НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Кафедра прикладной математики и информатики

Компьютерная лингвистика. Практика.

#11, #12, #13, #14, #15, #16

Нижний Новгород, 2018

- ~~Git. 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**
- **Векторное представление слов, skip-gram, w2vec. Введение в нейронные сети**

Практика #15.

Векторное представление слов.

- Skip-Gram

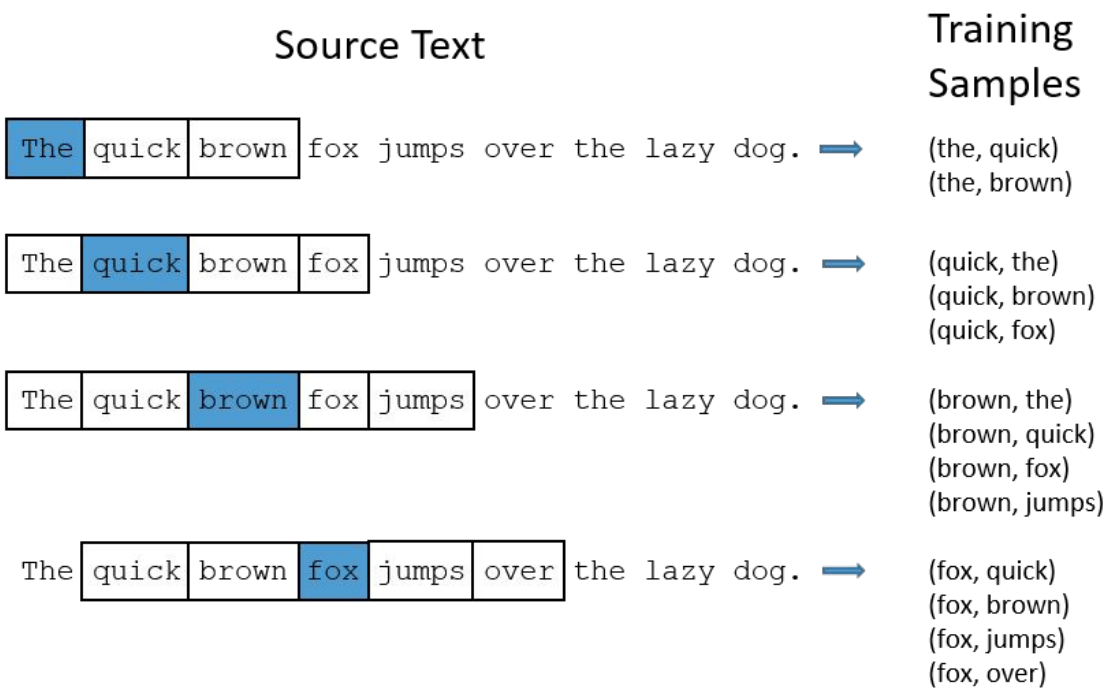

image credits:
http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

- CBOW (Continious Bag Of Words)

Center Words	Context Word
[0, 1, 0, 0, 0, 0, 0] [0, 0, 1, 0, 0, 0, 0]	[1, 0, 0, 0, 0, 0, 0]
[1, 0, 0, 0, 0, 0, 0] [0, 0, 1, 0, 0, 0, 0] [0, 0, 0, 1, 0, 0, 0]	[0, 1, 0, 0, 0, 0, 0]
[1, 0, 0, 0, 0, 0, 0] [0, 1, 0, 0, 0, 0, 0] [0, 0, 0, 1, 0, 0, 0] [0, 0, 0, 0, 1, 0, 0]	[0, 0, 1, 0, 0, 0, 0]
[0, 1, 0, 0, 0, 0, 0] [0, 0, 1, 0, 0, 0, 0] [0, 0, 0, 0, 1, 0, 0] [0, 0, 0, 0, 0, 1, 0]	[0, 0, 0, 1, 0, 0, 0]
[0, 0, 1, 0, 0, 0, 0] [0, 0, 0, 1, 0, 0, 0] [0, 0, 0, 0, 0, 1, 0] [0, 0, 0, 0, 0, 0, 1]	[0, 0, 0, 0, 1, 0, 0]
[0, 0, 0, 1, 0, 0, 0] [0, 0, 0, 0, 1, 0, 0] [0, 0, 0, 0, 0, 0, 1]	[0, 0, 0, 0, 0, 1, 0]
[0, 0, 0, 0, 1, 0, 0] [0, 0, 0, 0, 0, 1, 0]	[0, 0, 0, 0, 0, 0, 1]

original image credits:
http://www.claudiobellei.com/2018/01/07/backprop-word2vec-python/

Контрольная работа #4.

Программа классификации и анализа точности классификации

Задача.

1. Используя вектора TF-IDF и LSA, выполнить тематическое моделирование в выбранном корпусе
2. Найти оптимальное количество латентных тем и способ препроцессинга

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