Министерство Науки и Высшего Образования Российской Федерации

Национальный Исследовательский Институт Высшая Школа Экономики

Факультет Компьютерных Наук

Школа Анализа Данных и Искусственного Интеллекта

ДОМАШНЕЕ ЗАДАНИЕ №3

ЛЕКСИЧЕСКАЯ СЕМАНТИКА

Компьютерная лингвистика и анализ текстов

Вариант E

Студент *М.Д. Кирдин*

Преподаватель *Е.И. Большакова*

Москва, 2025г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Векторные представления слов, — также называемые эмбеддингами, — важная часть современной компьютерной лингвистики. Существует множество способов создания эмбеддингов, в их числе: *word2vec*, *fastText*. Особый интерес представляет их способность сохранять семантические связи между словами. Цель данной работы — провести исследование данного свойства у предобученных векторных представлений.

1 ХОД РАБОТЫ

Была поставлена задача провести эксперименты по кластеризации заданного набора слов русского языка, на основе векторного их представления, с целью разбиения набора на семантические/тематические классы.

Для этого был написан *Python* скрипт, использующий модель составленную из данных Википедии и НКРЯ за декабрь 2018 года представленную на ресурсе *RusVectōrēs* с идентификатором *ruwikiruscorpora_upos_skipgram_300_2_2019*. Данная модель была создана при помощи алгоритма Continuous Skipgram и содержит словарь объемом 248 978 слов, представленных эмбеддингами размерности 300. Токены в данной модели имеют структуру лемма_POS — это позволяет разрешить некоторые случаи омонимии (к примеру, Орел — город и орел — птица).

Кластеризация проводилась на 3 различных наборах слов:

- железный, серебряный, стальной, твердый, добрый, твердый, выносливый, терпеливый, алмазный, алюминиевый, верный, пластиковый, жестокий, отважный, высокомерный, надменный, деревянный, золотой, кожаный, медный, бронзовый, внимательный, раздражительный, хитрый, мудрый;

- киви, смородина, лиса, лисичка, ара, клубника, земляника, малина, черника, ежевика, огурец, облепиха, перец, яблоко, черешня, крыжовник, мандарин, мандаринка, ворон, сорока, беркут, орел, сокол, страус, эму, голубь, трясогузка, казуар;

- руль, штурвал, ручка, дверь, кабина, кокпит, шасси, трансмиссия, фара, элерон, тормоз, крыло, хвост, тяга, зеркало, колесо, шина, покрышка, кузов, оперение, крыло, планер, машина, капот, бак, закрылок, тангаж.

Первый набор был создан для исследования того как сохраняются семантические связи у имен прилагательных и состоит из слов описывающих черты характера и материал предмета. Также в нем присутствуют семантические омонимы, которые принадлежат обеим группам: «железный», «стальной», «золотой». Второй набор составлен аналогично первому. Он содержит существительные обозначающие животных и обозначающие съедобные предметы. Третий набор был создан для исследования того, как семантические связи сохраняются для узкоспециализированных слов в смежных направлениях. Он содержит слова, описывающие составные части автомобиля и самолета, а также пары гипоним-гипероним («оперение» и «элерон»).

Кластеризация слов в наборах была проведена с использованием двух алгоритмов: *K-Means* и *OPTICS*. Были использованы две меры семантической близости: косинусное расстояние и евклидово расстояние. Так как алгоритм K-Means предполагает использование только евклидовых и L_p метрик, косинусное расстояние использовалось только в алгоритме *OPTICS*.

РЕЗУЛЬТАТЫ

Метки классов полученных в результате работы программы можно увидеть в таблицах 1-3.

