Slovenská technická univerzita
Fakulta informatiky a informačných technológii
Ilkovičova 3, 812 19 Bratislava

Tag Cloud Vizualizácia

Vizualizácia Dát

Peter Válka

28.4.2018

Cvičiaci: Ing. Patrik Polatsek
Študijný odbor: Inteligentné Softvérové Systémy
Ročník: 1. Ing
Akademický rok: 2017/2018

Cieľ vizualizácie

Úlohou projektu vizualizovať slová v 3D priestore pomocou Tag Cloud (Word Cloud) techniky. Projekt je hlavne zameraný pre spracovanie datasetu youtube videí, ale pri miernych úpravách je možné vizualizovať iné dáta. Slová sú vizualizované v jednotlivých stenách, abecedne. Každá stena obsahuje 100 slov. Farba slova reprezentuje dominantnú jeho kategóriu (Kategória videí, ktorá mala najviac použití daného slova v datasete). Dĺžka bokov slova reprezentuje počet kategórií, v ktorom sa objavilo dané slovo. Veľkosť slova reprezentuje jeho počet v datasete. Veľkosť slova sa môže premapovať na reprezentáciu jeho lajkov a dislajkov. Vizualizácia poskytuje filtrovanie podľa jednotlivých kategórií. V priestore sa dá pohybovať lokálne(pomocou šípok a myšky), alebo podľa jednotlivých vrstiev (interakcia s oknom umiestnenom v spodnej časti scény). Pri kliknutí na ľubovoľný tag sa zobrazí okno s detailom. Okno poskytuje dodatočné informácie o tagu.

