Комплинг исследование научно-популярных текстов. Данные

Анастасия Кузнецова, Анна Лапидус,
Юлия Коломенская, Ксения Самойленко

НИУ ВШЭ

13 марта 2018 г.

План

1. Задачи по тематике текстов
2. Извлечение терминов
3. Извлечение имен ученых
4. Readability

План доклада

1. Задачи по тематике текстов
2. Извлечение терминов
3. Извлечение имен ученых
4. Readability

Задачи по тематике текстов

Задача: оценить степень похожести текстов разных тематик

- Text Similarity
- Тематическое моделирование
- Кластеризация

Рубрикация

- Сформирована рубрикация текстов в соответствии с тегами
- Правила сопоставления комбинаций тегов по рубрикам

author	tags	genre	mapped_rubrics	final_rubrics
Егор Задереев	География_Экология	Статьи	Науки о земле|Науки о земле	Науки о земле
Екатерина Боровикова	Российская наука_Антропология	Статьи	Мусор|История	История
Полина Лосева	Генетика_Медицина	Статьи	Физиология человека|Физиология человека	Физиология человека

Text Similarity

- Мера Жаккара;
- Косинусная мера близости tfidf векторов;
- Близость текстов с использованием эмбеддингов.

Сложности

- Большие объемы вычислений.

Примеры

Мера Жаккара = 0.91

- Возможно, не все об этом знают, но объемы воды в пресных и соленых озерах на планете примерно одинаковы. А вот научное внимание распределено неравномерно: исследований в области пресных водоемов в несколько раз больше.
- В бассейне очистных сооружений австрийского города Клостернойбурга биологи обнаружили доселе неизвестный гигантский вирус, получивший название в честь города — клоснойвирус.

Примеры

Косинусная мера = 0.93, мера Жаккара = 0.66

- Герой сегодняшнего дня — небольшая белая таблетка, ставшая одним из рекордсменов по производству и продажам за последнее столетие (в 1949 году ее внесли в книгу рекордов Гиннеса как самое продаваемое обезболивающее). Скорее всего, она есть и в вашей аптечке, но даже если нет, то ее содержимое — аспирин, он же ацетилсалициловая кислота,— наверняка найдется в составе какого-нибудь комплексного препарата.
- Во вторник, 15 марта, в Сеуле прошла последняя встреча между компьютерной программой AlphaGo и южнокорейскими профессионалом го Ли Седолем. По итогам пяти матчей машина, работающая на технологии нейронных сетей, выиграла со счетом $4-1$. Как это было и означает ли это, что искусственный разум окончательно превзошел человеческий?

План доклада

1. Задачи по тематике текстов
2. Извлечение терминов
3. Извлечение имен ученых
4. Readability

Извлечение терминов

Вручную размечено 72 текста.

Сложности

- Размытость понятия термин (субъективность аннотатора);
- Разная концентрация терминов в текстах разных жанров.

Примеры

	left_context	term
0	экологии составляет своеобразную гармонию с	правовым нигилизмом
0	тромб закупоривает сосуд поэтому с	системой свертывания крови
0	крови шутки плохи Что у	системы свертывания крови
0	выпуске рубрики Чем нас лечат	Ишемическая болезнь сердца
0	Ксарелто который назначают для профилактики	тромбоза

Что получилось

- Выделили термины из размеченных текстов в формате KWIC (размер контекстного окна - 5 слов)
- В процессе: разметка контекстов POS тэггером и изучение контекстов с опорой на последнее слово левого контекста
- Следующий шаг: на основе анализа контекстов написать грамматику для Томита-парсера

План доклада

1. Задачи по тематике текстов
2. Извлечение терминов
3. Извлечение имен ученых
4. Readability

Разметка

Полуавтоматически размечено 167 текстов.

```
– &Рузи Талейархан!& – выдающийся ученый. Я вместе с ним и моими выдающимися американскими
специалистами &Р. Лэхи!&, &Р. Блоком!& и &К. Вестом!& работаю над «пузырьковым термоядом» начиная
с 2000 года. Наши работы опубликованы в лучших международных журналах и докладывались на многих
профессиональных научных конференциях и семинарах в США и в России. Эти результаты обсуждались в
```

Извлечение контекстов для Томиты

- Извлекли правый и левый контекст для имен ученых;
- Отсортировали в формате KWIC-выдачи;
- Разметили контексты pos-теггером (Pymorphy, TreeTagger), чтобы упростить задачу выделения паттернов;
- **Далее:** написание грамматик для Томита-парсера.

