СЕМАНТИЧЕСКИЙ ПОИСК ФИЛЬМОВ ПО ОПИСАНИЮ СЮЖЕТОВ

Выполнила:
Студентка 19ФиПЛ-2
Меркулова Екатерина

Руководитель:
К.ф.н. Малафеев А.Ю.

Фильмы в десяти словах

Про некоторые фильмы нам было просто интересно узнать, какими словами чаще всего описывают их пользователи. В результате получились вот такие топы, по которым вы наверняка сможете угадать, что это за кино, если смотрели его.

ЖИВОТНЫЕ
ОКЕАН
ПЛЫТЬ
ЛОДКА
ТИГР
МАЛЬЧИК
ОСТАВАТЬСЯ
ЛЕВ
ВЫЖИВАТЬ
КОРАБЛЬ

Она 2013

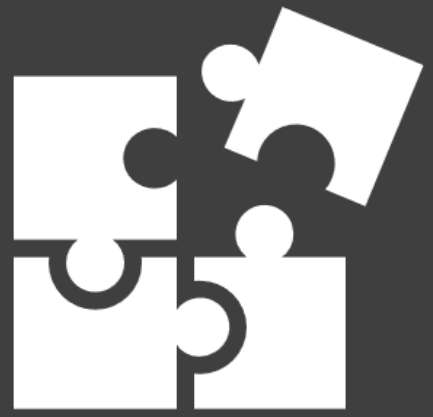
ВИРТУАЛЬНЫЙ
ОБЩАТЬСЯ
КОМПЬЮТЕР
РОБОТ
ОПЕРАЦИОННАЯ СИСТЕМА
ВЛЮБЛЯТЬСЯ
МУЖЧИНА
КОМПЬЮТЕРНЫЙ ИНТЕЛЛЕКТ
СКАРЛЕТ ЙОХАНССОН
ГОЛОС

Цель

Разработка прототипа системы семантического поиска: на входе – запрос на ЕЯ (описание сюжета/сцены), на выходе – ранжированный список фильмов

Проблема

Поиск лучшего способа представления коротких текстов для решения поставленной задачи



Существующие подходы

Методы:

- Tf-IDF [3, 4]
- Условные случайные поля [7]

Материал:

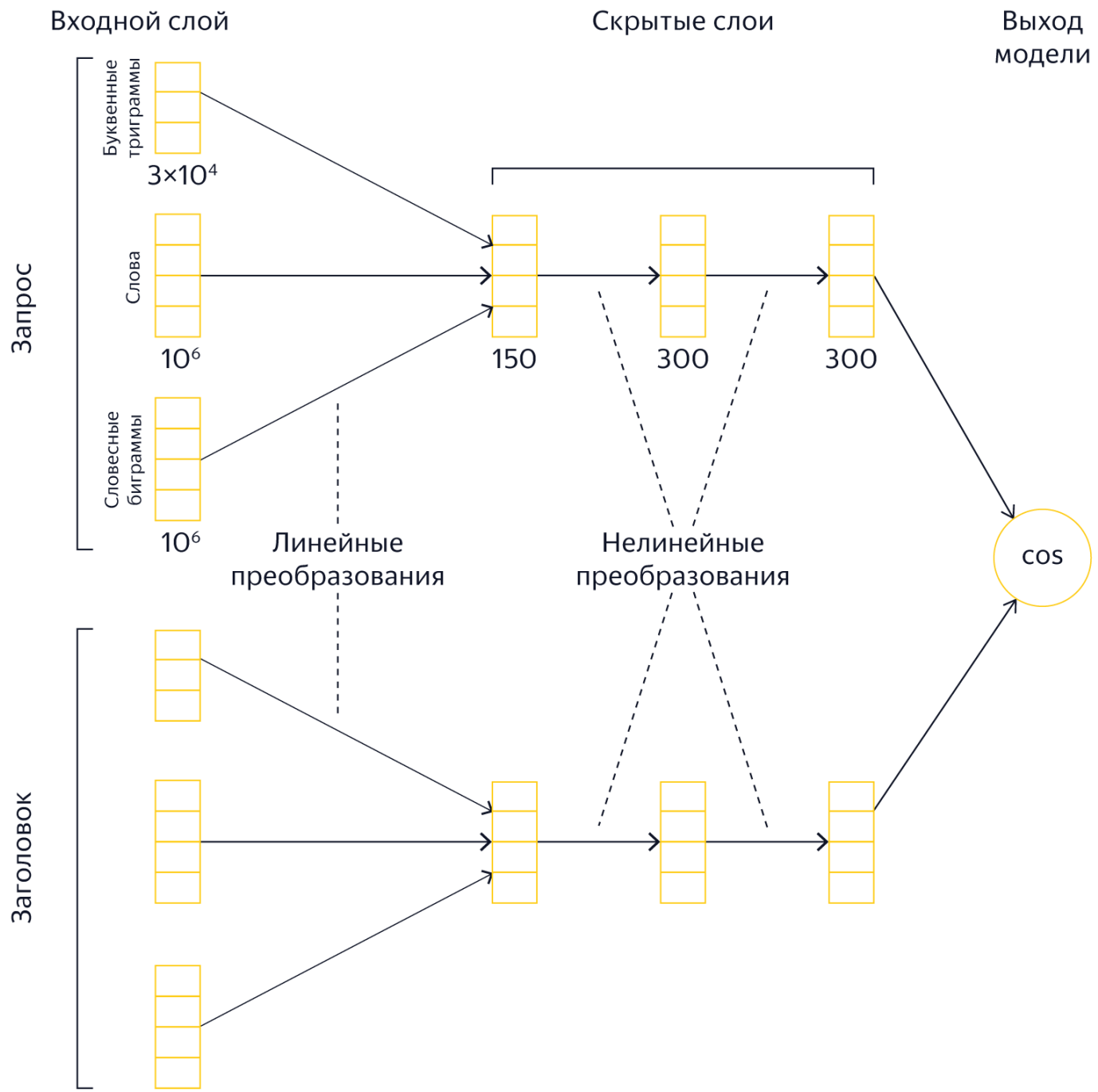
- онтологии [11, 12]
- пользовательские комментарии [4]
- метаданные и другие категории (актер, год, жанр и т.д.) [4, 7]
- ключевые слова [3]

