Компьютерные методы анализа текста: Программа курса

К. А. Маслинский

11 мая 2017 г.

1 Задачи курса

Курс «Компьютерные методы анализа текста» адресован студентам-социологам. Его главные задачи, — с одной стороны, познакомить слушателей с результатами, достигнутыми в области обработки естественного языка, а с другой, — стимулировать и подготовить их к аналитической работе с массивами текстовых данных в теоретических и прикладных социологических исследованиях.

Объем курса и его место в образовательной программе социологов, в которой отсутствуют базовые лингвистические курсы, а курсы по программированию в лучшем случае являются факультативными, не позволяют дать систематическое изложение всех разделов и методов автоматической обработки языка и компьютерной лингвистики.

В то же время, задачи курса предполагают возможность для слушателей пройти путь от теоретического обсуждения методов работы с текстом к их практическому применению. Поэтому в качестве основы для построения курса выбран принцип разбора кейсов — нескольких современных исследований, в которых проводился анализ большого объема текстовых данных. В рамках курса 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 их собственным текстовым коллекциям (как правило, мы работаем с русскоязычными текстами).

В рамках курса предполагается работа с текстовыми коллекциями с использованием статистического пакета R.

2 Построение курса

В рамках каждой темы отправной точкой для обсуждения является разбор *кейса* — современного (не старше 5 лет) опубликованного академического исследования, в рамках которого использовалась методология автоматического анализа текстовых данных. Исследования, выбранные в

качестве кейсов, могут быть выполнены в рамках любых социальных или гуманитарных дисциплин. Основными критериями отбора кейсов являются:

- доступность изложения для читателей, не имеющих специальной математической и лингвистической подготовки;
- иллюстративность — хорошая теоретическая и эмпирическая база исследования;
- относительная простота реализации и возможности широкого применения предложенной методологии.

Занятия в рамках данного курса предполагают следующий порядок работы по каждой из тем (по каждому кейсу):

1. Лекцию, в рамках которой излагаются понятия, методы и теоретические результаты, необходимые для понимания основной статьи (кейса) по данной теме.
2. Чтение и конспектирование основной статьи.
3. Обсуждение теоретических оснований, данных, методологии и результатов основной статьи на семинаре.
4. Доклады на семинарах по дополнительным темам к основной статье.
5. Применение методологии, использованной в основной статье, к предложенным в рамках курса или индивидуальным текстовым коллекциям в рамках практических занятий и самостоятельной работы по курсу.

3 Содержание курса

Курс включает четыре темы, каждая из которых предполагает взаимосвязанное обсуждение двух вопросов:

- формализованный анализ одного из аспектов текста (стиль, содержание, структура и т.п.);
- класс задач автоматической обработки языка и необходимые для их решения методы (классификация документов, анализ тональности, извлечение сущностей и т.п.).

NB: *Конкретные статьи приведены ниже для примера, перед началом курса список статей для чтения и разбора на занятиях может быть обновлен, чтобы адекватно отражать современный круг проблем и методологический инструментарий. Однако общая логика построения курса и основные темы сохранятся.*

Тема 1. Стиль — Классификация документов

Основная статья: *Koppel M.*, *Argamon S.*, *Shimoni A. R.* Automatically categorizing written texts by author gender // Literary and Linguistic Computing. 2002. т. 17, № 4. с. 401—412.

Темы для рассмотрения на лекции: Векторная модель документа. Матрица терминов—документов. Взвешивание терминов: нормализация по длине документа, TF-IDF. Проблема разреженных данных. Методы снижения размерности. Стоп-слова. Отбор значимых свойств (feature selection).

Задача машинного обучения. Машинное обучение с учителем. Обучающая и тестовая выборки. Алгоритм обучения.

Задача классификации текстов. Области применения классификации в обработке естественного языка. Оценка качества классификации. Точность. Кросс-валидация.

Понятие корпус. Лингвистическая аннотация. Иерархия языковых уровней.

Лексика. Частотный анализ текстов. Закон Ципфа. Открытые и закрытые классы слов. Морфологический анализ. Части речи. Стемминг и лемматизация. Полный и частичный синтаксический анализ. N-граммы.

Темы для докладов на семинарах:

1. Стилометрия. История дисциплины и классические результаты.
2. Алгоритмы классификации. Наивный Байес.
3. Алгоритмы классификации. Деревья принятия решений.
4. Алгоритмы классификации. Support vector machine (SVM).
5. Проблема переобучения (overfitting) и методы ее решения.

Задания для практического занятия:

Лабораторная работа № 1.

Задача: классификация текстов, оценка качества классификации, анализ наиболее значимых текстовых факторов (features), на которые опирался классификатор.

Пример преподавателя: дана коллекция текстов анекдотов на школьную тему, в части в качестве героя выступает Вовочка. Необходимо построить классификатор, выделяющий анекдоты про Вовочку среди остальных школьных анекдотов, оценить его точность и проанализировать набор наиболее значимых текстовых факторов.