Таблица 1. Метки классов слов из первого набора

слово	K-Means	OPTICS (евклидово расстояние)	OPTICS (косинусное расстояние)
железный	0	0	-1
серебряный	0	0	0
стальной	0	0	0
твердый	1	1	-1
добрый	1	1	-1
выносливый	1	1	-1
терпеливый	1	1	-1
алмазный	0	0	-1
алюминиевый	0	0	0
верный	1	1	-1
пластиковый	0	0	0
жестокий	1	1	-1
отважный	1	1	-1
высокомерный	1	1	-1
надменный	1	1	-1
деревянный	0	0	0
золотой	0	0	0
кожаный	0	0	0
медный	0	0	0
бронзовый	0	0	-1
внимательный	1	1	-1
раздражительный	1	1	-1
хитрый	1	1	-1
мудрый	1	1	-1

Таблица 2. Метки классов слов из второго набора

слово	K-Means	OPTICS (евклидово расстояние)	OPTICS (косинусное расстояние)
киви	0	-1	0
смородина	0	0	0
лиса	1	-1	0
лисичка	0	-1	0
ара	1	-1	0
клубника	0	0	0
земляника	0	0	0
малина	0	0	0
черника	0	0	0
ежевика	0	0	0
огурец	0	0	0
облепиха	0	0	0
перец	0	0	0
яблоко	0	0	0
черешня	0	0	0
крыжовник	0	0	0
мандарин	0	-1	0
мандаринка	1	-1	0
ворон	1	-1	0
сорока	1	-1	0
беркут	1	-1	0
орел	1	-1	0
сокол	1	-1	0
страус	1	-1	0
эму	1	-1	0
голубь	1	-1	0
трясогузка	1	-1	0
казуар	1	-1	0

Таблица 3. Метки классов слов из третьего набора

слово	K-Means	OPTICS (евклидово расстояние)	OPTICS (косинусное расстояние)
руль	0	0	0
штурвал	0	0	0
ручка	1	-1	0
дверь	1	-1	0
кабина	0	0	0
кокпит	0	0	0
шасси	0	0	0
трансмиссия	0	0	0
фара	0	0	0
элерон	1	0	0
тормоз	0	0	0
крыло	1	-1	0
хвост	1	-1	0
тяга	0	-1	0
зеркало	1	-1	0
колесо	0	0	0
шина	0	0	0
покрышка	0	0	0
кузов	0	0	0
оперение	1	0	0
планер	1	0	0
машина	0	0	0
капот	0	0	0
бак	1	0	0
закрылок	1	0	0
тангаж	1	0	0

Согласно таблице 1, на первом наборе слов алгоритмы использующие евклидово расстояние дают одни и те же метки классов, однако использование косинусного расстояния дает иной результат. Слова «алмазный» и «железный» попадают в один кластер со словами обозначающими черты характера, а не материал предмета. Так же представляет интерес тот факт, что слова «железный» и «стальной» при кластеризации с использованием евклидова расстояния определяется в группу слов, обозначающими материалы, а не черты характера. Это может быть вызвано тем, что корпус на котором была обучена имеет преимущественно научный стиль.

На втором наборе слов, согласно таблице 2, алгоритм *K-Means* дал наилучшие результаты, хотя стоит отметить, что он отнес слово «киви» в группу слов, обозначающих предметы пищи, а не животных. Алгоритм *OPTICS* с использованием евклидова расстояния в свою очередь отнес слова «лиса» и «лисичка», а также «мандарин» и «мандаринка» в один класс. При использовании косинусного расстояния алгоритм определил все слова принадлежащими к одному и тому же классу.

На третьем наборе слов каждый алгоритм кластеризации допустил хотя бы одну ошибку: *K-Means* определил термины связанные с авиацией такие как «кокпит» или «шасси» в ту же группу, что и составляющие автомобиля, алгоритм *OPTICS* при использовании косинусного расстояния не смог провести кластеризацию, а при использовании евклидова расстояния он определил гипоним и соответствующий ему гипероним в различные классы («оперение» и «элероны») и авиационные понятия такие как «шасси», «закрылок» и «тангаж» в один кластер с автомобильными терминами. Описанные явления могли быть вызваны недостаточным количеством данных в корпусе на котором обучалась модель, что часто приводит к некачественным эмбеддингам.

ЗАКЛЮЧЕНИЕ

Подытожив, хотелось бы отметить, что по результатам экспериментов наилучшим образом показал себя алгоритм *K-Means*, использующий евклидово расстояние для измерения семантической близости слов. Стоит отметить, что у всех рассмотренных алгоритмов качество кластеризации узконаправленных терминов было ниже, чем более общих групп слов. Это может быть вызвано недостаточным числом употреблений данных слов в корпусе, на котором была обучена модель и, следовательно, более низким качеством полученных эмбеддингов.