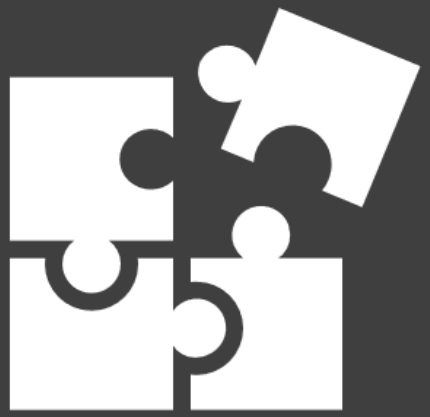
Алгоритм «Палех» от Яндекса 2016 [1]

Запрос -> Описание

Заголовок -> Текст

Интернет -> Тематический корпус





Отличия данной работы

Материал:

- тематически ограниченные датасеты: kinopoisk и Wikipedia

Методы:

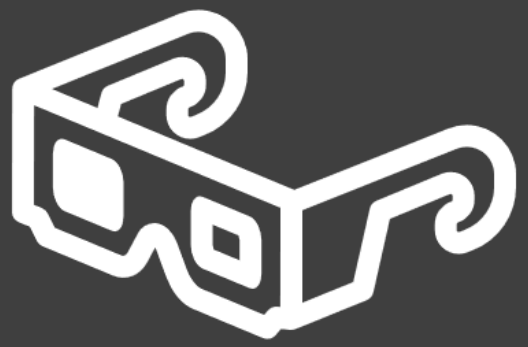
- TF-IDF
- Doc2Vec [6]

Пайплайн


Запрос
Предобработка
Векторное представление
Поиск «ближайшего»
Ранжированный список
Описание сюжетов фильмов
Предобработка
Векторное представление

Этапы

1. Сбор данных
2. Предобработка текстов
3. Обучение моделей, подбор гиперпараметров
4. Оценка качества



Сбор данных

	Кинопоиск	**Википедия**
Инструмент	httrack	wikipediaapi
Содержание	описание фильма + не более 10 рецензий	раздел «Сюжет»
Количество текстов / токенов	394 / 7 361 150	29 922 / 39 510 122
Средняя длина	2865	185
Жанровые особенности	эмоциональная, разговорная лексика, большая неинформативная часть	нейтральная общая лексика, текст более объективный

Пострадав в результате несчастного случая, богатый аристократ Филипп нанимает в помощники человека, который менее всего подходит для этой работы, – молодого жителя предместья Дрисса, только что освободившегося из тюрьмы. Несмотря на то, что Филипп прикован к инвалидному креслу, Дриссу удается привнести в размеренную жизнь аристократа дух приключений.

Рейтинг фильма

★ ★ ★ ★ ★ ★ ★ ★ ★ ★
1 2 3 4 5 6 7 8 9 10

8.807 1 162,027
IMDb: 8.50 738K

5
ТОП 250

+ Написать рецензию

Рецензии / 1+1
Intouchables, 2011

К описанию фильма »

Добавить рецензию...

Все: 543 | Положительные: 475 | Отрицательные: 36 | Процент: 90.4% | Нейтральные: 32

сортировать: по рейтингу | по дате | по имени пользователя

1 2 3 4 » »» 1—10 из 543 показывать: 10

Катя Ерч

рецензии | оценки | друзья | фильмы | звёзды

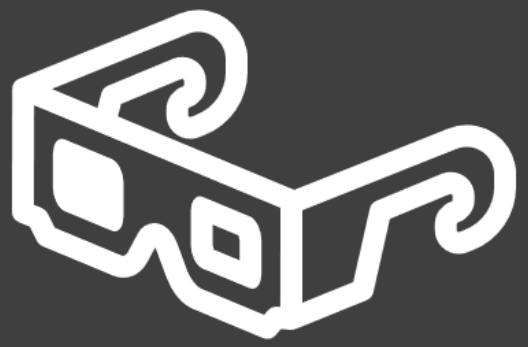
14 июля 2020 | 17:58

тип рецензии:

Дружба без условностей

Фильм рассказывает нам о двух уникальных жизнелюбах. Почему аристократ — инвалид Филипп, выбрал Дрисса, который только что вышел из тюрьмы? На протяжении всего фильма, мы можем наблюдать как столь, казалось бы разные люди, близки по духу и любви к жизни. Дрисс видел в Филиппе не только обладателя огромного состояния и шикарного дома, но и человека, который нуждается не только в медицинском уходе, но и в друге. Когда Филипп вынужден уволить Дрисса, ему не подходит никакой другой помощник. За это Филипп понял, что Дрисс искренне к нему относится. Каждому из людей хотелось бы иметь такого друга. Если вспомнить сцену, когда Дрисс танцует для Филиппа, сколько радости этот танец приносит им обоим и всем окружающим. И как профессионально был поставлен этот танец и музыка. Дрисс никогда не видел в Филиппе инвалида, это и послужило хорошим основанием для их долгой дружбы. Приятно было пересмотреть этот фильм еще раз. Как хорошо, когда люди не обращают внимания на цвет кожи. Как и Бог, который нелицеприятен.

