Летняя международная XXXIII молодежная Школа-конференция по параллельному программированию

# Оптимизация алгоритма представления слов в виде векторов на основе word2vec

**Исполнитель проекта:** Ертуяк Айбек., 3 курс, кафедра информатики
факультета информационных технологий
КазНУ им. аль-Фараби
**Руководитель проекта:** Мансурова М.Е.
**Дата доклада:** 13.07.2019

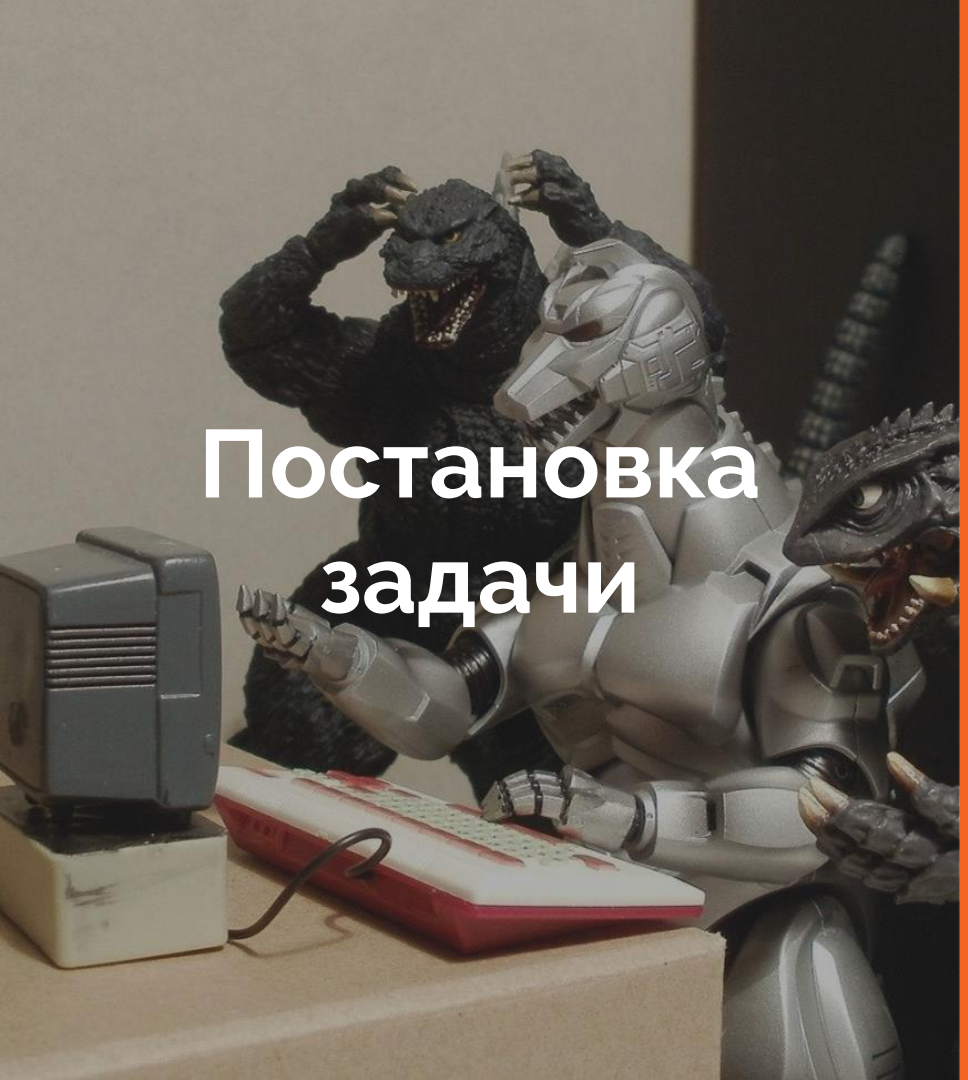
# План доклада

1. Введение
2. Постановка задачи
    - Задача №1
    - Задача №2
1. Идея решения
2. Реализация
3. Тестирование
4. Заключение

# Введение

**Word embedding** - одно из самых популярных представлений словарного запаса документов. Он способен захватывать контекст слова в документе, семантическое и синтаксическое сходство, связь с другими словами и т. д.

- ➔ **Проблема.** Из-за больших данных, обрабатываемых алгоритмом, скорость обработки текста и векторизации подвергается длительному процессу.



# Постановка задачи

## Задача №1

- Параллельная реализация препроцессинга текста. Сравнить результаты.

## Задача №2

- Параллельная реализация алгоритма word2vec. Сравнить результаты.

