

EXTRACTING KEYWORDS FROM MOVIE SUBTITLES

MATÚŠ KOŠÚT
MARIÁN ŠIMKO

matuskosut@gmail.com

HYPOTÉZA

- Využitie špecifík titulkov umožní zvýšiť presnosť extrakcie kľúčových slov z tituliek

CIELE

- Preskúmať možné vplyvy špecifík tituliek na výsledky extrakcie kľúčových slov
- Vytvoriť službu na extrakciu kľúčových slov z tituliek

NÁPADY

- Čas výskytu a doba zobrazenia titulku
- Detekcia konverzácií
- Počet slov za minútu
- Dĺžka úseku medzi titulkami
- Informácie pre nepočujúcich
- Pre jeden film môžu byť dostupné rôzne verzie titulkov od rôznych autorov

Batch processing

Method:

TF-IDF

TextRank

RAKE

View:

Visual

Hash

Limit:

30

GO!

• forrest_gump_1994.srt

forrest	218.29	gump	113.8	jenny	106.86	mama	101.18	shrimp	77.94	lieutenant	73.7	man	73.51	time	73.29	dan	69.3	bubba	62.22		
good	58.48	run	58.29	god	52.38	sir	49.44	back	47.52	boat	46.85	people	46.0	day	45.46	home	40.9	war	37.23	love	36.81
alabama	35.6	president	35.33	stupid	35.32	thought	33.89	told	33.77	school	33.74	ping	32.89	shrimping	32.48	vietnam	32.05	pong	30.77		
romantic	epic	comedy-drama	groom	love	braces	mother	1980s	alabama	team	stranger	president	watergate	platoon	success	lieutenant						
illness	school	football	marathon	apple	college	vietnam	business	home	wealth												

• home_2009.srt

water171.78

earth152.35

world123.85

years98.17

life87.34

energy76.7

species53.88

oil51.67

planet50.46

humans47.77

vocalizing43.13

people43.08

human42.88

ice41.15

sea39.64

billion39.07

country38.84

continues38.42

agriculture38.09

atmosphere37.9

time37.37

sun36.59

nature36.59

carbon36.32

faster35.66

population35.5

city35.29

thousands32.74

resources31.16

million30.64

trees29.9

earth

ecological footprint

environmental issues

environmental crime

environmental issue

environmental politics

green politics

environmental protection

well

water pollution

soy

solar power

oil

oil tanker

oil rig

greenhouse

fishing

environmental

easter island

dubai

deforestation

container ship

pollution

planet

planet earth

panorama

nature landscape

greenhouse gases

global warming

global climate change

environment

earth in peril

climate change

Precision: 6.25

Precision: 6.25

DÁTA

- OpenSubtitles corpus (10G v XML)
- Verejné úložiska tituliek
- Kľúčové slová z Loomie TV