Приятного просмотра и хорошего настроения!



Сбор данных

	Кинопоиск	**Википедия**
Инструмент	httrack	wikipediaapi
Содержание	описание фильма + не более 10 рецензий	раздел «Сюжет»
Количество текстов / токенов	394 / 7 361 150	29 922 / 39 510 122
Средняя длина	2865	185
Жанровые особенности	эмоциональная, разговорная лексика, большая неинформативная часть	нейтральная общая лексика, текст более объективный

Категория:Фильмы по алфавиту

Помощь

Материал из Википедии — свободной энциклопедии

В эту категорию **автоматически** помещаются статьи, содержащие шаблон {{Фильм}}.

Оглавление: *...в начало*

А Б В Г Д Е Ж З И К Л М Н О П Р С Т У Ф Х Ц Ч Ш Щ Э Ю Я

Абу Бар Ван Гам Дар Евл Жан Зай Иве Кан Лал Мак Най Обу Пас Рай Сан Так Уде Фак Хак Цар Чар Шап Щед Эвр Юди Ягу

Ада Без Вве Гек Дел Его Жев Зан Идо Кас Лас Мар Нат Ожи Пер Рац Сев Тва Уим Фед Хар Цви Чел Шва Щеп Эйф Юли Яко

Акт Бер Вес Геф Дец Ежо Жен Зах Ико Ким Лег Мау Ней Оку Пис Рел Сет Тер Улм Фид Хат Цен Чер Шен Щер Экт Юнг Яку

Алт Бит Вин Гно Дин Ели Жиг Зее Имп Кож Леп Мер Нет Опо Плу Рич Ско Тих Уни Фин Хим Цер Чес Шин Щит Эли Юри Ямб

Анд Бол Вла Гон Дов Епи Жир Зен Инт Кон Лин Мим Нил Орн Пор Рой Сод Топ Урб Фок Хок Циг Чиж Шма Щук Энг Юрк Яно

Апп Боя Вок Гра Дор Еро Жуа Зин Иос Кра Лов Мож Нов Осл Пре Рот Спе Три Урю Фоф Хоп Цин Чка Шта Щуч Эпо Юрь Яро

Арх Бул Вос Гру Дув Ест Жуп Зом Исм Кря Лук Мот Ном Отк Пуг Рут Стр Тум Утр Фро Хре Цна Чум Шув Щёг Эсс Юсу Яст

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Фильмы по алфавиту:

Цитаты в Викицитатнике

Подкатегории

В этой категории отображается 4 подкатегории из имеющихся 4.

*

- ► Мультфильмы по алфавиту (3764: 1 кат., 3763 с.)
- ► Телесериалы по алфавиту (3587: 1 кат., 3586 с.)
- [×] Телефильмы по алфавиту (904: 904 с.)

Т

- [×] Телефильмы СССР по алфавиту (542: 542 с.)

Страницы в категории «Фильмы по алфавиту»

Показано 200 страниц из 30 268, находящихся в данной категории.

(Предыдущая страница) (Следующая страница)

?

- ? (фильм)

0—9

- 11 сентября (фильм)
- 11-11-11
- 11:14
- 11.6
- 27 свадеб
- 27 Секунд памяти (фильм, 2019)
- 27 украденных поцелуев
- 28 дней

1+1 (фильм)

Материал из Википедии — свободной энциклопедии

[править | править код]

У этого термина существуют и другие значения, см. 1+1.

«1+1» (фр. *Intouchables — «Неприкасаемые»*) — французская трагикомедия 2011 года, основанная на реальных событиях[2][3][4][5][6][7] об успешном аристократе Филиппе, который в результате несчастного случая оказывается в инвалидном кресле и берёт себе в качестве помощника чернокожего бывшего преступника — Дрисса. Главные роли исполняют Франсуа Клюзе и Омар Си, удостоенный за эту актёрскую работу национальной премии «Сезар»[3]. Премьера во Франции прошла 2 ноября 2011 года[8]. В России фильм вышел в прокат 26 апреля 2012 под названием «1+1»[9].

В сентябре 2012 года Франция отправила «Неприкасаемые» бороться за статуэтку «Оскара» в номинации «Лучший фильм на иностранном языке»[10], но лента так и не вошла в шорт-лист. Несмотря на это, картина была удостоена номинаций на премии «Золотой глобус» и *BAFTA* в этой же категории.