Материалы:

1. Архив с текстами `data/anekdoty.zip`;
2. Скрипт `00corpus.R` — загрузка данных;
3. Скрипт `vovochka.R` — выделение анекдотов про Вовочку в коллекции, расстановка меток для классификации, удаление имени Вовочка из матрицы терминов;

4. Скрипт `01classify.R` — классификация, оценка качества и анализ факторов.

Варианты выполнения лабораторной работы:

1. Воспроизвести процедуру классификации текстов из примера преподавателя, изменив алгоритм классификации.

2. Воспроизвести процедуру классификации текстов из примера преподавателя, изменив набор свойств (features), используемых для классификации.

3. Выполнить аналогичную задачу классификации текстов на два класса, используя другие данные (и другую конкретную постановку задачи).

4. Воспроизвести процедуру классификации текстов из примера преподавателя, используя алгоритм классификации, описанный в статье [Koppel, Argamon, Shimoni, 2002].

Отчет: В отчете по результатам работы необходимо представить:

- Постановку задачи (выбранный тип задания из перечисленных выше).

- Если используются другие данные — описание данных.

- Описание методологии: всех произведенных изменений по сравнению с примером преподавателя.

- Если выбран другой алгоритм классификации — краткая характеристика алгоритма, обоснование выбора.

- Если выбран другой набор текстовых факторов (features) — краткое описание всех изменений в процедуре построения факторов. Обоснование для выбора таких факторов.

- Сравнить получившееся качество классификации в своей модели с качеством преподавателя, прокомментировать отличия по полноте, точности, каппа-статистике и другим показателям качества.

- Сравнить список наиболее значимых текстовых факторов в своей модели со списком факторов преподавателя (если использованы те же данные). Прокомментировать сходства и различия.

Сроки:
Мягкий дедлайн — 09.10.2015 Жесткий дедлайн — 26.09.2015

Тема 2. Содержание — Тематическое моделирование

Основная статья: *Jockers M. L.*, *Mimno D.* Significant themes in 19th-century literature // Poetics. 2013. т. 41, № 6. с. 750—769

Дистрибутивная гипотеза в семантике. Латентный семантический анализ. Вероятностный латентный семантический анализ (pLSA). Операционализация понятия «тема» как вероятностного распределения лексики. Латентное размещение Дирихле (LDA).

Процедура тематического моделирования. Препроцессинг. Сегментация текстов. Сэмплирование Гиббса. Интерпретация тем. Оценка качества модели.

Использование результатов тематического моделирования в задаче классификации текстов. Оценка качества классификации (продолжение). Таблица сопряженности. Точность, полнота, F-мера. Матрица неточностей. Каппа-статистика.

Темы для докладов на семинарах:

1. Обзор разновидностей тематических моделей. Twitter-LDA. Author-LDA. Диахронические модели.
2. Методы оценки качества тематических моделей. Perplexity. PMI.
3. Метрики качества отдельных тем.
4. Иерархические тематические модели. Pachinko allocation.
5. Тематическая кластеризация текстов.

Тема 3. Оценка — Анализ тональности

Основная статья: Narrative framing of consumer sentiment in online restaurant reviews / D. Jurafsky [и др.] // First Monday. 2014. т. 19, № 4

Автоматический анализ тональности текста. Извлечение мнений и оценок. Анализ отзывов как задача классификации. Словари оценочной лексики.

Извлечение характерной лексики. Метод контрастного корпуса. Отношение правдоподобия Даннинга. Критерий Манна-Уитни. Сравнение критериев для выделения лексики.

Темы для докладов на семинарах:

1. Обзор работ по анализу тональности текстов на русском языке.
2. Словарь оценочной лексики для области товаров Четверкина. Методология составления.
3. ??‘Polyanna hypothesis’.
4. Коллокации. Методы обнаружения коллокаций.
5. Сравнение методов выделения характерной лексики.

Тема 4. Структура — Извлечение сущностей

Основная статья: *Elson D. K.*, *Dames N.*, *McKeown K. R.* Extracting social networks from literary fiction // Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics. 2010. с. 138—147

Задачи ивзлечения информации (Data mining и information extraction). Извлечение и классификация именованных сущностей. Извлечение и классификация отношений. Анализ дат. Извлечение данных по шаблону.

Методы извлечения сущностей. Правила и словари. Статистические методы. Схема аннотации IOB. Извлечение сущностей как задача классификации. Извлечение сущностей как задача разметки последовательностей (Sequence labeling). Цепи Маркова. Скрытые марковские модели (HMM). Structured prediction. Conditional random fields.

Темы для докладов на семинарах:

1. Обзор работ по извлечению именованных сущностей из текстов на русском языке.
2. Регулярные выражения.
3. Tomita-парсер. Извлечение фактов с помощью контекстно-свободных грамматик.
4. Feature functions в пакете Stanford NER.
5. Задача извлечения отношений. Методы решения.
6. Алгоритм Витерби.