NLP && production

BigARTM: Генерируем плейлисты

Штех Геннадий
NAUMEN
28.04.2018

Задача

Как видим результат?

Плейлист

1. Сделать случайный набор треков, похожих на данный один или несколько треков
2. Сделать классный плейлист с плавным переходом по темам

Что такое “плавный плейлист”?

Трек 1: Linkin Park - Numb
….
….
….
….
….
….
….
Трек 2: Eminem - Slim Shaggy

Данные

lastfm360k + musikbrainz + lyrics.wikia

LastFM 360k

1. 360 000 плейлистов
2. В плейлистах только артисты(без треков) и количество прослушиваний
3. MBId(MusicBrainz)
4. Немного инфы о пользователе

MusicBrainz

1. Свободная подробная вики о музыке
2. Есть информация о всех треках и релизах, жанры, лейблы, ВСЁ!
3. ….
4. Кроме текстов песен

Lyrics.wikia

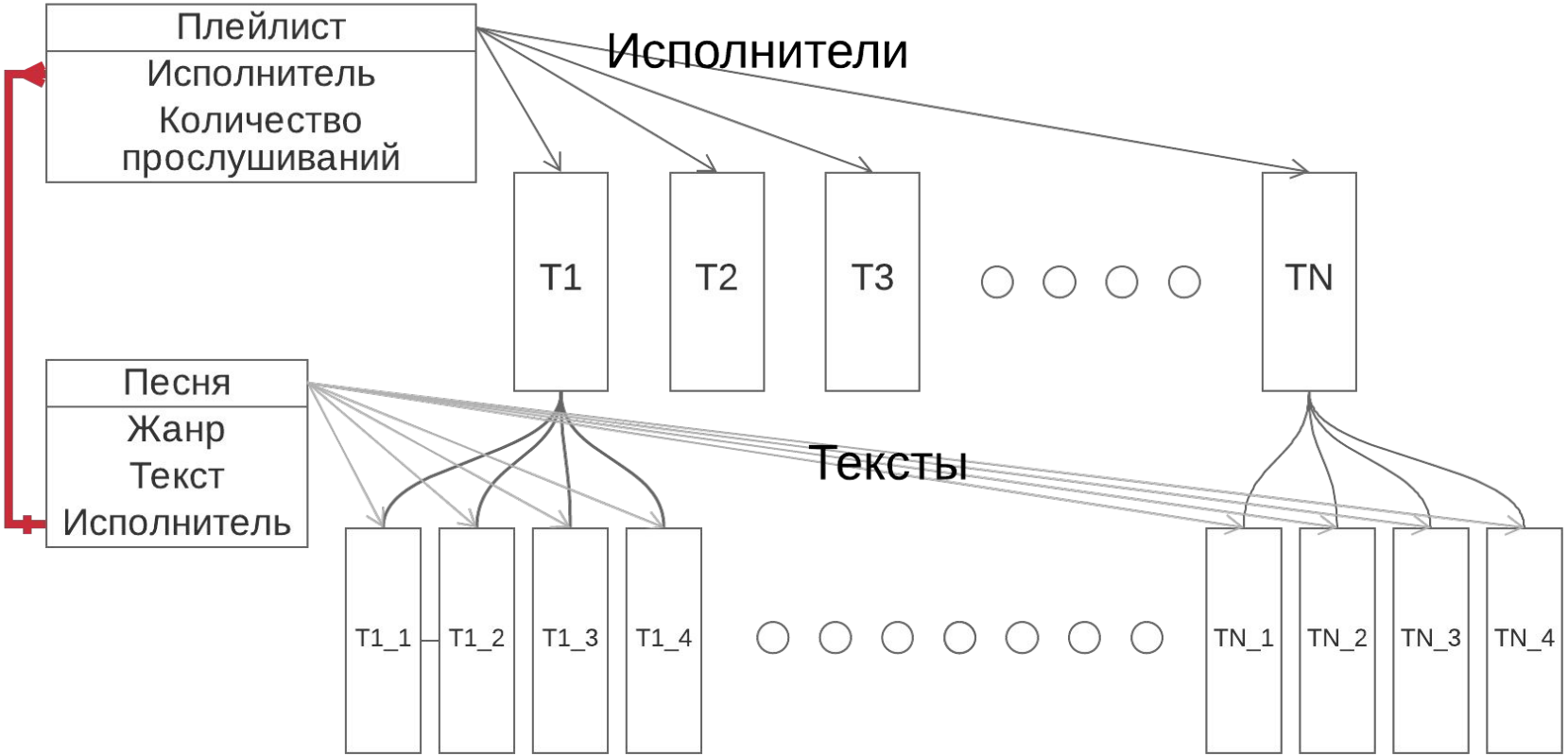
1. Несвободная вики о музыке
2. ….
3. Есть тексты песен

