# 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):

|  |  |  |  |
| --- | --- | --- | --- |
|  | мама | мыть | рама |
| текст1 | 1 | 1 | 0 |
| текст2 | 1 | 0 | 0 |

Матрица терм\*документ и обратный индекс (приведите пример)  
  
Матрица соответствий слов и документов, в которой на пересечении ставится 0, если слово не встречается в документе, и 1 или количество вхождений – если встречается.  
Например:  
  
По модели обратного индекса само слово является индексом, по которому можно получать информацию о содержащих его документах по их номерам.  
Например:

мама {1,2}  
мыть {1}  
рама {}

1. 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  
     
   Разделение текста на слова и другие элементы (знаки препинания, именованные сущности etc) для дальнейшей обработки текста.  
   Особые случаи для сент. анализа:
   1. Смайлики. Нужно будет считать одним токеном: "(", ":(", "):"
   2. при этом количество скобок больше одной подряд может означать другой оттенок эмоций, так что можно отдельно приводить "((((", ":(((((" к единому токену - ":((("
   3. особым случаем можно считать хештеги. Например, нужно решить приводить ли "#лето" к токену "лето", или хештег несёт в себе дополнительное значение
2. "О М Г", названия - "Бойцовский Клуб"
3. Персонажи - "Чудо-женщина", "что-то"
4. Приведите два примера разных решений относительно выделяемых частеречных тегов в разных системах автоматического морфологического анализа
5. 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

если тег предыдущего Det: Verb -> N  
 если тег предыдущего Verb: N -> Verb

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

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

Часть 2.

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

*Мой*:

лексема *мой* - 609 492 вхождения

лексема *мыть* - 269 496 вхождений

*три*:

лексема *три* - 181 016 вхождений

лексема *тереть* - 133 761 вхождение

Вероятность выше для (а)

1. 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-eb**e**z-gä –** нашим окнам

1. 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 выпускают разные стали.*

*Дети стали есть.*

*Вася бежал.*

*Пишите письма.*

*Бегайте по утрам.*