Программа экзамена курса “Математические методы анализа текстов” (осень 2020)

При подготовке билета можно пользоваться любыми материалами (в том числе и электронными). Незнание любого вопроса из теоретического минимума влечёт за собой неудовлетворительную оценку. Знание ответа на все вопросы из теоретического минимума не гарантирует положительную оценку :-)

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

Основная программа экзамена

1. Этапы решения NLP задачи. Основные операции предобработки текстовой коллекции: токенизация, стемминг, лемматизация, удаление стоп-слов. Выделение коллокаций при помощи меры ассоциации биграмм.
2. Этапы решения NLP задачи. Простейшие представления текста: bag of words и tf-idf. Модель логистической регрессии для бинарной и многоклассовой классификации.
3. Векторные представления слов. Гипотеза дистрибутивности. Count-based подходы для построения векторных представлений слов (SVD, Glove). Оценивание качества векторных представлений. Модификация MIMICK для работы с OOV словами.
4. Векторные представления слов. Модели Skip-gram и CBOW. Их модификации hierarchical softmax и negative sampling. Модификация FastText для работы с OOV словами.
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. Иерархическая RNN для учёта опечаток.
8. Задача машинного перевода. Оценивание качества модели перевода. Модель sequence-to-sequence. Модель sequence-to-sequence с механизмом внимания.
9. Модель трансформера. Self attention, устройство кодировщика и декодировщика. Особенности обучения трансформера.

10. Задача языкового моделирования. N-граммная языковая модель. Различные методы сглаживания модели. Алгоритм исправления опечаток при помощи модели шумного канала.
11. Задача языкового моделирования. Нейросетевые языковые модели (RNN, transformer). Модель GPT. Алгоритм BPE. N-shot learning.
12. Задача генерации естественного языка. Нейросетевые языковые модели для генерации текста. Особенности обучения и применения моделей. Гиперпараметры генерации текста (beam search, topK, topP, температура).
13. Задача переноса обучения. Языковое моделирование для переноса обучения. Модель ELMO. Модель ULMFIT.
14. Задача переноса обучения. Модель BERT. Обучение модели. Применение модели для разных задач. Модификации модели (достаточно рассказать про любые 2).
15. Задача классификации текстов. FastText классификатор. Свёрточные сети для классификации текстов. Рекуррентные сети для классификации текстов.
16. Этапы построения индустриальной ML-системы. Методы генерации выборки для классификации. Аугментация текстов. Модели doc2vec для представления документов.
17. Задача тематического моделирования. Тематическая модель ARTM, её обучение. Модель PLSA. Модель LDA, её интерпретация через регуляризацию.
18. Задача тематического моделирования. Мультимодальная регуляризованная модель. Модификации модели для задач классификации и регрессии. Разделение тем на фоновые и предметные.
19. Диалоговые системы. Виды диалоговых систем. Задачи, возникающие в диалоговых системах (intent detection, slot-filling). Retrieval-Based подход для ведения свободного диалога.
20. Вопросно-ответные системы. Два похода построения фактологических QA-систем (IR-based, KB-based), пайплайн выдачи ответа в каждой из них. Модели DrQA для IR-based системы. Применение BERT для QA.
21. Синтаксический анализ. Грамматика составляющих. Грамматика зависимостей. Свойство проективности. Применение синтаксиса в задачах сентимент-анализа и relation extraction.
22. Два подхода к обучению dependency-based синтаксического парсера: graph-based и transition-based, примеры архитектур моделей. Преимущества и недостатки подходов.
23. Информационный поиск. Модели DSSM, CDSMM. Персонализация поиска. Модель HBA.
24. Сессионные рекомендации. Модели GRU4Rec, BERT4Rec. Временные эмбеддинги в модели TASA.

25. Суммаризация текстов. Extractive подход. Алгоритм TextRank. Word Mover’s distance. Нейросетевая extractive суммаризация. Обучение контекстных эмбеддингов предложения.
26. Суммаризация текстов. Abstractive подход. Pointer-generator сеть и её модификации. Суммаризация, основанная на языковом моделировании.

Теоретический минимум

1. Сформулируйте (дано/найти/критерий качества), объясните зачем нужна и расскажите способы решения для каждой из следующих задач:
    a. Разметка последовательности (POS, NER)
    b. Классификация (сентимент-анализ, жанровая)
    c. Тематическое моделирование
    d. Машинный перевод
    e. Построение фактологической вопросно-ответной системы
    f. Построение чат-бота: определение интента, slot filing
    g. Языковое моделирование
    h. Генерация текста
    i. Информационный поиск
    j. Суммаризация
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