Схема данных, как связаны сущности

Немного о модели

Hierarchical ARTM

Иерархическая модель



Код?

workshop.g-sh.tech

github.com/sht3ch/public_workshop

Формат плейлистов

0000ee7dd906373efa37f4e1185bfe1e3f8695ae |artist stam na dream theater ac dc metallica iron maiden bob marley gentleman mötley crüe queen manowar tool killswitch engage slayer in flames skid row stone bullet for my valentine mokoma pantera brother firetribe hurriganes norther stratovarius kiuas ozzy osbourne |artist_f stam1na dream_theater ac/dc metallica iron_maiden bob_marley_&_the_wailers megadeth children_of_bodom tbullet_for_my_valentine mokoma pantera brother_firetribe hurriganes norther stratovarius kiuas ozzy_osbourne

Формат текстов песен

christ-alone-live-in-studio |text fly mortal earth measur depth girth father sai worth jesu death birth measur dollar sign brick mortar christ worthi golden crown worthi golden crown valu life live love forgiv treasur truli lai start end dai measur battl won good deed christ worthi golden crown worthi golden crown pauper king question beg belong sing resurrect song measur master hand truth stand christ worthi golden crown worthi golden crown |artist edens edge

Как выглядят темы

topic_48 ||| katy_perry:0.053, maroon_5:0.048, the_fray:0.042, robbie_williams:0.036, pink:0.035, lifehouse:0.035, james_blunt:0.034, avril_lavigne:0.034, dido:0.029, amy_macdonald:0.027, keane:0.024, coldplay:0.019, natalie_imbruglia:0.017, no_doubt:0.016, take_that:0.015
topic_49 ||| justin_timberlake:0.057, alicia_keys:0.044, mariah_carey:0.030, ne-yo:0.030, beyoncé:0.029, chris_brown:0.028, john_legend:0.026, usher:0.022, erykah_badu:0.018, kanye_west:0.017, janet_jackson:0.017, black_eyed_peas:0.016, timbaland:0.015, michael_jackson:0.014, lauryn_hill:0.014
topic_50 ||| oasis:0.112, coldplay:0.089, muse:0.076, radiohead:0.069, manic_street_preachers:0.036, gorillaz:0.035, the_verve:0.031, franz_ferdinand:0.026, the_killers:0.024, placebo:0.024, the_smashing_pumpkins:0.021, jamiroquai:0.018, r.e.m.:0.017, moby:0.016, the_beatles:0.016

Близость темы

Как построить “траекторию”?

Если бы это были *эмбеддинги*

1. Строим эмбеддинги
2. Берем вектора объектов A и B
3.
```
for alpha in np.linspace(0, 1, 100):
    get_similar(alpha*A + (1 -alpha)*B)
```

