Задачи и вопросы к экзамену по курсу Юрия Лифшица "Алгоритмы для Интернета"

Вопросы внутри билетов

1. Суффиксные деревья. Нарисуйте все необходимые структуры данных для алгоритма Укконена в момент окончания седьмой фазы работы на тексте *aababbababa*.

2. Суффиксные деревья. Какое максимальное количество суффиксных стрелок может быть в суффиксном дереве для текста длины n. Какое наибольшее число суффиксных стрелок может выходить из одной вершины?

3. BWT. Как примерно будет выглядеть результат MtF(BWT(T^{100})), где T^{100} — это 100 раз повторенный стобуквенный текст T

4. BWT. Из результата преобразования BWT стерли \$, и получился текст *bbbbaaaa*. Каким мог быть исходный текст?

5. PageRank. Чему равна сумма всех PageRank'ов? Пусть $\varepsilon = 1/2$, существует ли сеть, в которой одна из вершин имеет PageRank равный $1/\sqrt{2}$?

6. PageRank. Какую вероятность мы вычисляем в модели информационного поиска? А какие вероятности мы извлекаем (эмпирически) из учебной коллекции?

7. Сложные сети. Для какого распределения степеней в конфигурационной модели в итоге получится пуассоновская модель с $p = 1/2$?

8. Сложные сети. Немного изменим модель Прайса: каждая новая вершина с вероятностью $1/2$ соединятся с вершиной максимальной степени, и с вероятностью $1/2$ соединяется со случайной вершиной. Пусть в графе 1000 000 вершин. Оцените (сверху) вероятность того что все вершины имеют степень меньше 1000.

9. Классификация текстов. Два DNF-правила дают одинаковый итоговый показатель качества на проверочной коллекции. Нужно выбрать одно из них. Каким критерием воспользуемся?

10. Классификация текстов. Пусть учебная коллекция — это документы $(1,1), (3,3), (1,3), (3,1)$. Первые два — принадлежат категории, третий и четвертый — нет. Верно ли, что обученный на этой коллекции по методу регрессии классификатор не будет делать ошибок на самих этих документах?

11. Классификация текстов. Определите вычислительную сложность метода регрессии.

12. SVM. Точки (0,0), (1,2) относятся к первой категории, а (3,4), (3,3) и (5,2) — ко второй. Постройте классификатор по методу опорных векторов (без ошибок, без всяких ядер).

13. SVM. Какое ядро для метода опорных векторов позволит строить классифицирующее правило в виде окружности?

14. Semantic Web. Постройте онтологию для составления метаописаний учебных курсов.

15. Semantic Web. Напишите “в стиле RDF” это предложение.

16. Mechanism Design. Объявленные времена перевозки: по ребру AB — 10, по BC — 12, по CD — 8, по AE — 18, по ED — 21. Мы едем из A в D. Кому какую компенсацию надо заплатить?

17. Mechanism Design. Как следует изменить правила аукциона Викри, если одна из компаний участвует в аукционе с помощью двух подставных лиц?

18. Open Problems. Что произойдет, если в формуле распространения меток параметр затухания α окажется больше единицы?

19. Open Problems. Примените метод Фитча к полному бинарному дереву с восемью листьями, если каждое ДНК состоит из одного нуклеотида и на листьях написано A, G, G, T, G, A, T, C

Дополнительные вопросы

1. Что такое модель Прайса?
2. Что такое “ядро для метода опорных векторов”?
3. Напишите систему уравнений для PageRank’ов интернет страниц.
4. Напишите систему уравнений для распространения меток по веб-графу.
5. Что делает алгоритм Фитча?
6. Что делает алгоритм Клейнберга?
7. Перечислите все распределения вероятностей на графах (модели), которые мы изучали в лекции “Структура сложных сетей”.
8. Какую целевую функцию при каких ограничениях мы минимизируем в методе опорных векторов (без ошибки)? А методе с ошибками и штрафами?
9. Напишите формулу оптимальной оплаты для задачи о поиске кратчайшего пути.
10. Примените преобразование Move-to-Front к тексту *aababababaacaca*.
11. Верно ли, что обученный на некоторой коллекции (DNF, метод регрессии) классификатор не делает ошибок на самой этой коллекции?
12. Чему может быть равна глубина вложенности тегов RDF-документов?
13. Какие стандарты лежат в основе семантического Веба?
14. В какой из двух сетей PageRank вершины A больше? Первая сеть — A ссылается на B; вторая — A ссылается на B, а B ссылается на A и на себя.
15. Как можно оценить приближение распределения заданий “по минимальному предложенному времени” к оптимальному распределению?
16. Из каких этапов состоит архивирование при помощи преобразования Берроуза-Вилера?
17. Какое количество опорных векторов требуется в двумерном случае?
18. Почему метод опорных векторов так называется?
19. Мы строили алгоритм Укконена в три приема. В чем они заключались?
20. Какие методы классификации текстов вы знаете?

Задачи на пятерку

1. Как реализовать Move-to-Front за время $O(n \log |\Sigma|)$, где Σ — используемый алфавит, а n — длина текста?

2. Докажите, что расстояние между векторами $PR_k(i), PR(i)$ экспоненциально быстро (по k) стремится к нулю

3. Пусть мы узнали, что вероятности принадлежности документов к категории равны $p_1 \geq \ldots \geq p_n$. По какому порогу надо принять решение о принадлежности, чтобы ожидание функции эффективности
$$u_{TP} \cdot \#TP + u_{TN} \cdot \#TN + u_{FP} \cdot \#FP + u_{FN} \cdot \#FN$$
было максимально (мы считаем, что $u_{TP}, u_{TN} > u_{FP}, u_{FN}$)?

4. Докажите, что гиперплоскость с максимальной шириной разделяющей полосы единственна

5. Пусть $|v| < |u|$, докажите что с вероятностью не менее $\frac{1}{2}$ для случайного вектора r выполнено $r \cdot v < r \cdot u$

6. Докажите, что система уравнений для PageRank имеет единственное решение

7. Найдите PageRank для сети из трех вершин: A ссылается на B, B ссылается на C.

8. Докажите, что предельные значения PageRank'а не зависят от начальных значений.

9. На плоскости расположено n точек. Выбирая случайную прямую и проектируя на нее точки, можно получить некоторый порядок на точках. Оцените сверху количество всех возможных порядков при проектировании.

10. Рассмотрим алгоритм Фитча, работающий на полном бинарном дереве с 128 листьями. В каждом листе записан один из четырех нуклеотидов A, G, C, T. Найдите точную верхнюю оценку количества мутаций в дереве, которое построит алгоритм Фитча.