Содержание [скрыть]

Сюжет [править | править код]

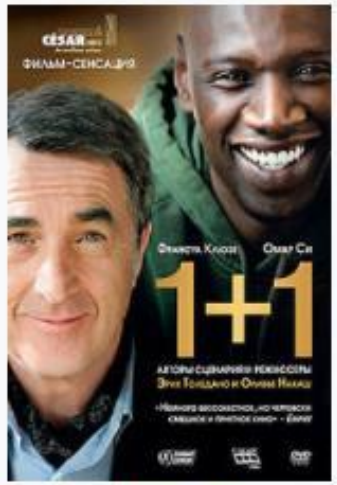
Парализованный богатый аристократ Филипп, ставший инвалидом после того, как разбился на параплане, ищет себе помощника, который должен за ним ухаживать. Одному из кандидатов, чернокожему Дриссу, не нужна работа — он хочет письменный отказ, чтобы получать пособие по безработице. Но неожиданно именно его Филипп берёт на работу. Выходцу из Сенегала с криминальными наклонностями, любителю марихуаны, женщин и ритмичной музыки совершенно неизвестны хорошие манеры — он груб, бестактен и чужд всяких условностей. Но именно его естественность и непосредственность привлекают Филиппа. Страдая от заключения внутри собственного тела, жалости окружающих и внутреннего одиночества, Филипп хочет чего-то нового. В роскошный и чопорный дворец Филиппа Дрисс приносит частичку хаоса, а в жизнь Филипа — дух приключений, спонтанности и лёгкости отношения к любым проблемам. Несмотря на сложную жизнь, Дрисс оказывается хорошим человеком. Между ним и Филиппом завязывается крепкая дружба.

Однажды Дрисс узнаёт об Элеоноре, подруге Филиппа по переписке, которая не знает, что он парализован. В результате Дрисс уговаривает Филиппа позвонить Элеоноре. Та просит Филиппа прислать его фото. Дрисс находит в альбоме два фото Филиппа: на одном видно инвалидное кресло, на другом нет. Дрисс и Филипп сначала решают отправить первое фото, но потом Филипп пугается и просит домоуправляющую Ивонну поменять фотографии. Филипп отправляется на свидание в ресторан, но в последний момент передумывает и просит Ивонну срочно увезти его, разминувшись в дверях с Элеонорой.

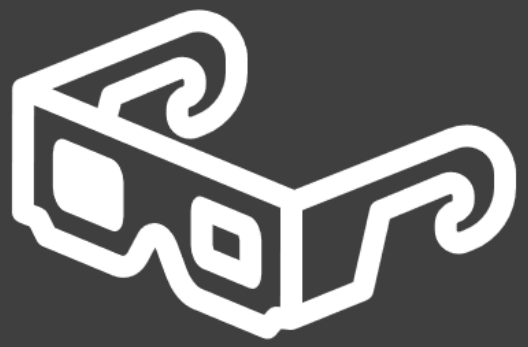
Через некоторое время по семейным обстоятельствам Дрисс вынужден покинуть Филиппа, но тот уже не может без него обходиться. Его не устраивают французские помощники с хорошими манерами и безупречными рекомендациями. Жизнь начинает казаться ему пустой, он даже задумывается о самоубийстве, но в этот момент Дрисс возвращается. Он увозит Филиппа на берег моря, и к аристократу вновь приходит радость жизни. Дрисс приводит Филиппа в кафе, где сообщает, что обедать с ним не будет: компанию Филиппу составит Элеонора.

В финале ленты сообщается о дальнейшей судьбе реальных прототипов главных героев фильма. Филипп переехал в Марокко, снова женился и обзавёлся двумя дочерьми. Абдель Селлу (Дрисс) открыл собственный бизнес, тоже женился и имеет троих детей. И по сей день они с Филиппом остаются близкими друзьями.

1+1
Intouchables



Жанр	трагикомедия
Режиссёр	Оливье Накаш Эрик Толедано
Продюсер	Николя Дюваль-Адассовски Лоран Зейтун Ян Зену
Автор сценария	Оливье Накаш Эрик Толедано
В главных ролях	Франсуа Клюзе Омар Си
Оператор	Матьё Вадпье
Композитор	Людовико Эйнауди
Кинокомпания	Quad Productions Chaocorp Gaumont TF1 Films Production
Дистрибьютор	Gaumont
Длительность	112 мин
Бюджет	$ 11 500 000
Сборы	$ 426 588 510[1]
Страна	Франция
Язык	французский
Год	2011
IMDb	ID 1675434



Сбор данных

	Кинопоиск	**Википедия**
Инструмент	httrack	wikipediaapi
Содержание	описание фильма + не более 10 рецензий	раздел «Сюжет»
Количество текстов / токенов	394 / 7 361 150	29 922 / 39 510 122
Средняя длина	2865	185
Жанровые особенности	эмоциональная, разговорная лексика, большая неинформативная часть	нейтральная общая лексика, текст более объективный

Предобработка данных

	Кинопоиск	**Википедия**
Лемматизация	pymorphy2 VS mystem (контекст)	
Стоп-слова	nltk + что, это, так, вот, быть, как, в, к, на	
Нормализация	500 < текст < 5000	
Итоговое количество текстов	337	28181
Разделение на предложения	+	-

Частоты слов

Метод преобразовывает входной текст в вектор; используется в качестве бейзлайна.

Data = ['The', 'quick', 'brown', 'fox', 'jumps', 'over', ' the', 'lazy', 'dog']

	The	quick	brown	fox	jumps	over	lazy	dog
Data	2	1	1	1	1	1	1	1

TF-IDF

преобразовывает входной текст в матрицу, значениями которой, являются слово с некоторым весом;
вес слова пропорционален частоте употребления этого слова в документе и обратно пропорционален частоте употребления слова во всех документах коллекции.

