Программа экзамена курса “Математические методы анализа текстов” (2019)

При подготовке билета можно пользоваться любыми материалами (в том числе и электронными). Незнание любого вопроса из теоретического минимума влечёт за собой неудовлетворительную оценку.

Оценка за экзамен выставляется по 10-ти балльной шкале. Итоговая оценка выставляется по формуле, указанной на странице курса.

Основная программа экзамена

1. Основные операции предобработки текстовой коллекции: токенизация, стемминг, лемматизация, удаление стоп-слов. Выделение коллокаций с помощью мер ассоциации биграмм. Решение задачи сентимент-анализа без использования размеченных текстов.
2. Задача классификации текстов. Простейшие представления текста: bag of words и tf-idf. Hashing Trick. Модель логистической регрессии для бинарной классификации. Методы One-vs-Rest и One-vs-One для многоклассовой классификации.
3. Векторные представления слов. Гипотеза дистрибутивности. Count-based подходы для построения векторных представлений слов (SVD, Glove). Интерпретация модели Skip-gram как count-based метода.
4. Векторные представления слов. Модели skip-gram и cbow. Их модификации (hierarchical softmax и negative sampling).
5. Задача разметки последовательности. Модель линейного CRF. Нахождение оптимальной последовательности с помощью алгоритма Витерби. Обучение модели на размеченных данных.
6. Рекуррентные нейронные сети (RNN). Детали обучения RNN. Проблема взрывающихся и затухающих градиентов. Gradient clipping. LSTM.
7. Задача разметки последовательности. BIO-нотация. Разметка последовательности с помощью RNN. Модель RNN-CRF. Модификации для учёта опечаток: иерархическая RNN, посимвольная RNN.
8. Задача языкового моделирования. Биграммная языковая модель. Сглаживание Лапласа. Сглаживание через откат. Интерполяционное сглаживание. Модель шумного канала для задачи исправления опечаток.

9. Задача генерации естественного языка. Обучение языковой модели с помощью RNN, разница на этапах обучения и применения (teacher forcing). Генерация текста с помощью языковой модели. Beam search. Генерация текста с помощью модели трансформер.
10. Задача машинного перевода. Метрика качества BLEU. Модель sequence-to-sequence. Модель sequence-to-sequence с механизмом внимания.
11. Модель трансформера. Self attention, устройство кодировщика и декодировщика. Особенности обучения.
12. Использование языкового моделирования для transfer learning. Модель ELMO. Модель ULMFIT.
13. Модель BERT. Обучение модели, применение модели для разных задач. Алгоритм BPE.
14. Модификации моделей эмбеддингов для работы с опечатками (FastText, Mimick). Неглубокие модели представления предложений (skip-thoughts, infersent).
15. Задача классификации текстов. FastText классификатор. Свёрточные сети для классификации текстов. Рекуррентные сети для классификации текстов. Методы генерации выборки для классификации. Аугментация текстов.
16. Задача тематического моделирования. Тематическая модель PLSA, её обучение. Модель LDA как регуляризованная модель PLSA.
17. Задача тематического моделирования. Мультимодальная регуляризованная модель. Модификации модели для задач классификации и регрессии. Разделение тем на фоновые и предметные.
18. Задача тематической сегментации. Методы TextTiling и TopicTiling. Оценка качества методов сегментации.
19. Задача extractive суммаризации. TextRank для суммаризации. Тематическое моделирование для суммаризации. Нейросетевая модель extractive суммаризации с использованием self-supervised. Метрика ROUGE.
20. Задача синтаксического анализа. Грамматика составляющих. Грамматика зависимостей. Свойство проективности. Применение синтаксического анализа для решения практических задач.
21. Два подхода к обучению dependency-based синтаксического парсера: graph-based и transition-based, архитектуры моделей. Преимущества и недостатки подходов.
22. Вопросно-ответные системы. Два похода построения фактологических QA-систем (IR-based, KB-based), пайплайн выдачи ответа в каждой из них. Модели DrQA, BiDAF для IR-based системы.

Теоретический минимум

1. Сформулируйте (дано/найти/критерий качества), объясните зачем нужна и расскажите способы решения для каждой из следующих задач:
    a. Языковое моделирование.
    b. Разметка последовательности (пос-теггинг, распознавание именованных сущностей)
    c. Классификация (сентимент-анализа, жанровая)
    d. Тематическое моделирование
    e. Тематический поиск
    f. Тематическая сегментация
    g. Extractive суммаризация
    h. Машинный перевод
    i. Построение фактологической вопросно-ответной системы
    j. Построение чат-бота: определение интента, slot filing
2. Этапы предобработки текстов
3. Векторные представления слов: модели skip-gram и cbow
4. Алгоритм Витерби
5. Устройство RNN. Использование RNN для решения различных задач.
6. Устройство блока self-attention в трансформере
7. Модель BERT: обучение и применение
8. Модель PLSA