НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Кафедра прикладной математики и информатики

Компьютерная лингвистика. Практика.

#7, #8, #9, #10

Нижний Новгород, 2018

- ~~Clt. Python. Pandas~~
- ~~XML. JSON. Парсинг новостного сайта~~
- ~~Предобработка корпуса. Лемматизация. Дистанция редактирования~~
- ~~Обратный индекс~~
- ~~TF-IDF. Разреженные матрицы~~
- ~~Косинусное расстояние. BIRM~~
- **Анализ моделей классификации. Ошибки 1го, 2го рода, точность, полнота, F-score… (sklearn)**
- **Марковский процесс. N-граммы. Правдоподобие и перплексия. OpenCorpora**
- Тегирование и POS. HMM, TreeTagger, Spacy
- Тематическое моделирование (x2)
- Векторное представление слов, skip-gram, w2vec. Введение в нейронные сети (x2)

Практика #10.

Перплексия.

- Модель языка - *распределение слов в документах*

Перплексия коллекции D для языковой модели $p(w|d)$ (чем меньше, тем лучше):

$$\mathcal{P}(D) = \exp\left(-\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw}\ln p(w|d)\right)$$

$$n = \sum_{d\in D}\sum_{w\in d} n_{dw}$$

- Модель языка - *распределение слов в документах*

- **Перплексия** ~ правдоподобие, усредненное по всем словам и документам
- **Перплексия** - мера различности слов в тексте
- **Перплексия** - степень *ветвления* текста (сколько слов ожидается после каждого другого слова)

Контрольная работа #3.

Программа классификации и анализа точности классификации

Задача.

1. Используя вектора TF-IDF, построить классификатор новостных категорий
2. Logistic Regression, SVM, Random Forest, Gradient Boosting Trees + ансамбли и feature selection

Домашняя работа #3.

Выполнение заданий OpenCorpora

Задача.

1. Выполнить разметку именованных сущностей 20 текстов OpenCorpora

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