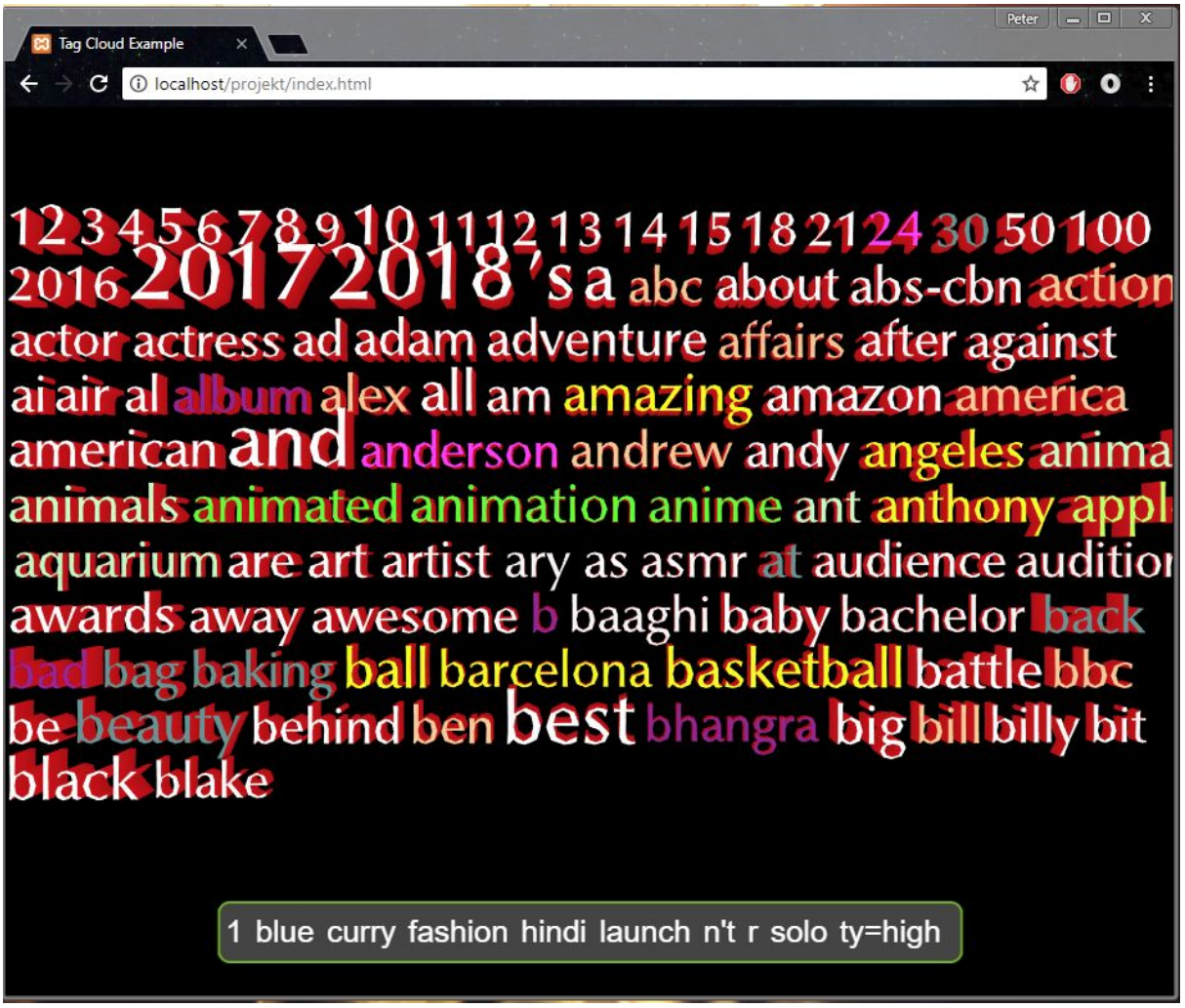
Opis dát a spracovanie

Pre vstupné dáta bol použitý dataset youtube, ktorý obsahuje informácie trendujúcich videí. Zameral som sa hlavne na stĺpec Tags, ktorý obsahuje tagy pre video. Všetky jednotlivé tagy boli spracované a pre každý tag sa vypočítala jeho početnosť (počet videí ktoré obsahuje daný tag), počet lajkov (súčet lajkov videí, ktoré obsahujú tag), dislajky (rovnako ako lajky, ale dislajky), jeho dominantná kategória, a zapísané ostatné kategórie tiež. Do metadát sa zapísali maximálne hodnoty lajkov, početnosti, dislajkov, ktoré sú použité pre