Квантитативный анализ текста

Автор курса:
Кирилл Александрович Маслинский (kmaslinsky@pushdom.ru)

15 февраля 2021 г.

Задачи курса, — с одной стороны, познакомить слушателей с методологией использования текстовых данных в количественных исследованиях в области социальных и гуманитарных наук, а с другой стороны, — стимулировать и подготовить к самостоятельной аналитической работе с коллекциями текстов. Курс включает обсуждение важнейших результатов, достигнутых в области автоматической обработки языка (успехи стилометрии и дистрибутивной семантики), а также особенностей применения статистических методов к текстам (распределение языковых единиц, проблема размерности данных, статистические тесты, моделирование).

Знакомство с методологией количественных исследований текстовых данных построено по принципу разбора кейсов — нескольких современных статей, в которых проводился анализ большого объема текстовых данных. На семинарских занятиях подробно обсуждаются теоретические основания, методология и программный инструментарий, необходимые для проведения аналогичных исследований.

На практических занятиях и в ходе самостоятельной работы по курсу слушатели получают возможность применить изученные методы к предложенным в рамках курса или к собственным текстовым коллекциям (как правило, мы работаем с русскоязычными текстами). В рамках курса предполагается работа с текстовыми коллекциями с использованием статистического пакета R.

1 Содержание курса

В рамках курса рассматривается семь тем, каждая из которых предполагает взаимосвязанное обсуждение двух вопросов:

- формализованный анализ одного из аспектов текста (стиль, жанр, тематика и т.п.);
- класс задач автоматической обработки языка и необходимые для их решения методы (кластеризация и классификация документов, тематическое моделирование и т.п.).

1.1 Слово — Лексическая статистика

Частотное распределение лексики в языке. Закон Ципфа. Доля hapax legomena. Скорость роста словаря. Меры лексического разнообразия и их применимость. Открытые и закрытые классы слов. Токенизация и нормализация текста. Стемминг и лемматизация. Морфологический анализ. Части речи. Инструменты морфологического анализа для русского языка: mystem. Построение частотных списков лексики.

1.2 Стиль — Стилометрия

Извлечение ключевых слов. Метод контрастного корпуса. Отношение правдоподобия. Взаимная информация. Стилометрические методы определения авторства. Метод Delta Барроуза и его модификации. Возможности и пределы применимости стилометрических методов. Меры расстояния и кластеризация документов. Иерархическая кластеризация. Чтение кластерных дендрограмм.

Литература:

- Understanding and explaining Delta measures for authorship attribution / S. Evert [и др.] // Digital Scholarship in the Humanities. 2017. Т. 32, suppl_2. С. ii4—ii16

Дополнительная литература:

- *Stamatatos E.* A survey of modern authorship attribution methods // Journal of the American Society for information Science and Technology. 2009. Т. 60, № 3. С. 538—556
- Surveying stylometry techniques and applications / T. Neal [и др.] // ACM Computing Surveys (CSUR). 2017. Т. 50, № 6. С. 1—36

1.3 Корпус — Векторная модель текста

Векторная модель документа. Матрица терминов—документов. Взвешивание терминов: нормализация по длине документа, TF-IDF. Проблема разреженных данных. Стоп-слова.

Литература:

- *Gavin M.* Is there a text in my data? (Part 1): on counting words // Journal of Cultural Analytics. 2020. 25 янв. DOI: 10.22148/001c.11830

1.4 Жанр — Статистическое моделирование. Классификация

Задача машинного обучения. Машинное обучение с учителем. Обучающая и тестовая выборки. Задача классификации текстов. Методы классификации: логистическая регрессия. Методы снижения размерности. Регуляризация. Оценка качества классификации. Точность. Кросс-валидация.

Литература:

- *Underwood T.* The Life Cycles of Genres // Journal of Cultural Analytics. 2016. 23 мая. Т. 1, № 1. DOI: 10.22148/16.005
- *Bamman D.*, *Eisenstein J.*, *Schnoebelen T.* Gender identity and lexical variation in social media // Journal of Sociolinguistics. 2014. Т. 18, № 2. С. 135—160

1.5 Тема. Идея. Дискурс — Тематическое моделирование

Дистрибутивная гипотеза в семантике. Латентный семантический анализ. Операционализация понятия «тема» как вероятностного распределения лексики. Латентное размещение Дирихле (LDA). Процедура тематического моделирования. Препроцессинг. Сегментация текстов. Сэмплирование Гиббса. Интерпретация тем. Темы и дискурсы. Оценка качества модели.

Литература:

- *Jockers M. L.*, *Mimno D.* Significant themes in 19th-century literature // Poetics. 2013. Т. 41, № 6. С. 750—769
- *Rule A.*, *Cointet J.-P.*, *Bearman P. S.* Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014 // Proceedings of the National Academy of Sciences. 2015. Т. 112, № 35. С. 10837—10844. ISSN 0027-8424. DOI: 10.1073/pnas.1512221112

Дополнительная литература:

- *Волошинов В. Н.* Марксизм и философия языка. Москва : Лабиринт, 1993. (Бахтин под маской). ISBN 5-87604-016-9

1.6 Смысл — Дистрибутивная семантика

Пространственная метафора лексического значения. Word embeddings. Модель word2vec (SGNS). Векторная алгебра семантических отношений. Семантические вектора и социальные категории. Сфера применения и ограничения статических моделей word embedding. Социальные стереотипы и смещение (bias) в моделях.

Литература:

- *Kozlowski A. C.*, *Taddy M.*, *Evans J. A.* The geometry of culture: Analyzing the meanings of class through word embeddings // American Sociological Review. 2019. Т. 84, № 5. С. 905—949

1.7 Формула — Языковые модели

Вероятностные языковые модели. N-граммы. Цепи Маркова. Контекстное окно. Коллокации. Статистические меры для извлечения коллокаций. Синтаксический анализ. Синтаксис зависимостей: Universal dependencies. Инструменты синтаксического анализа: udpipe. Нейросетевые языковые модели. Порядок слов и механизм attention. Трансформеры (BERT, ELMo и др.) и сфера их применения для исследовательских задач.

2 Формы занятий и порядок оценивания

Курс включает лекционные, семинарские и практические занятия. Участие в семинарах предполагает чтение основной литературы к семинарским занятиям, участие в общем обсуждении, по каждому исследовательскому кейсу планируется один-два доклада слушателей. На практических занятиях слушатели тренируются в применении рассмотренных на лекциях и семинарах методов анализа текста на материале предложенных преподавателем данных.

Помимо этого каждый из слушателей выбирает собственный материал для анализа (приветствуется использование данных, связанных с исследовательской темой слушателя, однако можно использовать и любые другие опубликованные наборы данных). В ходе курса слушатели выполняют два практических домашних задания и финальный проект с использованием выбранного набора данных. Домашние задания и финальный проект представляют собой письменные отчеты, в которых приводится описание данных, аналитической задачи, методологии анализа и результатов.

Итоговая оценка формируется следующим образом:

- 40% — работа на семинарах и практических занятиях;
- 30% — домашние задания;
- 30% — финальный проект.