

# Отчёт по четвёртому практическому заданию курса "Автоматическое извлечение информации из текстов"

Выполнил: Алексей Грищенко, 209 группа

# Постановка задачи

1. Использовать предобученную на основе fasttext модель word2vec, рассчитанную на большом интернет-корпусе Common Crawl, для подсчета косинусных близостей слов в датасетах wordsim-similarity и wordsim-relatedness (см. ссылку в следующем разделе).
2. Посчитать корреляцию полученных значений близости человеческими оценками из датасетов с помощью корреляции Спирмена.
3. Сделать вывод на основании полученных значений корреляции

## Ресурсы используемые в практической работе

Ссылка на практическую работу:

[https://github.com/grishchenkoalexey2004/fasttext\\_word2vec\\_usage](https://github.com/grishchenkoalexey2004/fasttext_word2vec_usage)

Датасеты **wordsim-similarity** и **wordsim-relatedness** содержащие в себе пары слов и человеческие оценки их лексической близости можно найти по ссылке: <http://alfonseca.org/eng/research/wordsim353.html>

Модель word2vec обученную на основе fasttext можно найти по ссылке: <https://fasttext.cc/docs/en/english-vectors.html>

Для подсчёта косинусной близости используются функции **dot** и **norm** из библиотеки **numpy**, отвечающие соответственно за вычисление скалярного произведения и нормы векторов.

Коэффициент корреляции Спирмена считается с помощью функции **spearmanr** из модуля **numpy.stats**

## Результаты

Пары слов из файлов **wordsim\_similarity\_goldstandart.txt** и **wordsim\_relatedness\_goldstandart.txt** с посчитанной для них косинусной близостью можно найти в файлах **results\_similarity.txt** и **results\_relatedness.txt**.

Корреляция Спирмена между человеческими оценками из **wordsim\_similarity** и **wordsim\_relatedness** и косинусными расстояниями равняется 0.835 и 0.64 соответственно.

## Выводы

Между человеческими оценками и косинусными расстояниями наблюдается сильная связь (согласно шкале Чеддока). Результаты говорят о том, что из сходства контекстов, в которых два слова, с большой вероятностью следует лексическое сходство слов и наоборот.