výpočet jednotlivých veľkostí. Ako vstupný formát pre vizualizáciu, bol zvolený formát JSON.

Vizuálne Mapovanie

V kapitole opíšem všetky atribúty, ktoré sa mapujú na vizuálne prvky.

Farba Tagu:



Farba tagu reprezentuje jeho dominantnú kategóriu (ako môžeme vidieť, tagy about a adventure majú rovnakú kategóriu).

Veľkosť tagu:

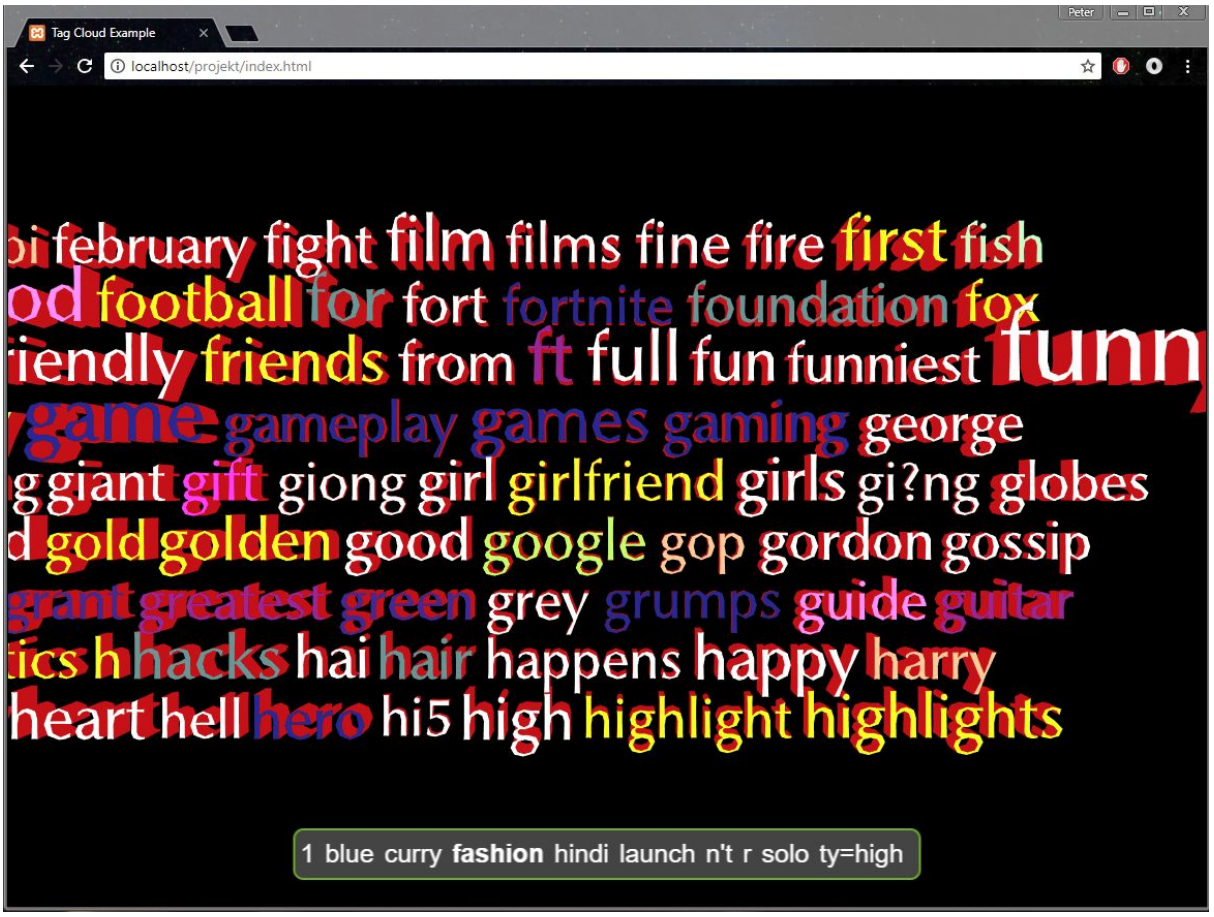
Veľkosť je možné mapovať pre viacero atribútov. Začiatočný atribút je zvolený jeho počet vo videách. Používateľ môže premapovať veľkosť na lajky, alebo dislajky. Nasledujú obrázky znázorňujúce použitie všetkých možností mapovania

