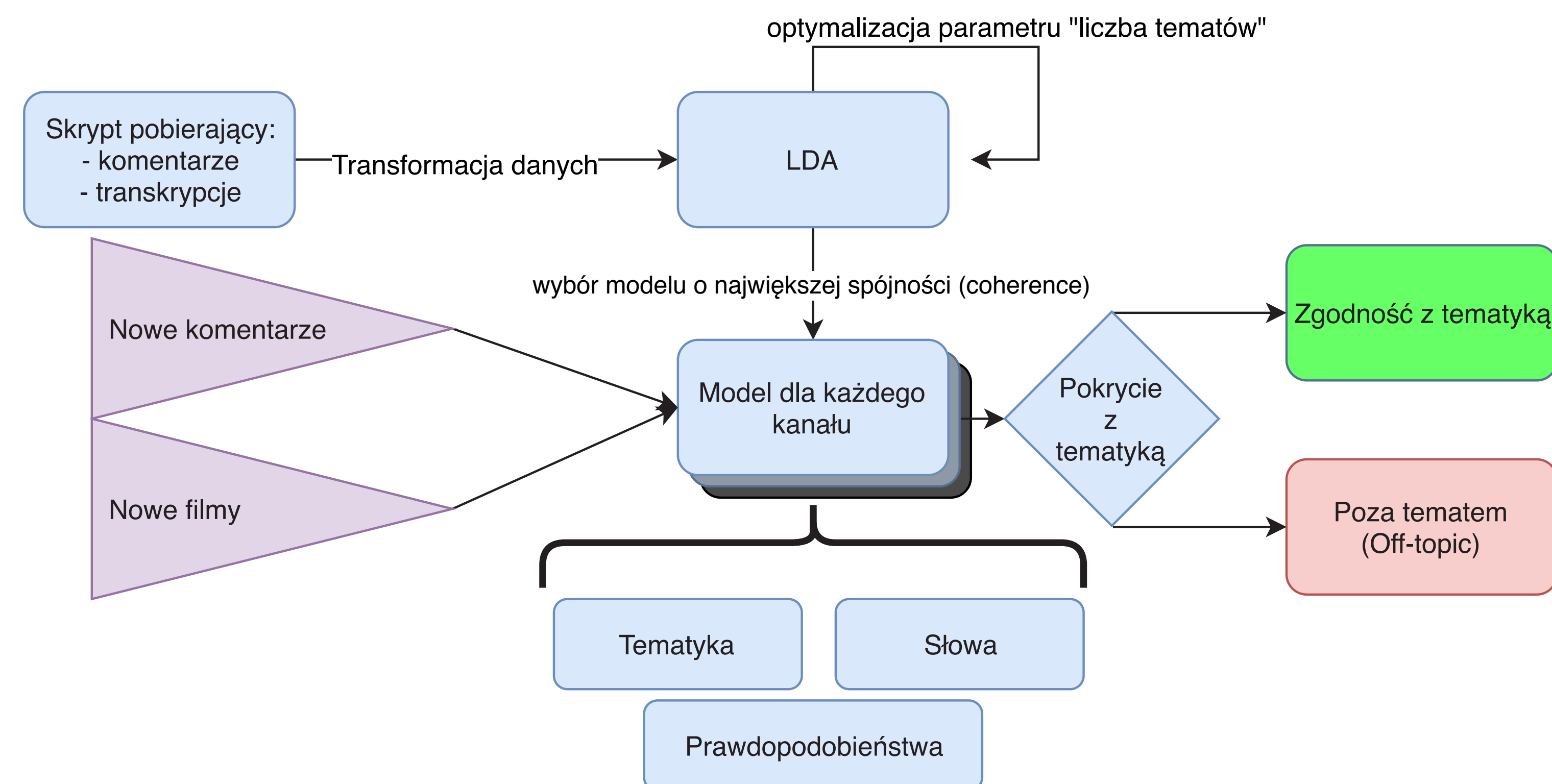


Figure 2: Schemat rozwiązania oparty na modelu LDA z pakietu gensim

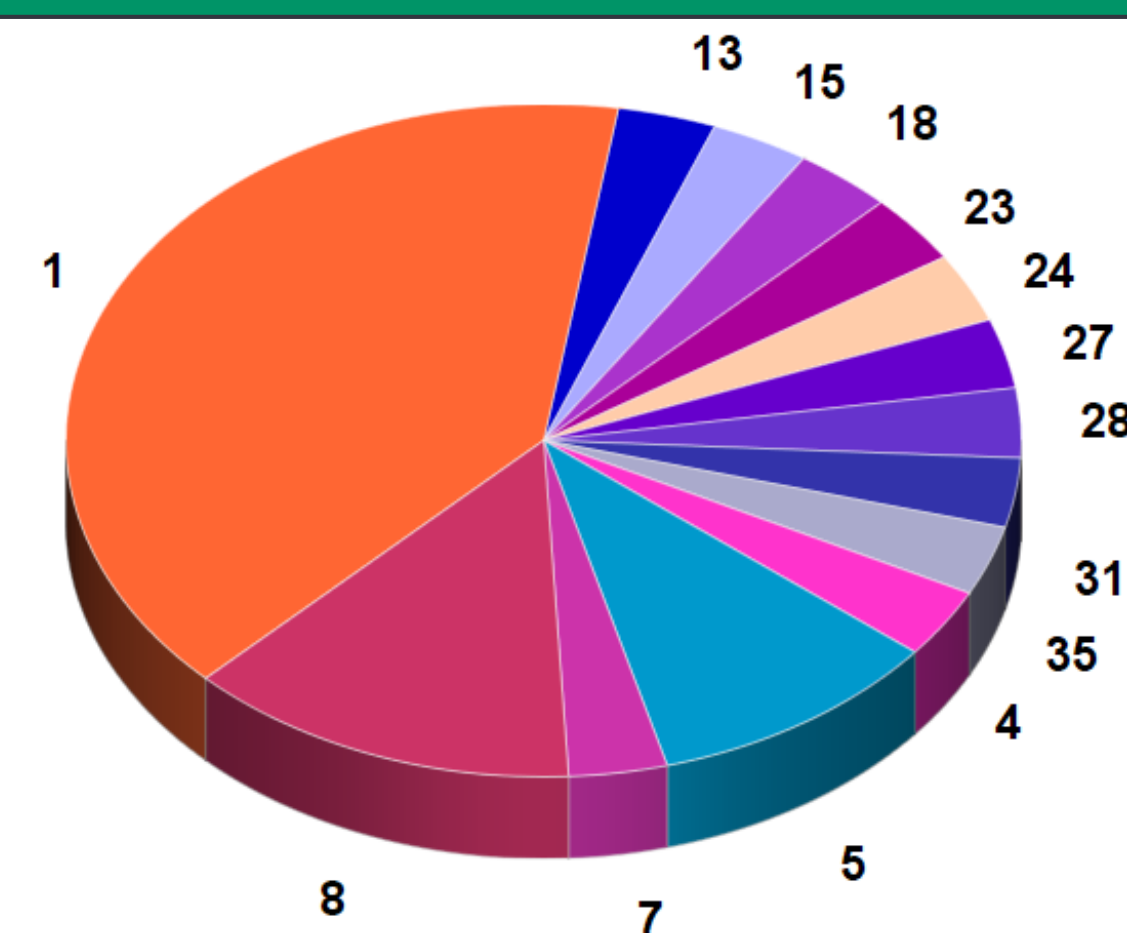
Przykładowe tematy z modelu kanału motoryzacyjnego:

numer tematu	słowa składające się na temat
34	"car", "one", "get", "like", "back"
40	"glad", "thermal", "line", "sort", "heat"
33	"model", "vag", "channel", "image", "nightmare"

Przykładowe tematy z modelu kanału o modzie:

numer tematu	słowa składające się na temat
23	"omg", "subscriber", "gosh", "little", "fine"
4	"pretty", "make", "tutorial", "nice", "check"
7	"look", "brows", "usual", "something", "seen"

ROZKŁAD TEMATÓW



Najbardziej prawdopodobne słowa w tematach:

1 "love" "video"	15 "job" "concealer"
4 "pretty" "make"	23 "omg" "subscriber"
5 "best" "want"	24 "always" "unique"
7 "look" "brows"	27 "girl" "great"
8 "like" "beatuy"	28 "stunning" "absolutely"
13 "eyeliner" "flawless"	31 "talented" "liner"
18 "please" "could"	35 "product" "btw"

Figure 3: Rozkład tematów w kanale(mody).