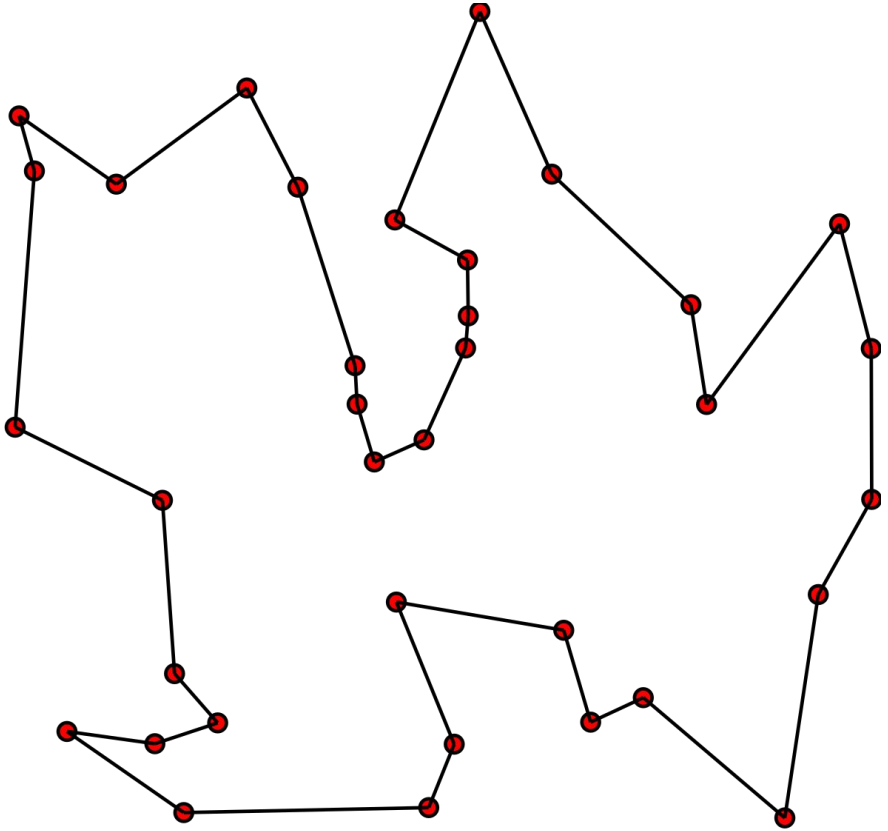
Близость на *векторах тем*

1. Строим тематическую модель
2. Дополняем “путём” по темам с помощью **Traveling Salesman Problem** с весами

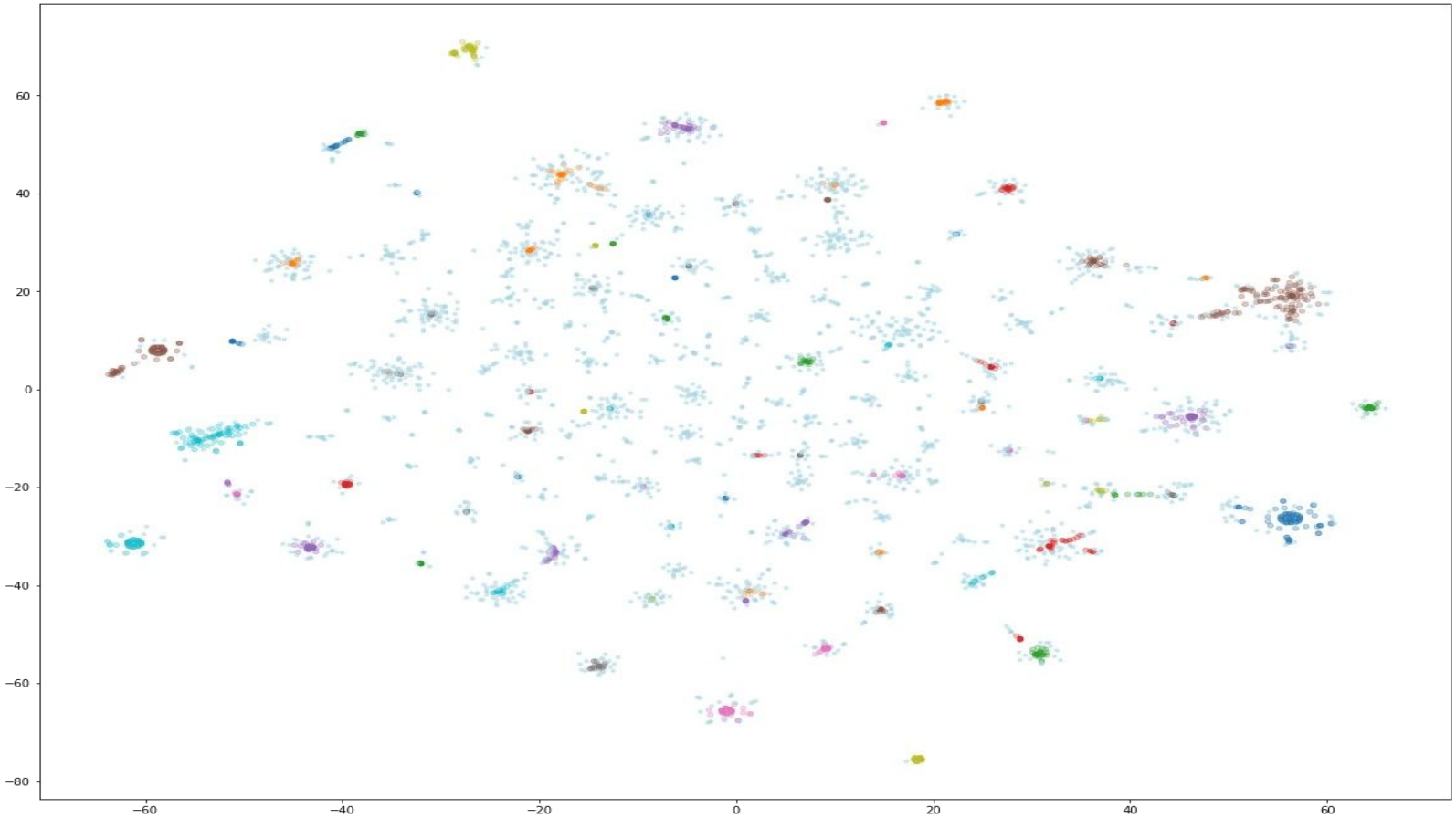
$$G(E, T):$$
$$e(t\_1, t\_2) = cosine(topic\_1, topic\_2)$$