Podľa počtu videí

Podľa lajkov

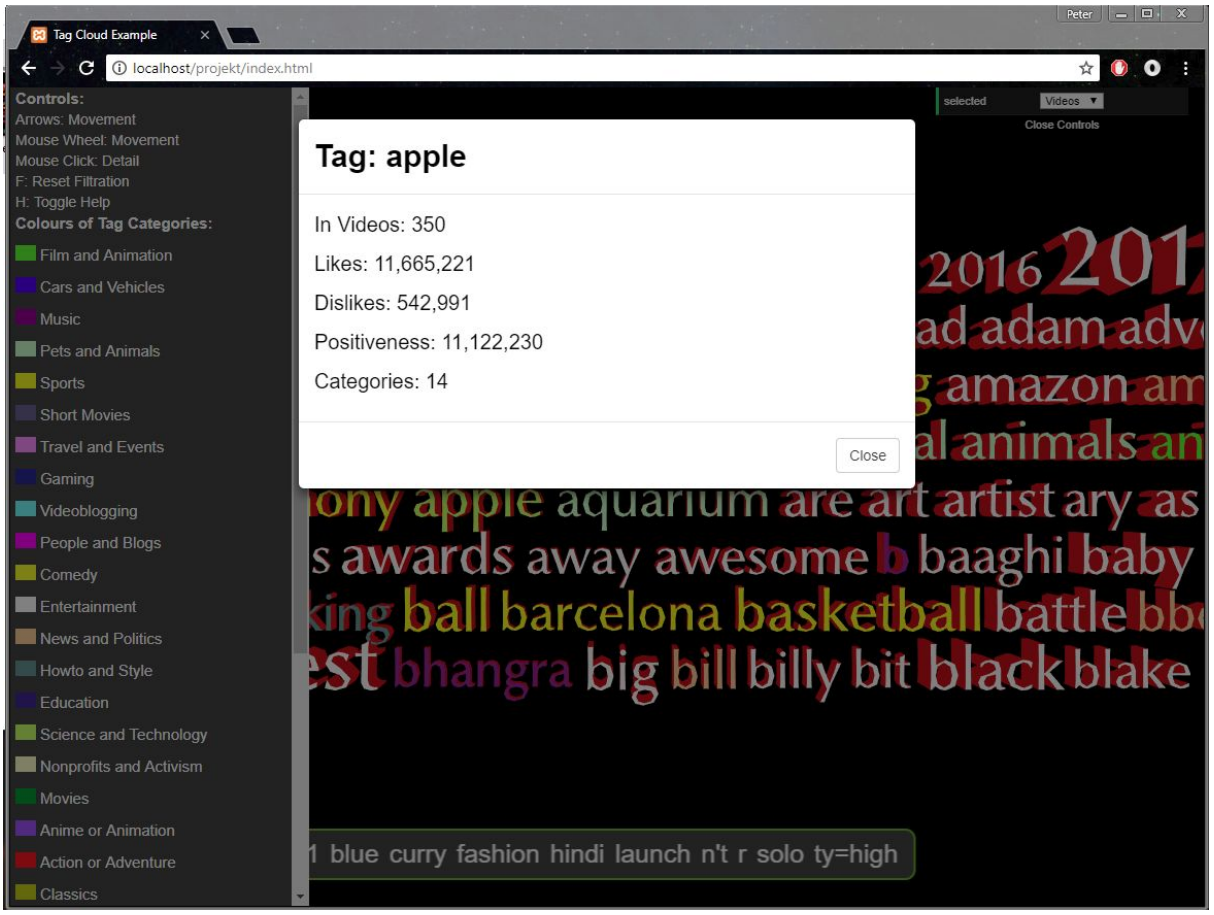
Podľa dislajkov

Hĺbka tagu:



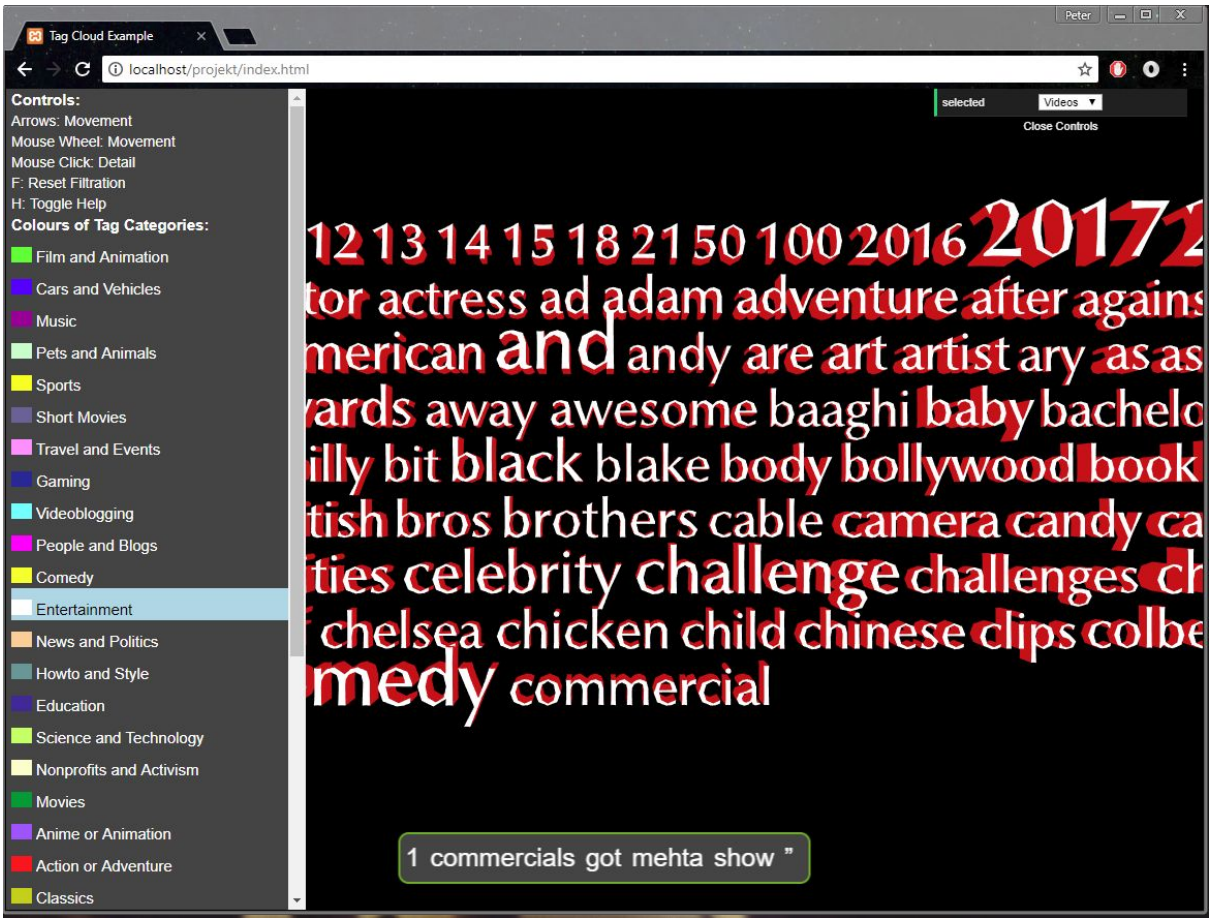
Hĺbka tagu sa mapuje na počet rôznych kategórií, v ktorých sa tag vyskytol. Ako môžeme vidieť, tag game má väčšiu hĺbku ako tag games. Tag gameplay sa objavil iba v 4 kategóriách, ale tag game sa objavil v 14.

