

## NAKI-II-USTR-UKONCENE - Task #3802

Task # 3633 (Closed): Etapa 01 - Příprava dat a datových struktur, testy existujících metod

### Témata - unsupervised

09.03.2016 13:09 - Zajíc Zbyněk

<b>Status:</b>	Closed	<b>Start date:</b>	09.03.2016
<b>Priority:</b>	Normal	<b>Due date:</b>	01.05.2018
<b>Assignee:</b>	Ircing Pavel	<b>% Done:</b>	0%
<b>Category:</b>		<b>Estimated time:</b>	0.00 hour
<b>Target version:</b>			
<b>Description</b>			
LSk na dodaných prepisech od ÚSTRu zkusí unsupervised shlukování			
<b>Related issues:</b>			
Related to NAKI-II-USTR-UKONCENE - Task #3931: Test ASR		<b>Closed</b>	<b>30.05.2016</b> <b>31.08.2016</b>

### History

#### #1 - 31.05.2016 14:19 - Skorkovská Lucie

- Due date changed from 31.05.2016 to 30.06.2016

podívám se na dodané prepisy na disku a co se s nimi dá dělat

#### #2 - 07.06.2016 15:17 - Zajíc Zbyněk

- Related to Task #3931: Test ASR added

#### #3 - 07.06.2016 15:27 - Zajíc Zbyněk

- Due date changed from 30.06.2016 to 31.08.2016

#### #4 - 03.10.2016 15:37 - Zajíc Zbyněk

- PI předá prepisy a kontakt na doktoranta co se zabývá tématy, LSk diplomanta zaúkoluje a společně nashlukují témata z dat z [#3931](#)

#### #5 - 07.10.2016 15:34 - Skorkovská Lucie

Domluvili jsme se s Jaromírem Novotným na vyzkoušení sumarizačních metod z jeho diplomové práce. Vymysleli jsme varianty prvních pokusů, na co by se to dalo použít, podle prvních výsledků se pak vymyslí jestli je to reálně použitelné:

- sumarizovat automaticky spojené větší množství otázek a odpovědí (třeba 10, nebo aby to mělo nějaký počet slov...)
- sumarizovat dlouhé výpovědi v kuse, ty pak zkusit shlukovat - mohlo by pomoci shlukovacímu algoritmu
- sumarizovat celý hotový shluk - získat tak jeho lepší popis než jednotlivými "důležitými" slovy co se získají z lsa a podobných metod

#### #6 - 06.02.2017 12:39 - Skorkovská Lucie

- Podle J. Novotného je jeho sumarizační algoritmus nepoužitelný.
- Shlukování jednotlivých textů selhává na jejich délce, je třeba je nějak rozdělit.
- Texty je třeba rozdělit podle témat, testuji na to použití LDA podle rady Honzy Lehečky. Pak se budou dát možná shlukovat, nebo hledat podobné k vybranému textu - úseku.

#### #7 - 06.02.2017 18:26 - Zajíc Zbyněk

Ze schůzky [[<https://wikky.zcu.cz/redmine/projects/naki-ii-ustr/wiki/Schuzka16-10-03>]]:

- **LSk** prepisy audiodat, obsahují teoreticky mnoho témat (životní výpověď' očitých svědků)
  - supervised metody nemáme data
  - unsupervised moc dlouhé dokumenty
  - budeme hledat spíše tématicky podobné dokumenty k danému doc.
- **dočkat** -> rozsekat automaticky na tématicky ucelené bloky a ty rozpoznat (přes LDA)
  - > nastudovat LDAtoVec

#### #8 - 08.11.2017 15:55 - Zajíc Zbyněk

- Due date changed from 31.08.2016 to 01.05.2018

- Assignee changed from Skorkovská Lucie to Ircing Pavel

- vstupní data rozsekat na cca 2.min. dokumenty s nějakým překryvem
- uživatel klikne do nahrávky a z okolí +-1min se vytvoří hledaný dokument
- nalézt N nejpodobnějších dokumentů k hledanému
- aplikovat jednu z metod z **LSK** disertačku
- s **JS** domluvit formát uložených indexů slov d DB

**#9 - 18.11.2019 09:08 - Zajíc Zbyněk**

- Status changed from New to Closed