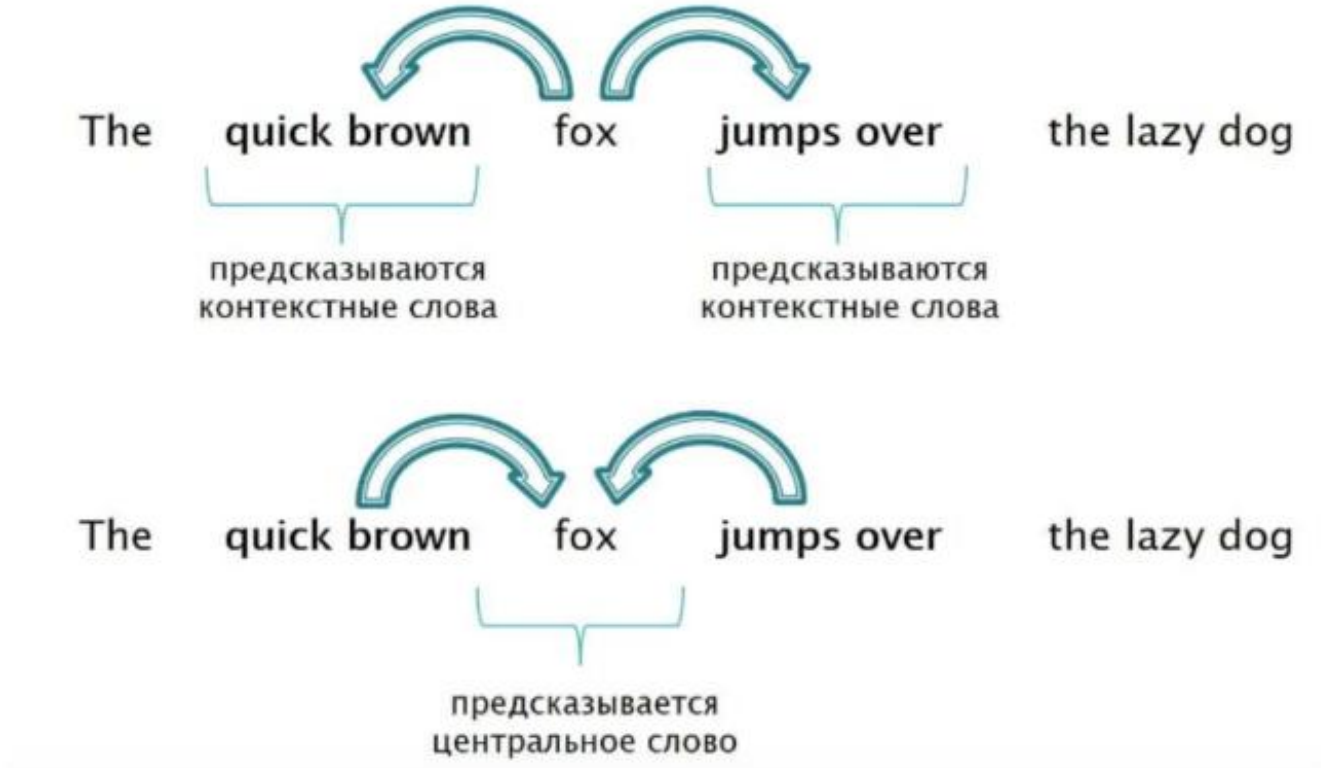
$$\mathrm{TF}\,(\mathrm{w},\mathrm{j}) = \frac{(\text{Number of times term } w \text{ appears in a document})}{(\text{Total number of terms } w \text{ in the document})}$$

$$\mathrm{IDF}\,(\mathrm{w}) = \log\frac{(\text{Total number of documents})}{(\text{Number of documents with term } w \text{ in it})}$$

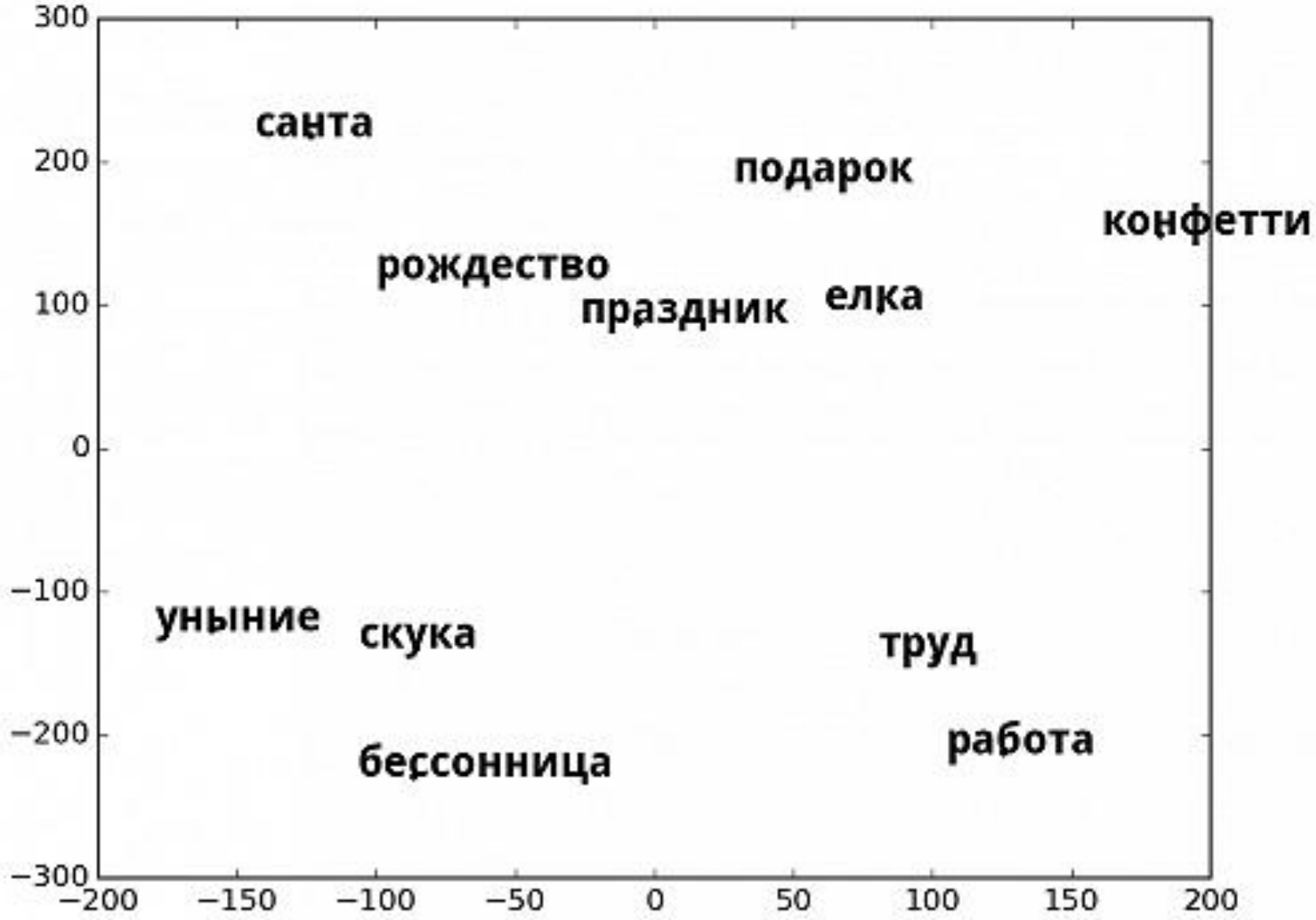
$$\mathbf{TF\text{-}IDF}(\mathbf{w},\mathbf{j}) = \mathbf{TF}(\mathbf{w},\mathbf{j}) \times \mathbf{IDF}(\mathbf{w})$$