Selekcia:



Po zakliknutí ľubovoľného tagu sa zobrazí okno s detailom.

Filtrácia:



Filtrovanie tagov spočíva v zakliknutí želanej kategórie, podľa nej sa začne filtrovať. Na obrázku sú prefiltrované slová. Ostali iba slová, ktoré majú kategóriu Entertainment.

Scenáre vizuálnej analýzy

1. Úloha 1: Je potrebné zistiť, ktorý tag má najviac videí v kategórii Music.
    a. Je potrebné zmeniť vizuálne mapovanie na Videá, pokiaľ je použité iné mapovanie.
    b. Zakliknúť Music kategóriu pre filtrovanie.
    c. Prejsť všetky vrstvy pričom najväčšie tagy zakliknúť a zistiť ich presný výskyt.
    d. Výsledok, by mal byť "music" tag.
2. Úloha 2: Obsahuje kategória Entertainment viac tagov ako kategória Sports?
    a. Prefiltrovať vizualizáciu podľa obydvoch kategórií.
    b. Pre každú kategóriu zistiť, koľko má vrstiev v spodnom okne.
    c. Pokiaľ jedna kategória má viac slov v prehľade, vieme povedať, že má viac tagov.
    d. Pokiaľ majú rovnako slov v prehľade, tak musíme sa presunúť na posledné vrstvy a spočítať počet tagov, pričom vyhrá vrstva, ktorá má viac slov v poslednej vrstve
    e. Kategória Entertainment má viac tagov ako kategória Sports

3. Úloha 3: Má tag “the” viac dislajkov ako tag “paul”?
    a. Prepneme vizuálne mapovanie na dislajky
    b. Nájdeme použitím prehľadu tagy “the” a “paul”.
    c. Selekciou zistíme ich počty dislajkov.
    d. “paul”: 26,534,251 “the”: 9,626,833

Použité vývojové prostredie a knižnice

Grafická knižnica pre tvorbu scény bola zvolená knižnica Three.js. Three je napísaná v Javascripte, preto ako vývojové prostredie, bola použitá kombinácia HTML, CSS, Javascript. Výhodou je spustenie vizualizácie priamo v internetovom prehliadači. Pre vytvorenie okna s detailom tagu sa použila knižnica bootstrap. Pre umožnenie načítania JSON súboru a následnú prácu s ním sa použil XAMPP (server apache pre windows). Natívna stránka nedokáže načítať súbor v počítači.
Výpočet veľkosti tagu:
Nasledujúca sekcia obsahuje vysvetlenie vypočítania veľkosti každého tagu. Veľkosť môže byť daná podľa počtu videí, lajkov, alebo dislajkov, ako základná metrika je použitý počet videí.

```
function calculateTextSize(tag_value, maxVal, minVal){
      var result;
      result = ( (( tag_value - minVal) * (maxLetterSize -
minLetterSize)) / (maxVal-minVal) ) + minLetterSize;
      return Math.min(result, maxLetterSize);
}
```

Premenné funkcie:
tag_value: aktuálna hodnota tagu podľa príslušnej metriky
maxVal: maximálna hodnota tagu podľa príslušnej metriky
minVal: minimálna hodnota tagu podľa príslušnej metriky
minLetterSize: minimálna veľkosť tagu
maxLetterSize: maximálna veľkosť tagu
Výsledkom je veľkosť tagu interpolovaná medzi maxLetterSize a minLetterSize hodnotami na základe aktuálnej tag_value.

