# 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):
2. Матрица терм\*документ и обратный индекс (приведите пример)

Матрица терм-документ: матрица вида D\*T, где D – количество документов в корпусе, а T – количество уникальных термов в этих документах. Ячейки матрицы содержат встречаемость данного терма в данном документе.

Обратный индекс: мэппинг термов к спискам документов (vs. мэппинг документов к спискам содержащихся в них термов). Позволяет по терму находить документы, в которых они содержатся.

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

Токенизация – разбиение текста на единицы, которые мы считаем неделимыми (токены). На практике это чаще всего слова и приравнённые к ним категории граждан. Сентимент-анализ: 1. Отдельные сочетания знаков препинания, как токены (:((, :) и т.д.), 2. Не всегда можно использовать пробел/дефис для токенизации 3. Ссылки и т.д. как один токен?

Пробел: Аддис Абеба, У Ж А С Н Ы Й фильм

Дефис: М. С. Прохудин-Горский, :-)

1. Приведите два примера разных решений относительно выделяемых частеречных тегов в разных системах авматического морфологического анализа
2. 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

1. (?Det) V -> N [[(?X) = positive lookbehind]]
2. (? N V) N -> V

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

D1: Mankind abandons hope (N V N) -> (N V V)

D2: Cthulhu shreds reality to pieces (N V N Pp V) -> (N V V Pp V)

D3: Sanity is heresy (N V N) -> (N V V)

D4: Shub-Niggurath devours Manhattan(N V N) -> (N V V)

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

R1: (.\*? .+?ал .+?)

R2: (.\*? .+?ит$)

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)

Без информации о том, что V или Num – “три” и без учёта согласования, выше вероятность (а). При учёте лексической информации или синтаксических признаков, выше вероятность (б).

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

bala – Pl – Poss – Dat

bala – l[unrounded]r – [unrounded]b[unrounded]z-g[unrounded]

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.*

Pron = Они

N\_A = дети /вася/ письма / есть

N\_P = письма

N\_I = утрам

V\_tr = стали / бежал / выпускают

V\_imp = пишите / бегайте

Adj = разные

Pp = по

S -> NP VP

S -> NP VP NP

S -> VP NP

NP -> Pron