Word2Vec

Представление текста в векторном пространстве посредством преобразования слов в числовые векторы



Семантическое пространство

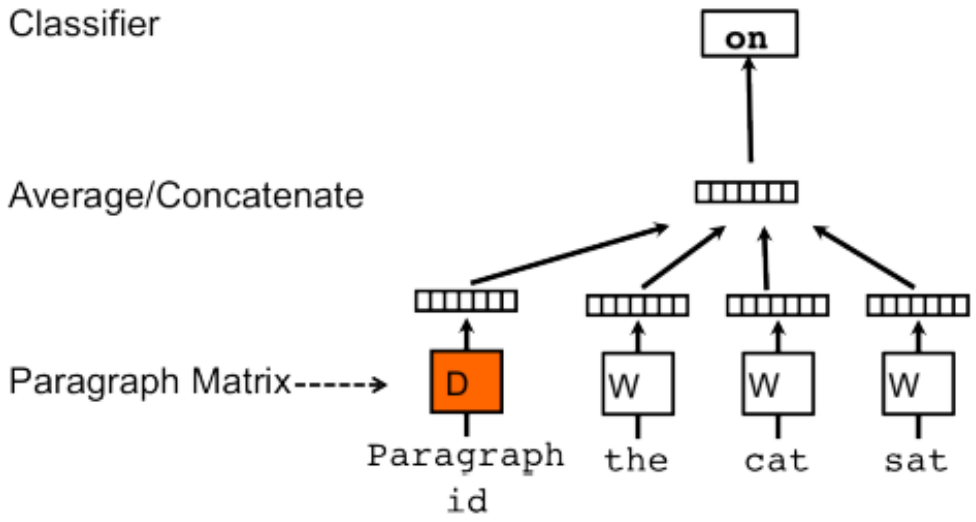


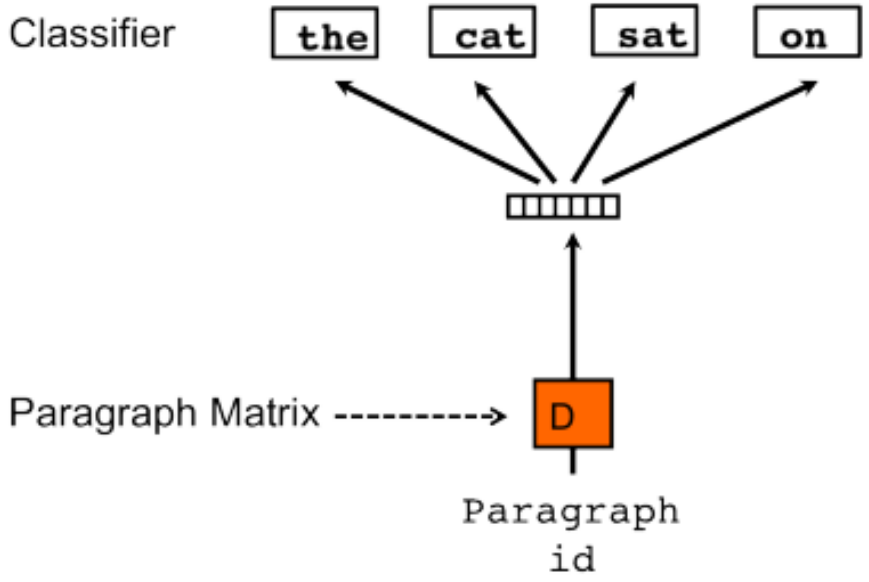
Doc2Vec

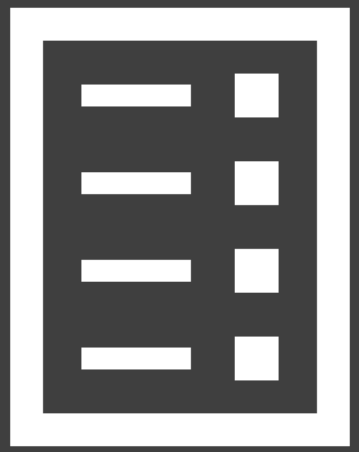
Представление коллекции текстов в векторном пространстве посредством преобразования текстов в числовые векторы