Примеры

НИИ физико-химической медицины и академик	академик	Скрябин
половине XX века станет академик	академик	Геннадий Месяц
и биогаза. По мнению академика	академика	Алфёрова
Отцы наших атомных бомб академики	академики	Зельдович

НИИ(Ncmsnn) физико-химической(Afpfsgf) медицины(Ncfsgn) и(C) академик(Ncmsny)

половине(Ncfsln) XX(Mo---d) века(Ncmsgn) станет(Vmif3s-a-p) академик(Ncmsny)

и(C) биогаза(Ncmsgn) .(SENT) По(Sp-d) мнению(Ncnsdn) академика(Ncmsgy)

Сложности

- Скорость и точность ручной разметки. Пытались как можно больше автоматизировать;
- Большое количество контекстов, сложно обобщить для написания правил.

План доклада

1. Задачи по тематике текстов
2. Извлечение терминов
3. Извлечение имен ученых
4. Readability

Readability

Readability — сумма всех элементов текстового материала, которые влияют на понимание текста, скорость прочтения и уровень интереса к материалу.

Основные методики — Обновление!

- Стандартные меры: длина слов, длина предложений, количество сложных слов, и т.д.
- Статистические метрики:различные комбинации метрик с разными коэффициентами (FRE, SMOG, Gunning fox, etc.)
- Дополнительные признаки: количество абстрактных слов, количество общих и специальных слов, тональность текста, и т.д.

Readability — Проблемы

Почему сложно оценить читабельность русских научно-популярных текстов?

- Есть готовые решения с набором статистических метрик.
  Но они для английского языка и на русском дают ошибки;
- Есть разные подходы к коэффициентам для р.я.
  в статистических метриках;
- Все метрики в основном ранжируют тексты по сложности
  от "для первоклассника"до "выпускника университета".
  Большинство русских научно-популярных текстов
  не рассчитаны на школьников младших и средних классов.
  Поэтому у них по определению будет высокий индекс
  сложности по всем этим метрикам.

Readability — Проблемы

Феминизм первой волны, ассоциируемый с движением суфражисток, — это конец XIX и начало XX века. Центральным вопросом, болевой точкой здесь является неравенство между мужчинами и женщинами, существующее на уровне законодательства. На повестке дня — реформы в области права и политики, борьба женщин за доступ к образованию, обретение прав собственности и избирательных прав. Вторая волна феминизма (конец 60-х годов XX века) предлагает новое, расширенное понимание неравенства. Активисты и исследователи полагают, что несправедливость не сводится к сфере легального, но коренится в самом общественном устройстве, в том, как мужские и женские опыты организованы социально.

Разные метрики дают этому тексту оценку в пределах 10 – 25 пунктов, что равносильно уровню "выпускник университета".

Readability — Решение

Становятся ли от этого статистические метрики бесполезными? Нет. Ведь текст может быть написан хорошо, а может плохо, и от этого меняется его сложность. Зашкаливающе длинные предложения с большим количеством сложных слов все равно будут восприниматься тяжело, и наши метрики это покажут. Но все равно необходимо введение дополнительных признаков. Что делать?

- Создать классификацию текстов по сложности;
- Создать корпус классифицированных тектов;
- Написать небольшой классификатор на основе статистических метрик, проверить, как его оценка совпадает с классами в корпусе;
- Реализовать алгоритм классификации с учетом дополнительных признаков.