- Каладзе В.А.
- Ворсунов А.А.

Формирование релевантного множества слов в текстах бытового юмора методом Word2Vec

- Исследование относится к области компьютерной лингвистики, связанной с идентификацией алгоритмов формального описания текстов естественных языков.
- Цель работы: поиск контекстных слов в текстах бытового юмора с использованием модели Word2Vec.

Принцип модели: проектирование слов в пространственные вектора и формирование схожих слов с учётом их встречаемости в контексте.

$$\frac{\text{Мужчина}}{\text{Женщина}} \sim \frac{\text{Король}}{\text{Королева}}$$

Релевантное, ожидаемое слово определяется через вероятность появления этого слова в указанном контексте

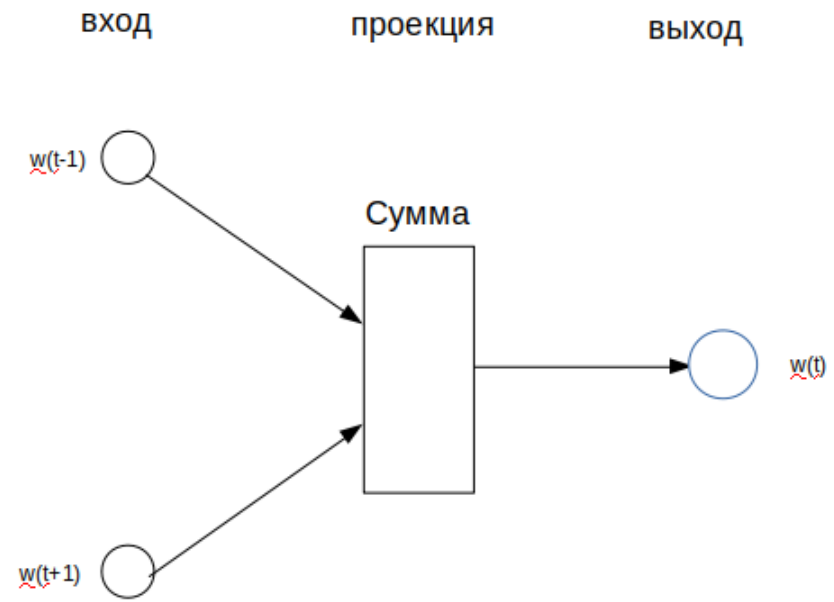
$$p(w_0 / w_i) = \frac{\exp(v_{w_i} v_{w_0}^T)}{\sum_{j=1}^{N} \exp(v_{w_i} v_{w_j}^T)}$$

Близость между двумя векторами в пространстве оценивается через косинус угла между ними

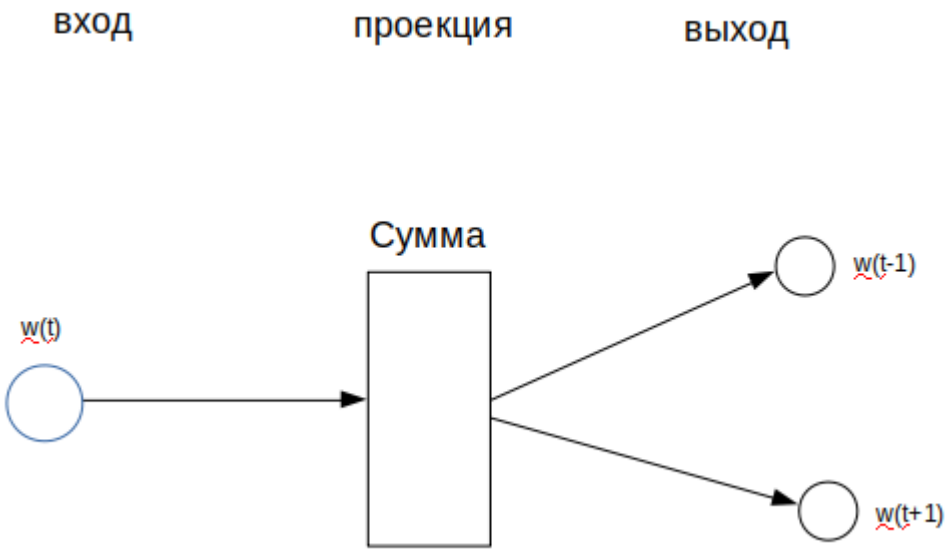
$$\cos\theta = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Численная модель Word2Vec состоит из двух процедур: CBOW и Skip-gram.

Архитектура CBOW представлена схемой



Процедура Skip-gram действует в противоход алгоритму CBOW

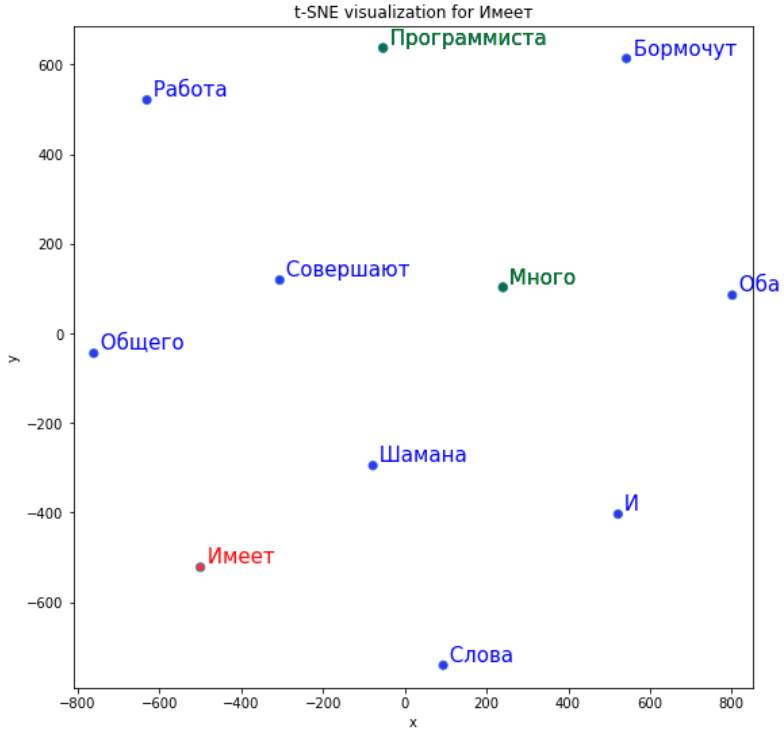


- Модель была обучена на алгоритме Skip-gram, на 150 эпохах и с 300-мерной размерностью векторов.

```
[24]: model.wv.most_similar("имеет")
```

```
[24]: [('шамана', 0.9917073845863342),
       ('много', 0.9909415245056152),
       ('общего', 0.986379206180572 5),
```

- Полученные векторы были визуализированы с помощью библиотеки TSNE:



Word2Vec является эффективным методом представления текстовой информации, способным сохранять семантику естественных языков в векторном пространстве.