Доклад по домешнму заданию на тему: “Лексическая Семантика”

Подготовил: Кирдин М.Д.

Преподаватель: Большакова Е.И.

Магистратура “Науки о Данных” ФКН ВШЭ

Москва, 2025

Содержание

- Цель работы
- Используемая модель
- Методы
  - Описание данных
  - Алгоритм *K-Means;*
  - Алгоритм *OPTICS;*
- Результаты
- Заключение

Цель работы

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.

Используемая модель

- Идентификатор: `ruwikiruscorpora_upos_skipgram_300_2_2019`, модель взята с ресурса RusVectōrēs.
- Корпус: данные Википедии и НКРЯ за декабрь 2018 года, 788 млн слов.
- Алгоритм: *Continuous Skipgram*.
- Словарь: 248 978 слов.
- Размерность эмбеддингов: 300.

Методы

Для исследования способности эмбеддингов сохранять семантические сходства между словами была проведена кластеризация эмбеддингов для нескольких наборов слов. Были использованы алгоритмы *K-Means* и *OPTICS*. В качестве метрик расстояния между эмбедингами использовались косинусное и евклидово расстояния.

Описание данных

Исследование проводилось на трех наборах слов:

- Первый набор слов был направлен на исследование того как сохраняются семантические связи между **именами прилагательными** 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 слов составлен аналигочно первому и содержит существительные обозначающие животных и существительные обозначающие съедобные предметы.
- Третий набор был направлен на исследование того как сохраняются связи между узконаправленными терминами в смежных напревлениях. Он содержит слова, описывающие составные части автомобиля и самолета, а также **пары гипоним-гипероним** («оперение» и «элерон»).

Алгоритм *K-Means*

В работе используется алгоритм, предоставленный библиотекой *scikit-learn*. Значения входных параметров:

- n_clusters: 2 — количество кластеров(все наборы слов составлены так, что подразумевается наличие лишь двух кластеров);
- init: “k-means++” — метод задания начальных положений центроид. Данный вариант подразумевает выбор точек данных центроидами учетом их вклада в инерцию.

Алгоритм *OPTICS*

Аналогично *K-Means*, в работе использовался *OPTICS* предоставляемый библиотекой scikit-learn. Были выбраны следующие значения параметров:

- metric: “cosine” и “euclidean” — евклидово и косинусное расстояния;
- для остальных параметров взяты значения по умолчанию.

Результаты

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

слово	K-Means	OPTICS (евклидово расстояние)	OPTICS (косинусное расстояние)
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

Результаты

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

слово	K-Means	OPTICS (евклидово расстояние)	OPTICS (косинусное расстояние)
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

Результаты

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

слово	K-Means	OPTICS (евклидово расстояние)	OPTICS (косинусное расстояние)
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

Выводы

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.