Стырина Соня

Контрольная работа по автоматической обработке текста

Вариант 3.

Часть 1. Без компьютера

Комментарий: не требуется очень развернутого ответа, ответы должны быть краткие и «по делу». Если вопрос совпадает с вопросом из вывешенного на сайте примера теста, ответ, в котором совпадает текст ответа или пример, будет оцениваться как 0 баллов. На первую часть отводится 30 мин.

1. Дайте определение / краткую характеристику следующим терминам (приведите пример):
   1) Матрица терм*документ и обратный индекс (приведите пример)

Ответ:

матрица терм*документ: таблица, где столбец — это название текстового документа, строка — словоформы. В пересечении строки и столбца стоит 1/кол-во вхождений слова в документы, или 0 если в документе слова нет.

	text1	text2	text3
mouse	1	0	0
dog	0	1	0
leg	0	1	1
soup	1	0	1

обратный индекс: каждому документу приписывается уникальный номер, и каждому терму приписываются все номера документов, в которых он встречается.

mouse	156, 9903, 98
dog	320, 2990, 9
leg	11, 67
soup	129, 29

2. Что такое токенизация. Приведите 3 примера особых случаев при токенизации твитов для задач сентимент анализа Приведите по два примера на случаи (а) когда пробел не должен служить разделителем на токены, (б) когда дефис не должен служить разделителем на токены

Ответ:
токенизация —автоматическая обработка текста, которая разбивает текст на токены — отдельные словоформы, знаки припинания, etc.

Примеры токенизации твитов: эмодзи — для анализа тональности; тэги (#blacklivesmatter) для определения темы/тренда, к которой/которому твит имеет отношение; название юзера (@theloudduck).

(a) именованные сущности (New Orleans), идиомы (raining cats and dogs)
(b) сокращения типа **г-н**, в именованных сущностях (Anne-Marie)

3. Приведите два примера разных решений относительно выделяемых частеречных тегов в разных системах авматического морфологического анализа

4. Приведите пример 2-х правил (patch-и) в методе Эрика Брилла (метод дизамбигуации, основанный на автоматическом извлечении правил), которые можно вывести из следующего фрагмента:

   Золотой стандарт:
   The fly can fly
   Det N V V

   Первичная аннотация
   Det Verb Verb N

Ответ:
V —> N | Det __
N —> V | V __

Приведите пример миникорпуса (4 предложения), на котором одно из полученных правил увеличит количество ошибок, а не уменьшит?

Ответ: I bought forks last night. I love ripples in the water. I boiled it for two hours. Bring me the fruit.

(N —> V | V __)

5. Как в КС-грамматике можно реализовать допустимость предложений *Мальчик бежит* и *Мальчик порвал книгу* при запрете: **Мальчик бежит книгу* и **Мальчик порвал*

Ответ: сделать отдельные тэги для переходных и непереходных глаголов.

VP —> V (intrans)

VP —> V (trans) NP

Часть 2.

6. С помощью информации из НКРЯ рассчитайте, вероятность какой цепочки тегов выше для *Мой три окна*: (а) A-Pro V N или (б) V Num N (с учетом лексической вероятности)

Ответ:
(а)
APRO
на расстоянии 1 от V
на расстоянии 1 от S — **694 014 вхождений**.

мой (APRO) — **609 491 вхождение.**
три (V,imper,2p) — 127 514 вхождений.

(б)
V
на расстоянии 1 от NUM
на расстоянии 1 от S — **375 317 вхождений**

мой (мыть V,imper,sg) — **128 220 вхождений.**
три (NUM) — 181 015 вхождений.

цепочка тегов (а) встречается почти в 2 раза чаще, чем (б). Также значительно чаще разбор словоформы мой встречается как APRO, из чего следует предположить, что выше будет вероятность цепочки (а).

7. Приведите глубинное, промежуточное и поверхностное представление для словоформ татарского языка (исходя из принципа двухуровневой морфологии: символу алфавита на одном уровне соответствует только один символ алфавита на другом уровне, грамматический тег – один символ):

bala-**lar-ɤbɤz-ga –** нашим детям

täräz-**lär-ebez-gä –** нашим окнам

8. Даны четыре предложения. Постройте для них деревья НС. Извлеките из полученного корпуса грамматику. Переведите ее в нормальную форму Хомского.

Распишите применение алгоритма Кока-Янгера-Касами для разбора предложения *Такие типы стали есть в цехе*. Если Вам не хватает правил построенной Вами грамматики для разбора предложения, допишите необходимые правила.

Предложения:
Они (N) выпускают (V) разные (Adj) стали (N).
Дети (N) стали (V) есть (V).
Вася (N) бежал (V).
Пишите (V) письма (N).
Бегайте (V) по (Prep.) утрам (N).

Ответ:
S —> NP VP
S —> VP
NP —> Adj N
NP —> N
VP —> V NP
VP —> V PP
VP —> V V
VP —> V
PP —> Prep N

N —> они | стали | дети | Вася | письма | утрам | цехе
V —> выпускают | стали | есть | бежал | пишите | бегайте
Prep —> по | в
Adj —> разные | такие

Такие типы стали есть в цехе

добавить правила:
NP —> N N
NP —> Adj NP
VP —> V V PP