Dosiahnuté výsledky

Najviac tagov obsahovala kategória Entertainment. Ako jedinej kategórií sa podarilo prekonať hranicu 100 tagov. Úspech tejto kategórie pripisujem súčasnému trendu Youtuberov. Youtuberi musia

reagovať na aktuálny stav spoločnosti často formou vtipných videí, pričom často vytvárajú obsah, ktorý sa prelína s ostatnými.
Tag “the” mal najväčší počet videí, konkrétne 7,252. Skoro prekonal hranicu 50% výskytu vo videách. Prekvapivo jeho počet dislajkov dosiahol iba 9,626,833, pričom počet lajkov je 136,724,145. Tag sa objavil v 16 rôznych kategóriách.
Tag “video” má najviac lajkov: 144,706,328. Je to prekvapivé, pretože video obsahovalo iba 2,707 videí na rozdiel od tagu “the”, ktoré vyhralo počet. Počet dislajkov má 9,137,315, čo je rovnako slušne nízky počet.
Tag “paul” získal najviac dislajkov: 26,534,251. Objavil sa v 14 rôznych kategóriách a v 1,065 videách. Počet lajkov, má prekvapivo veľký: 107,517,923. Z daného záznamu usudzujem, že aby video bolo označené ako trendujúce, je dôležité, aby malo pozitívny stav lajkov a dislajkov.

Škálovateľnosť

Z pamäťových dôvodov sa nevykresluje kompletný dataset. Vždy sú v scéne vykreslené aktuálne 3 vrstvy, ktoré sa prepočítavajú, v prípade pohybu, alebo zmene vrstvy. Prehľad datasetu sa mi nepodarilo spraviť, pretože vykreslenie 900 slov zaberie 3 GB ramky. Dataset som rozdelil na 3 skupiny. Prvá skupina obsahuje iba tagy, ktorých počet videí je minimálne 100. Konkrétny dataset obsahuje 995 záznamov. Daný dataset bol použitý v ostatných kapitolách projektu. Druhá skupina obsahuje tagy, ktorých počet je minimálne 50. Dataset obsahuje 2108 tagov. Posledný obsahuje tagy, ktorých minimálny počet je iba 10. Daný dataset je preto najväčší a obsahuje 8369 záznamov. Všetky datasety išlo bez problémov vizualizovať, keďže sa znázorňuje iba ich časť.

Prílohy

Moje Konzultácie nad kolegovim projektom:

Konzultácia 2 s Markom Ondrušom, písal Peter Válka:

Cieľom vizualizácie je zobraziť porovnanie medzi premávkou, ktorá obsahuje, alebo neobsahuje DOS útok.
Identifikoval som objekt vizualizácie: cieľový port premávky, pričom pozícia v grafe označuje početnosť.
Vedel som identifikovať 2 rôzne porty a podľa ich pozície ich porovnať využitie v premávke.

Pridal by som legendu, ktorá zobrazuje ktorý atribút premávky je zobrazovaný. Rovnako extra informácie by som navrhol zobraziť pri nastavením kurzora na určitý vrchol.

Konzultácia 3 s Matejom Uhlíkom, písal Peter Válka:

1.1 Otvor filter a zavri. Vyskúšaj si označiť niektoré guličky na grafe.

Guličky sa zvýrazňujú, a aj čiary. Problém je, že sa vyznačujú aj čiary, ktoré nesúvisia z guličkou. Treba opraviť algoritmus zvýrazňovania čiar.

1.2 Vyznač a zisti koľko pripojených čiar je ku ľavej dolnej najväčšej guličke?

Identifikácia prebehla rýchlo a výsledok je presný. Bohužial nepresný pri väčšom počte čiar.

1.3 Označ guličku a pozri sa na ňu z blízka.

Označenie a ovládanie bolo jednoduché.

1.4 Početnosť predstavuje veľkosť guličky. Nájdi najväčšiu a zisti, koľko čiar ku nej smeruje?

Bez prítomnosti filtrácie táto úloha, nie je jednoduchá. Po krátkom čase ale subjekt identifikoval správnu guličku, a spočítal jej počet čiar.

