Вопросы к экзамену

1. Задача языкового моделирования. Биграммная языковая модель. Сглаживание Лапласа. Katz backoff. Interpolation smoothing.
2. Задача языкового моделирования. Нейросетевая вероятностная языковая модель (Bengio, 2003). Лог-билинейная и иерархическая лог-билинейная языковые модели.
3. Задача языкового моделирования. Рекуррентные нейронные сети (RNN). Языковое моделирование с помощью RNN. Генерация текста с помощью RNN. Beam search. Scheduled Sampling.
4. Задача разметки последовательности. Рекуррентные нейронные сети (RNN). Детали обучения RNN. Проблема взрывающихся и затухающих градиентов. Gradient clipping. Разметка последовательности с помощью RNN.
5. Задача разметки последовательностей. Скрытая марковская модель. Алгоритм Витерби. Обучение HMM по размеченным данным.
6. Основные операции предобработки текстовой коллекции: токенизация, стэмминг, лемматизация, удаление стоп-слов. Выделение коллокаций с помощью меры ассоциации биграмм и TextRank. Выделение ключевых слов по TF-IDF.
7. Модели признакового представления текста: Bag of words, TF-IDF. Hashing Trick. Модель логистической регрессии для бинарной классификации. Методы One-vs-Rest и One-vs-One для многоклассовой классификации.
8. Задача поискового ранжирования. Методология оценивания качества ранжирования. Метрики DCG, NDCG, pFound.
9. Задача поискового ранжирования. Три подхода к построению алгоритма ранжирования. Ranking SVM. RankNet. LambdaRank.
10. Векторные представления слов. SVD разложение для построения векторных представлений. Модель Skip-gram. Модель Skip-ram Negative Sampling.
11. Расширения модели Skip-gram: FastText, Paragraph2vec, Sent2vec, Starspace. Оценивание качества векторных представлений слов и документов.
12. Тематические модели PLSA и ARTM. Регуляризаторы для разделения тем на фоновые и предметные. Регуляризатор де-коррелирования тем.
13. Тематические модели PLSA и ARTM. Мультимодальная тематическая модель. Мультиязычные тематические модели ML-P и ML-TD.
14. Тематические модели PLSA и ARTM. Модель сети слов WNTM. Иерархические тематические модели.
15. Задача машинного перевода. Концепция зашумленного канала. Языковая модель и модель перевода. Задача выравнивания слов. Модели IBM-1, IBM-2.
16. Задача машинного перевода. Модель sequence to sequence для перевода. Механизмы внимания (attention) для улучшения перевода. Метрика BLEU.
17. Модели представления предложений: Skip-thoughts, InferSent, ELMO.
18. Вариации сверток (depthwise, lightweight, dynamic convolutions) и само-внимания (self-attention). Способы параметризации весов, вычислительная сложность методов по памяти и по времени.
19. Типы диалоговых систем. Проблемы, возникающие в task-oriented системах и методы их решения. Deep Structured Semantic Model (DSSM) для поиска ответа.
20. Задача обучения с подкреплением. Алгоритм REINFORCE. Применение для оптимизации метрик BLEU или ROUGE.

Теоретический минимум

1. Задача языкового моделирования.
2. Задача определения частей речи (POS-tagging).
3. Задача распознавания именованных сущностей (NER).
4. Задача классификации текстов.
5. Задача ранжирования.
6. Задача тематического моделирования.
7. Задача тематического (разведочного) поиска.
8. Задача машинного перевода.
9. Биграммная языковая модель.
10. Сглаживание Лапласа для языковых моделей.
11. Мягкий максимум (softmax). Иерархический мягкий максимум.
12. Рекуррентные нейронные сети.
13. Скрытая марковская модель.
14. Алгоритм Витерби.
15. Операции предобработки текстовой коллекции: токенизация, стэмминг, лемматизация, удаление стоп-слов.
16. Выделение коллокаций с помощью меры ассоциации биграмм.
17. Модель логистической регрессии для бинарной классификации.
18. Метрики качества бинарной классификации accuracy, precision, recall.
19. Алгоритм Ranking SVM.
20. Задачи близости и аналогий слов. Оценка качества векторных представлений слов.
21. Модель представлений Skip-gram.
22. Тематическая модель PLSA.
23. Подход аддитивой регуляризации тематических моделей. Примеры регуляризаторов.
24. Мультиязычное моделирование. Параллельные корпуса текстов.
25. Проблемы оценивания качества в машинном переводе, формула BLEU.
26. Нейросетевая архитектура sequence-to-sequence.
27. Механизм внимания для задачи машинного перевода.
28. Типы диалоговых систем (chit-chat и task-oriented, генеративные и ранжирующие).
29. Способы векторных представлений предложений (название и идея для любых трёх).
30. Алгоритм REINFORCE и пример его применения в NLP.