max_epochs = 20

vec_size = 50







Тестовая выборка

110 объектов по типу запрос-фильм

Семантический поиск фильма

Приветствую, уважаемый респондент! Я создаю систему семантического поиска фильма по описанию его сюжетных элементов. Буду благодарна, если Вы поможете мне создать тестовую выборку для тестирования работы модели! (Количество ответов не ограничено, можно сериалы, аниме и пр.)

* Обязательно

Представьте, что Вы когда-то давно смотрели фильм, но забыли его название и имена главных героев. Опишите, пожалуйста, в паре предложениях его сюжет (без упоминания имен героев и названия фильма). Например, фильм про мальчика, который на Рождество остался один в отеле. Он боролся с жуликами и устраивал им пакости. *

Фильм о предпринимателе, который перевернул понимание о цирке

Название фильма. Например, Один дома 2: затерянный в Нью-Йорке. *

Величайший шоумен

1+1 (фильм)

Материал из Википедии — свободной энциклопедии

[править | править код]

У этого термина существуют и другие значения, см. 1+1.

«1+1» (фр. *Intouchables — «Неприкасаемые»*) — французская трагикомедия 2011 года, основанная на реальных событиях[2][3][4][5][6][7] об успешном аристократе Филиппе, который в результате несчастного случая оказывается в инвалидном кресле и берёт себе в качестве помощника чернокожего бывшего преступника — Дрисса. Главные роли исполняют Франсуа Клюзе и Омар Си, удостоенный за эту актёрскую работу национальной премии «Сезар»[3]. Премьера во Франции прошла 2 ноября 2011 года[8]. В России фильм вышел в прокат 26 апреля 2012 под названием «1+1»[9].

В сентябре 2012 года Франция отправила «Неприкасаемые» бороться за статуэтку «Оскара» в номинации «Лучший фильм на иностранном языке»[10], но лента так и не вошла в шорт-лист. Несмотря на это, картина была удостоена номинаций на премии «Золотой глобус» и *BAFTA* в этой же категории.

Содержание [скрыть]

Сюжет [править | править код]

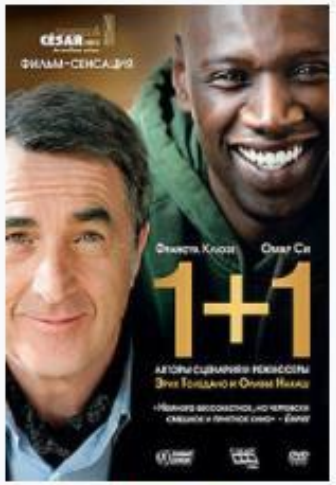
Парализованный богатый аристократ Филипп, ставший инвалидом после того, как разбился на параплане, ищет себе помощника, который должен за ним ухаживать. Одному из кандидатов, чернокожему Дриссу, не нужна работа — он хочет письменный отказ, чтобы получать пособие по безработице. Но неожиданно именно его Филипп берёт на работу. Выходцу из Сенегала с криминальными наклонностями, любителю марихуаны, женщин и ритмичной музыки совершенно неизвестны хорошие манеры — он груб, бестактен и чужд всяких условностей. Но именно его естественность и непосредственность привлекают Филиппа. Страдая от заключения внутри собственного тела, жалости окружающих и внутреннего одиночества, Филипп хочет чего-то нового. В роскошный и чопорный дворец Филиппа Дрисс приносит частичку хаоса, а в жизнь Филипа — дух приключений, спонтанности и лёгкости отношения к любым проблемам. Несмотря на сложную жизнь, Дрисс оказывается хорошим человеком. Между ним и Филиппом завязывается крепкая дружба.

Однажды Дрисс узнаёт об Элеоноре, подруге Филиппа по переписке, которая не знает, что он парализован. В результате Дрисс уговаривает Филиппа позвонить Элеоноре. Та просит Филиппа прислать его фото. Дрисс находит в альбоме два фото Филиппа: на одном видно инвалидное кресло, на другом нет. Дрисс и Филипп сначала решают отправить первое фото, но потом Филипп пугается и просит домоуправляющую Ивонну поменять фотографии. Филипп отправляется на свидание в ресторан, но в последний момент передумывает и просит Ивонну срочно увезти его, разминувшись в дверях с Элеонорой.

Через некоторое время по семейным обстоятельствам Дрисс вынужден покинуть Филиппа, но тот уже не может без него обходиться. Его не устраивают французские помощники с хорошими манерами и безупречными рекомендациями. Жизнь начинает казаться ему пустой, он даже задумывается о самоубийстве, но в этот момент Дрисс возвращается. Он увозит Филиппа на берег моря, и к аристократу вновь приходит радость жизни. Дрисс приводит Филиппа в кафе, где сообщает, что обедать с ним не будет: компанию Филиппу составит Элеонора.

В финале ленты сообщается о дальнейшей судьбе реальных прототипов главных героев фильма. Филипп переехал в Марокко, снова женился и обзавёлся двумя дочерьми. Абдель Селлу (Дрисс) открыл собственный бизнес, тоже женился и имеет троих детей. И по сей день они с Филиппом остаются близкими друзьями.

