OBSAH

VZ vyhledávání a indexace

Český rozhlas hodlá provést automatizovanou transkripci všech svých pořadů dle dostupných záznamů kontinuálního vysílání od roku 2003. Tuto transkripci požaduje zpracovat pro rychlé a přehledné vyhledávání s návazností na konkrétní čas v rámci audia.

1. Popis situace

Český rozhlas (ČRo) hodlá zpřístupnit některé obsahy vysílání v textové podobě pro snadné hledání i sledování obsahu živého vysílání v textové podobě. Pro tento účel je vyvíjí Analytiku mluveného slova (AMS) a Analytiku vysílání (AV). Obě analytiky pracují souběžně při zpracování živého vysílání, AV identifikuje zvuky, jejich začátky a konce. AMS identifikuje řeč, provádí transkripci, diarizaci a identifikaci mluvčích s tím, že promluvy přiřazuje k hlasovým profilům - identitám anonymním i identifikovaným. Tím obě analytiky vytvářejí archiv vysílání ihned z vysílání dostupné online se zpožděním v řádu nižších minut. AMS živého vysílání bude (pravděpodobně) probíhat dvoufázově. V první fázi půjde o rychlost AMS, v druhé po odvysílání celé epizody ve vysílání bude provedena AMS hloubková s důrazem na přesnost.

Předmětem zpracování je nejen živé vysílání s okamžitými výstupy, ale také zpracování archivních záznamů kontinuálního vysílání ČRo od roku 2003. AMS archivu bude probíhat průběžně od aktuálního vysílání do minulosti. Zpracování živého vysílání bude přímo navazovat na zpracování archivu.

Předpokládáme, že s dalším vývojem AMS a AV dojde ke zpřesňování obou analytik a bude proto nutné indexovat již zpracované vysílání opakovaně.

Použité zkratky a výrazy:
ČRo - Český rozhlas
AMS - Analytika mluveného slova, zjednodušeně transkripce vysílání
AV - Analytika vysílání, zjednodušeně identifikace zvuků ve vysílání
SERP - Search engine result page, stránky s výsledky hledání
HW - hardware
SW - software
FE - Front end, vizuálně obslužný prostor v prohlížeči uživatele
Podcast - audio vytvářené primárně pro internet, neprošlo vysíláním

2. Popis zakázky

2.1 Co je úkolem dodavatele

1. Dodavatel dodává návrh - architekturu řešení s ohledem na další možný vývoj.
2. Dodavatel vyvíjí a dodává systém pro automatizované zpracování výstupů AMS a AV (text - transkripce, osoby, pořady vysílání takové, které umožňuje uživatelům rychlé a snadno parametrizovatelné hledání v obsahu vysílání v rámci projektu mujROZHLAS.

3. Dodavatel vyvíjí a dodává API pro výstupy na SERP či jiných stránkách, na kterých bude probíhat zadávání hledání i jeho výsledky.
4. Dodavatel definuje architekturu, HW s OS na kterém systém poběží. Dodavatel může pro zpracovávání využívat vlastní či i cloudová řešení externích dodavatelů. Základní procesy budou probíhat na HW ČRo.
5. Dodavatel definuje nároky na správu systému za období platnosti zakázky.

2.2 K čemu ČRo hodlá systém využívat

Využití

- Podrobné vyhledávání v celém archivu.
- Dohledání v právě probíhajícím vysílání (AMS živého vysílání)
- Hledání osob, o nichž se mluví (téma / štítek)
- Hledání osob, jež mluví ve vysílání i s jejich promluvami ve vysílání
- Hledání pořadů / epizod / bloků vysílání
- Automatické štítkování obsahu pro snadnější kategorizaci a dohledání
- Výsledky dle práv uživatelů (práva řeší ČRo) - systém budeme chtít využívat i pro interní účely nebo speciální přístupy - “badatelské” a pak běžné public - normální návštěvníci

2.3 Předmět zpracování systémem

Předmětem zpracování jsou transkripce, jména osob, názvy pořadů a epizod.
Předměty ke zpracování jsou

- Výstupy z AMS archivu kontinuálního vysílání, dodávané průběžně
- Výstupy z AMS živého vysílání
    - Výstupy z online AMS
    - Výstupy z hloubkové AMS
- Výstupy z AMS obsahů vyvíjených pro mujROZHLAS (Podcasty apod.)

2.4 Systém - jeho funkce

a) Analytika a indexace
    1. archivních transkripcí odvysílaných pořadů,
    2. transkripcí postupně dodávaných dle možností dodávající třetí strany
    3. transkripcí vytvářených online ze živého vysílání
    4. osob (autoři promluvy, autoři obsahu, zmíněné osoby)
    5. tematizace obsahu (téma je všeobecnější, které zahrnuje štítky, např. vesmír, zdraví a umožňuje nabízet obsahy na HP jako sekci) - zde je otázka, zdali neřešit obecnějším štítkováním
    6. štítkování obsahu
b) Slučování výsledků (jedna epizoda)
c) Kategorizace indexovaného obsahu dle přístupových práv uživatelů (uživatel výsledek vyhledá, ale nebude pro něj dostupný v plném rozsahu / uživateli se nepřístupné výsledky nezobrazí) - autentizaci řeší ČRo.
d) Vyhledávání s možnou filtrací jak pro vyhledávání, tak pro výsledky hledání
    a. obsahy / osoby / témata? / štítky viz FE systému
e) Výstup vyhledávání dle práv uživatele

f) Search engine prohledává i v obsazích, které nemají povolenou transkripci nebo již není možné je spustit a nabízí je na SERP
g) Rozšíření na další projekty ČRo (iROZHLAS, ROZHLAS.cz) - spíše jde o otevřenou možnost rozšířit systém o další obsahy mimo projekt mujROZHLAS.
h) API
i) Personalizace výsledků
j) Opětovná analytika a indexace výběru vysílání při:
    a. novém zpracování transkripce
    b. opravě transkripce
    c. novém klíčovém slovu (třeba czexit)
    d. štítky (klíčová slova)
        i. téma zahrnuje skupinu klíčových, v podstatě jde o kategorii, pod níž se slučují klíčová slova
    e. hloubkové analýze AMS
k) Slovníky
    a. zařizuje dodavatel a jsou součástí nabídkové ceny

