**Практическая работа №6**

Язык программирования: Python

Библиотеки: natasha, pymorphy3 (для предобработки текста), gensim

**Задание:**

Создать модель векторного представления слов

Пункты 1-5 – 5 баллов

Пункты 6-7 – 5 баллов

**Решение:**

1. В качестве исходного набора взять любой набор текстов на выбор.

<https://www.kaggle.com/datasets/joebeachcapital/restaurant-reviews>

<https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews>

<https://www.kaggle.com/datasets/andrewmvd/trip-advisor-hotel-reviews>

<https://www.kaggle.com/datasets/arushchillar/disneyland-reviews>

Все эти наборы содержат отзывы (на разные тематики), поэтому целевой текст необходимо искать в соответствующем столбце.

Их необходимо предобработать (убрать лишнее, привести к нижнему регистру, лемматизировать и тд).

Все тексты необходимо преобразовать в список предложений.

1. Для построения модели используется конструктор с большим количеством параметров:

gensim.models.word2vec.Word2Vec(sentences=None, corpus\_file=None, vector\_size=100, alpha=0.025, window=5, min\_count=5, max\_vocab\_size=None, sample=0.001, seed=1, workers=3, min\_alpha=0.0001, sg=0, hs=0, negative=5, ns\_exponent=0.75, cbow\_mean=1, hashfxn=<built-infunctionhash>, epochs=5, null\_word=0, trim\_rule=None, sorted\_vocab=1, batch\_words=10000, compute\_loss=False, callbacks=(), comment=None, max\_final\_vocab=None, shrink\_windows=True)

Например,

word2vec = Word2Vec(all\_words, min\_count = 2, vector\_size=20, sg=0)

Предложения передаются в параметр sentences.

Необходимые параметры:

* vector\_size
* window
* min\_count
* sg

Вывести список похожих слов можно с помощью команды:

word2vec.wv.most\_similar('word')

1. Используя различные значения параметров (указанных выше) построить модели и сравнить результаты.
2. Использовать визуализацию для представления результатов.

Необходимо для 3 разных слова (выбрать самостоятельно) найти наиболее контекстно близкие (для каждого слова 10 близких) и отобразить их. Для этого можно использовать метод уменьшения размерности t-SNE.

1. Ответить на вопросы:

Какие модели используются для построения векторов слов в Word2Vec и в чем их различие?

Что означают параметры, указанные в задании?

Какие задачи можно решить с помощью модели Word2vec?

Какие еще есть модели векторных представлений слов и в чем их преимущества?

1. Преобразовать тексты в векторное представление с помощью векторных представлений слов. Для получения единого вектора для всего текста можно использовать среднее значение векторов отдельных слов. Сделайте разные векторы для разных векторных представлений слов.
2. Обучить модели классификации (в качестве целевого признака выбрать рейтинг). Оценить качество моделей. Обоснуйте выбор модели классификации, сравните результаты и сделайте выводы.