Exploratívna Analýza, písal: Peter Válka

Vo vizualizácií sa zobrazovali údaje z pobočiek Juke, Avol, Perly, Jubano, Einsty.
Každá transakcia má pri sebe opis, čiže je jasné, čo prezentuje.
Pri rozkliknutí určitého stavu sa zobrazí viacero detailov o transakcií.
Zistil som, že veľkosť stavu (guličky) symbolizuje počet transakcií, ktoré sú spojené s ostatnými.
Použitie filtrov je veľmi vhodné, pretože pri zisťovaní konkrétnych údajov, môžu ostatné vrstvy prekážať.
Rekonfigurácia spôsobila “pád” celého grafu, ciže zamenili sa osi x a y.

Kolegove Konzultácie nad mojim Projektom:

Konzultácia 2 vykonával: Marko Ondruš

Word Cloud, video tagy on YT, veľkosť = pocetnost, farba = kategória
zoradené podľa abecedy
riadky nesúvisia, keďže ideme abecedne

pri malom zoome su farby niektorých kategórií nejednoznačne a teda je potrebné použiť zoom pre identifikovanie kategórií podľa farby
chýba zoznam kategórií aké majú farby
neviem identifikovať kategóriu podľa farby, napr. vzťah comedy a chicken
christmas a cinema, podľa farby viem že sú v iné kategorie, ale ako spomenuté, neviem určiť, o aké kategórie sa jedná
viem porovnať aj ich vzajomnu početnosť v danej kategórií
veľkosť sú absolútne a nezávisle od početnosti elementov v danej kategórií
početnosť elementu je na on-demand, aktuálne neviem určite, len zo základného pohľadu jeho početnosť
ktoré elementy majú početnosť nad x? to sa nedá určiť, toto bude riesene pravd. pomocou selekcie - ktoré slovo je najviac z danej kat.
malá rozlíšiteľnosť pri podobných hodnotách - chyba samotnej technologie
prinos? - podla použiteľnosti slov sa zamerať marketingovo, alebo na použiť na vlastné video a tak získať potenciálne viac videní
scrollovanie medzi blokmi/obrazovkami je realizované cez 3D prechod vrstvy aktualneho bloku - neviem podľa oka porovnať dve slová v rôznych blokoch

Konzultácia 3 vykonával: Matej Uhlík

Scenáre:

1. Skri legendu

Samému sa nepodarilo vykonať scenár, pretože info o ovládaní nebolo v popredí, ale bolo treba scrolovať (fix info o ovládaní treba dať do popredia)

2. Vyber si určitý tag, pričom zisti, či má viac lajkov, alebo dislajkov

Scenár dopadol úspešne, neboli žiadne pripomienky

3. Nájdi posledný tag podľa abecedy, ktorý je zobrazený

Scenár dopadol úspešne

4. Zvoľ si určitú kategóriu tagov, a prefiltruj vizualizáciu, podľa zvolenej kategórie (legenda)

Scenár dopadol úspešne

5. Nájdi najpopulárnejší tag z filtrovaných

Scenár dopadol úspešne

6. Resetuj vizualizáciu (legenda)

Scenár dopadol úspešne

Exploratívna Analýza, vykonával: Matej Uhlík

Objavil som základnú funkcionalitu celkom rýchlo. Po krátkom používaní som zistil novú možnosť, pohyb po rozsahu dát na lište dole na obrazovke. Vizualizácia zobrazuje početnosť videí podľa tagu, ale je aj schopná to zmeniť a zobrazovať početnosť „like“ alebo „dislike“ atribútov. Pri základnom zobrazení, je vidno aj legendu, v ktorej sa nachádza aj základná funkcionalita pre ovládanie vizualizácie. Jednoduchá selekcia podľa kategórií je veľmi intuitívna, a vhodne zvolená.
Pri zmene prideľovania veľkosti, podľa iného atribútu sa ale veľkosť písma niekedy prelína s textom na vyššom riadku. Podľa môjho názoru je tento efekt nie veľmi príjemný na pohľad. Možnosť zmeny určovania veľkosti pri jednotlivých „tagoch“, je trošku neviditeľné a jeho hľadanie by mohlo trvať dlhšie, ako je očakávané.
Inak je vizualizácia vhodne zvolená, a pohyb v nej je rýchly. Informácia je reprezentovaná vhodne a početnosť je rýchlo identifikovateľná aj pri zmene konfigurácie.