DOBÓR LICZBY TEMATÓW

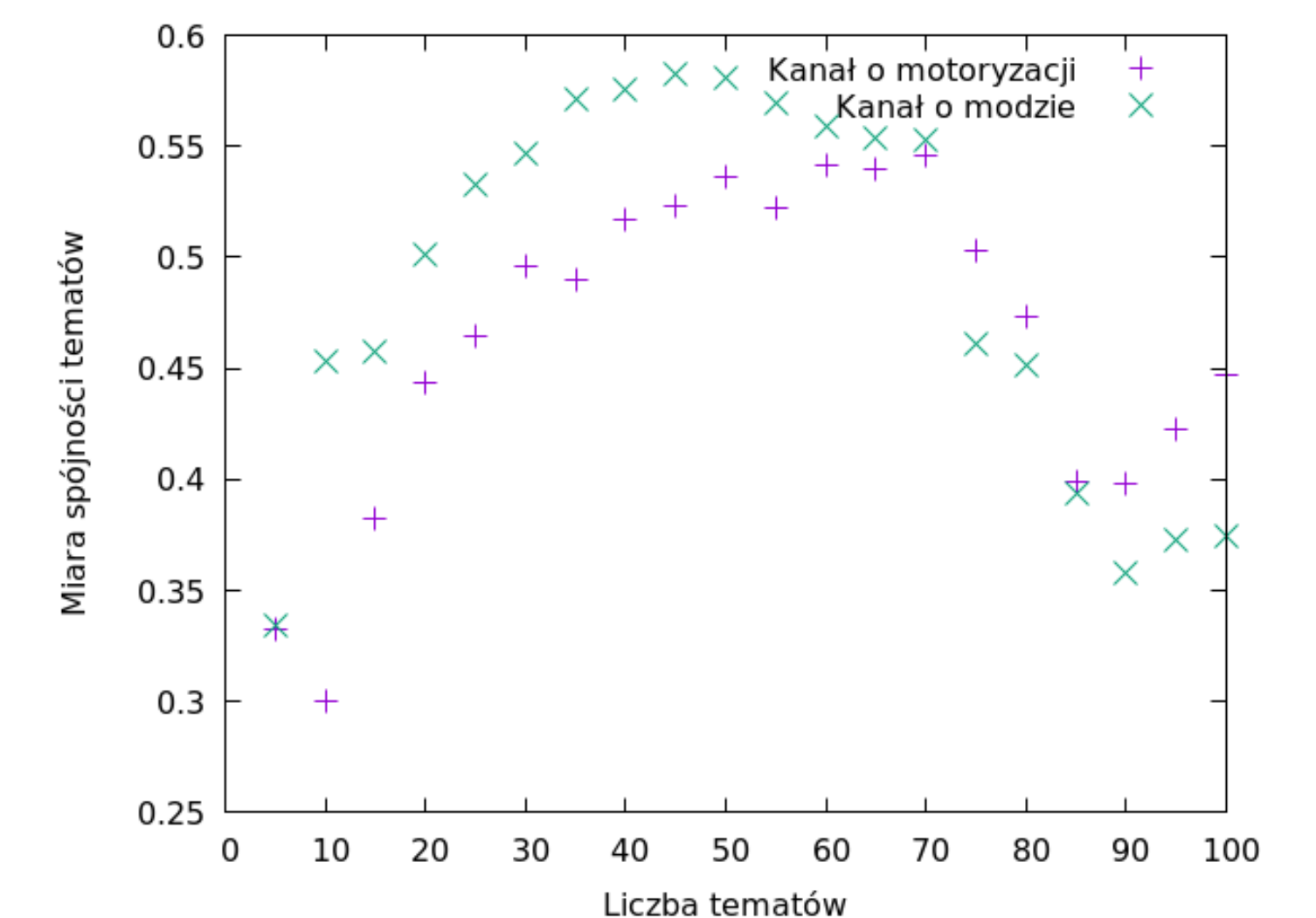


Figure 4: Dobranie parametru odbywa się poprzez szukanie wartości maksymalnej miary spójności dla różnej liczby tematów. Przy czym im mniej tematów tym są one bardziej ogólne.

PRZYKŁAD DZIAŁANIA

"You need to check out the Topshop straight jeans they are unreal! definitely my fav jean at the moment"

Największe prawdopodobieństwo tematu 3 z kanału o modzie. Zawiera on w większości komentarze z powodu często używanych słów przez komentujących.

"Are you really sure you new car has actual tire pressure sensors/monitors in the wheels? Mine hasnt. Gets the info from the ABS system, Im told."

Największe prawdopodobieństwo tematu 34 z kanału motoryzacyjnego. Jest to najczęściej występujący temat w komentarzach, gdyż użytkownicy często piszą o swoich doświadczeniach z samochodami.



Wrocław University
of Science and Technology

Autorzy: Łukasz Drewniak i Maciej Urbaniak
Akredytacja obrazu: Green Simple Abstract Background
from Vector.me (by Vectorportal.com)