2.5 FE systému

Vyvíjí ve spolupráci s ČRo. Dodavatel dodává API a ČRo definuje požadavky a vyvíjí FE.
Níže je popis FE pro informaci dodavateli, s jakými daty na FE ČRo počítá.

2.5.1 Zadání hledaného výrazu

Input s našeptávačem

2.5.2 Parametrizované hledání

Kategorie
- Pohádky
- Detektivky a krimi
- Povídky
- Četba
- …

Téma
- Věda
- Příroda
- Vesmír
- Zdraví
- Sport
- ...

Čas
- Bez omezení (od nejnovějších)
- Termín od do
- 24 hodin (dnes a včera od 0:00)

Délka
- nastavit délku
- do 15 minut
- do 30 minut
- do 1 hodiny

Stanice / projekty

- mujROZHLAS
- Radiožurnál
- Plus
- Dvojka
- Vltava
- Wave
- …

Pořady
- Našeptávač
- Podcasty (audia pro internet)

Osoby
- Našeptávač

Štítky / klíčová slova
- Našeptávač

Práva
- Pouze výsledky s možností přehrání, defaultně zaškrtnuté)

2.5.3 Filtrování a řazení výsledků

Filtry
- Osoby
  - Konkrétní osoba / osoby
- Zprávy
- Datum - rozmezí
- Premiéry / poslední reprízy
- Štítky
- Subkategorie
  - Děti malé
  - Děti školní
  - Teenageři
- Délka epizody

Řazení
- Podle data vydání
- Podle oblíbenosti
- Personalizované

2.5.4 Výsledky hledání

Na výsledné stránce by měl být mix hledání dle váhy, kterou by měl určit dodavatel ze svých zkušeností s vyhledáváním uživateli.

Obsah výsledků hledání na SERP
- Možnosti
  - Zobrazit počet výsledků
  - Zobrazit / skrýt i výsledky bez možností přehrání
- Objekty (vše co nevede přímo ke konkrétní audiostopě):
  - Pořady
    - do autoplaylistu
    - do mujROZHLAS playlistu
    - hlídací pes

    - Projekty
    - Osoby
    - Stanice
    - Štítky
- Konkrétní úsek audia v epizodě / bloku
    - Obsahující hledaný výraz
    - Odkazující na konkrétní úsek v audiu
        - Kliknutím na takovýto výsledek přenese uživatele do audia na konkrétní časový úsek, řeší ČRo
    - Je možné přehrát daný úsek přímo na SERP
    - Položky
        - Název epizody
        - Název pořadu
        - Datum vysílání
        - Délka epizody
        - Díl epizody
        - Odkaz na všechny díly v případě, že jde o pokračování
        - Část textu s hledaným výrazem
        - Hlídat nové epizody pořadu
        - vložit do autoplaylistu
            - tuto epizody
            - tento pořad
        - Vložit do mujROZHLAS playlistu
            - tuto epizodu
            - tento pořad
        - Hlídací pes
            - tohoto pořadu
- Sloučené úseky v rámci epizody
    - Pokud je hledaný výraz v rámci jedné epizody, na SERP bude vypsána tato jedna epizoda s možností rozbalení výsledků a výběru toho požadovaného s kliknutím do epizody a přehráváním, nebo přehráváním přímo na stránce hledání.

2.6 API

- Rozhraní pro získávání výsledků hledání podle
    - zadaných parametrů,
    - práv uživatele.
- REST API

2.7 Budoucnost

Níže uvedené není v rámci této zakázky požadováno ani řešeno:
ČRo bude mít zájem na vytváření automatických shrnutí z epizod a to nikoliv ve formě vybraných vět, ale souvislého textu do ca 500 znaků.
V druhém, paralelním kroku, se chce ČRo podílet na vývoji text-to-speech systému, který by umožňoval našim posluchačům / uživatelům shrnutí vybraných pořadů (personalizované /customizované nastavení).

3. Hardwarové požadavky

Dodavatel navrhne optimální řešení viz bod 2.1.4.

4. Use case vyhledávání

1. Hledání mluvčího a tématu v určitém období
    a. - O rozšíření Temelína mluvil Miloš Zeman už koncem prvního desetiletí.
2. Hledání pořadu v určitém období
    a. - Byl to pořad myslím s Čechem někde na Dvojce
3. Hledání epizody na téma
    a. - Včera jsem slyšel někde v radiu, jak se bavili o schizofrenii
4. Hledání pořadu
    a. Chci poslední Meteor.
5. Hledání osoby / moderátora
    a. Kovařík, ten uvádí nějaký pořad.
6. Hledání podle délky a kategorie
    a. Potřebuji nějakou četbu na hodinu do letadla
    b. Pohádku pro malé děti do 30 minut
7. Hledání konkrétního dílu
    a. Ve třetím díle četby Husáka to myslím bylo.