1+1
Intouchables



Жанр	трагикомедия
Режиссёр	Оливье Накаш Эрик Толедано
Продюсер	Николя Дюваль-Адассовски Лоран Зейтун Ян Зену
Автор сценария	Оливье Накаш Эрик Толедано
В главных ролях	Франсуа Клюзе Омар Си
Оператор	Матьё Вадпье
Композитор	Людовико Эйнауди
Кинокомпания	Quad Productions Chaocorp Gaumont TF1 Films Production
Дистрибьютор	Gaumont
Длительность	112 мин
Бюджет	$ 11 500 000
Сборы	$ 426 588 510[1]
Страна	Франция
Язык	французский
Год	2011
IMDb	ID 1675434

Оценка качества

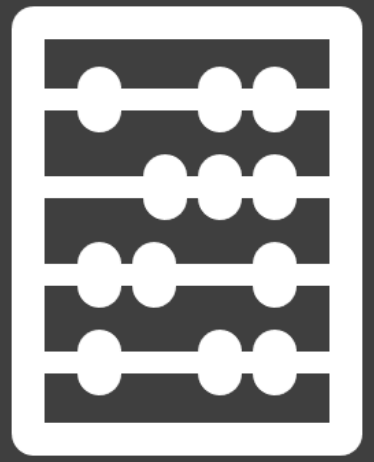
$$\mathrm{AveP} = \frac{\sum_{k=1}^{n} (P(k) \times \mathrm{rel}(k))}{\text{number of relevant documents}}$$

$$\mathrm{MAP} = \frac{\sum_{q=1}^{Q} \mathrm{AveP(q)}}{Q}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

Качество моделей на кинопоиск-датасете

preprocessing	embedding	recall	MAP
None	CountVectorizer	0.22	0.134667
stop-words, lemmatization	CountVectorizer	0.58	0.446333
split_sentence, stop-words, lemmatization	CountVectorizer	0.24	0.138056
None	TF-IDF	0.67	0.457833
stop-words, lemmatization	TF-IDF	0.78	0.565833
split_sentence, stop-words, lemmatization	TF-IDF	0.44	0.278486
split_sentence	Doc2Vec	0.31	0.198000
split_sentence, stop-words, lemmatization	Doc2Vec	0.56	0.380000
split_sentence, stop-words, lemmatization	Doc2Vec & TF-IDF	0.56	0.413000



TF-IDF – ключевые слова и имена собственные

query	precision	average_precision
Сериал про детектива и доктора	0.0	0.0
Сериал про Шерлока Холмса	0.2	1.0

Doc2Vec – имена нарицательные

query	precision	average_precision
сериал про детектива и доктора	0.2	0.333333
сериал про шерлока холмса	0.0	0.000000

Качество моделей на википедия-датасете

preprocessing	embedding	recall	MAP
stop-words, lemmatization	CountVectorizer	0.036364	0.021212
stop-words, lemmatization	TF-IDF	0.090909	0.048788
stop-words, lemmatization	Doc2Vec	0.000000	0.000000

Качество моделей на википедия-датасете 1000

preprocessing	embedding	recall	MAP
stop-words, lemmatization	CountVectorizer	0.118182	0.086667
stop-words, lemmatization	TF-IDF	0.218182	0.136667
stop-words, lemmatization	Doc2Vec	0.009091	0.003030

Перспективы

1. автоматически сгенерированная тестовая выборка из Википедии;
2. парсинг субтитров;
3. другие модели: Word2Vec, GloVe, Fasttext, BERT;
4. предобученные модели

Список литературы

1. Фильм, в котором был грунт. Исследование Яндекса и краткая история поиска по смыслу // URL: https://habr.com/ru/company/yandex/blog/464315/
2. Фильм, в котором // URL: https://yandex.ru/company/researches/2019/whatsthemovie
3. Oh, Sung-Ho, and Shin-Jae Kang. "Movie Retrieval System by Analyzing Sentimental Keyword from User's Movie Reviews." Journal of the Korea Academia-Industrial cooperation Society 14.3 (2013): 1422-1427.
4. Kim, Hyung W., et al. "Moviemine: personalized movie content search by utilizing user comments." IEEE Transactions on Consumer Electronics 58.4 (2012): 1416-1424.
5. Kutuzov A., Andreev I. Texts in, meaning out: neural language models in semantic similarity task for Russian // Proceedings of the Dialog Conference – 2015.
6. Le Q., Mikolov T. (2014), Distributed representations of sentences and documents, n Proceedings of the 31st International Conference on Machine Learning (ICML 2014), pp. 1188-1196.
7. Liu, Jingjing, et al. "A conversational movie search system based on conditional random fields." Thirteenth Annual Conference of the International Speech Communication Association. 2012.
8. Manning C. D., Raghavan P., Schütze H. Introduction to information retrieval. – Cambridge university press, 2008. – 581 p.
9. Mikolov T. et al. Distributed representations of words and phrases and their compositionality // Proceedings of NIPS 26. – 2013. – p. 3111-3119.
10. Mikolov T. et al. Efficient estimation of word representations in vector space // Proceedings of Workshop at ICLR – 2013.
11. Priya, P., and R. Rajalaxmi. "Ontology based semantic query suggestion for movie search." 2013 International Conference on Information Communication and Embedded Systems (ICICES). IEEE, 2013.
12. Wang, Ruofan, et al. "Re-ranking search results using semantic similarity." 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). Vol. 2. IEEE, 2011.