3. Берем вектора объектов A и B
4.
```
for alpha in np.linspace(0, 1, 100):
    get_similar(path(G, A, B) * alpha)
```

*Красивая картинка про “путь”



*Красивая картинка про “путь”



Актуальные алгоритмы

Представление
- Tf-idf, nPMI, hashing trick

Факторизация (декомпозиция)
- PCA, LSI-LSA, pLSA

Тематическое моделирование
- pLSA, LDA, HDP, ARTM

Поиски
- BM25, HNSW, LSH

Эмбеддинги
- word2vec, glove, ~~paragraph2vec,~~ fasttext, starspace, poincaré

Нейросетевые подходы
- LSTM, GRU, Attention, siamese network, similarity learning

Полезный NLP-софт

Предобработка текста (нормализация, токенизация)

- pymorphy2(ru), snowball stemmer(en), Stanford NLP(en)

Фреймворки

- sklearn, NLTK, gensim, spaCy

Узкоспециализированные фреймворки

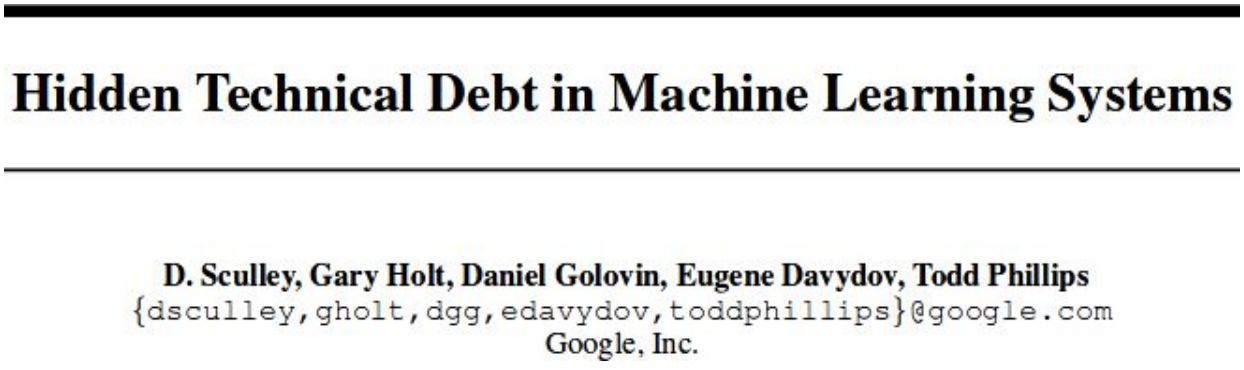
- BigARTM, Vowpal Wabbit, Fasttext, Starspace, faiss, annoy, NMSLib, lucene, sphinx, elastic

Нейросетевые фреймворки

- Pytorch, Keras

Заключение

Владения инструментами недостаточно для построения эффективных решений.
Важно не забывать о процессах.



Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.

https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

Контакты

Штех Геннадий*
@ NAUMEN
gshtekh@naumen.ru

Gennady Shtekh
shtechgen@gmail.com
t.me/sht3ch
github.com/ShT3cH

*R&D Data Usage Department Executive