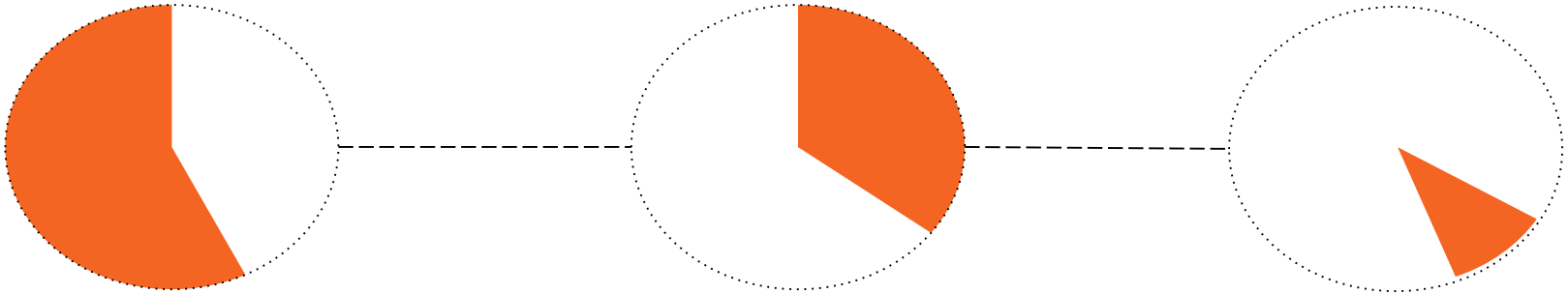
# Идея решения

OpenMP - механизм написания параллельных программ для систем с общей памятью.

# Реализация

- **Анализ алгоритма**
  Что уже сделано, а над чем ещё предстоит работать?
- **Распараллеливание алгоритма**
  Использовать openMP для алгоритма
- **Сравнение результатов**
  Сравнить последовательную реализацию и параллельную реализацию

# Алгоритм



Препроцессинг

Создание модели с помощью word2vec

Использование других алгоритмов (классификация, кластеризация и т.д.)

# Тестирование

2k новостей

~14.3k слов

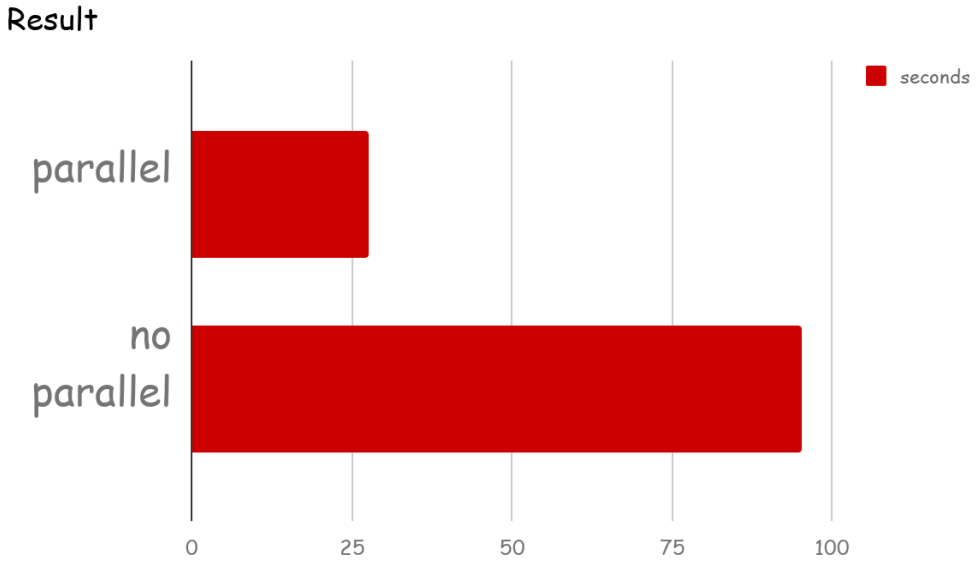
| Последовательная,sec | Параллельная, sec |
| --- | --- |
| 1.5552 | 1.51 |

| | Последовательная,sec | Параллельная, sec |
|---|---|---|
| 2k новостей<br>~14.3k слов | 1.5552 | 1.51 |
| с помощью кластера | 0.95 | 8 cores 0.68 |

| | Последовательная,sec | Параллельная, sec |
|---|---|---|
| 2k новостей ~14.3k слов | 1.5552 | 1.51 |
| с помощью кластера | 0.95 | 8 cores 0.68 |

| | Последовательная,sec | Параллельная, sec |
|---|---|---|
| 50k новостей ~34.8kk слов | 95.21 | 8 cores 27.5 |

# Заключение




- Проведен анализ алгоритма
- Изучены основы openMP и возможности стандарта для распараллеливания программ
- Проведены тесты

# СПАСИБО!