Реализация семантического поиска на корпусе новостных документов

Курсовая работа
студента 351 группы Янченко В. А.

Саратовский государственный университет
им. Н. Г. Чернышевского

Кафедра математической кибернетики
и компьютерных наук

Научный руководитель: доцент, к. ф.-м. н. С. В. Папшев

2025г.

Полнотекстовый поиск не улавливает смысловую близость слов, что снижает точность при нечетких запросах. Семантический поиск помогает, но универсальные эмбеддинги не отражают особенности конкретной области, а расширение запроса через общеязыковой тезаурус добавляет слишком общие слова, снижая релевантность результатов.

Цель работы: разработать систему семантического поиска по корпусу новостных документов с использованием эмбеддингов.

Были взяты новости СГУ за период с 2007 по 2022 года.

News_Id	News_Link	News_Date	News_Title	News_Text	News_Tokens	NumWords	NumUnicTokens
1	https://www.sgu.ru	10.05.2022	Студенты и сотруд	9 мая состоялась об	май состояться общ	207	152
2	https://www.sgu.ru	09.05.2022	В.В. Володин поздр	Председатель Госу	председатель госу,	48	41
3	https://www.sgu.ru	09.05.2022	Д.А. Медведев поз	Заместитель предс	заместитель предс	79	74
4	https://www.sgu.ru	09.05.2022	Ректор А.Н. Чумаче	Уважаемые ветера	уважаемый ветера	129	98

Рис.: Пример представления новостей

Обучены модели:

1. Word2Vec
2. GloVe

Параметры моделей:

- размер получаемых эмбеддингов — 300;
- размер контекстного окна — 5;
- минимальная частота встречаемости слова — 5.

Таблица: Синонимы для 5 случайных слов для моделей GloVe и Word2Vec

Word	GloVe	Word2Vec
баян	аккордеон	флейта, балалайка, аккордеон, саксофон, ария, барабан, гитара, бас, фортепиано, укулела
голод	холод	холод, вражеский, танк, убийство, отважный, наносить, жуткий, фашист
орфография	пунктуация	-
внимательность	сосредоточенность	быстрота, коммуникабельность
альпинизм	скалолазание	скалолазание

Таблица: Сравнительная характеристика графов моделей

Характеристика	GloVe	Word2Vec
Кол-во вершин	18402	18401
Кол-во ребер	2793	45709
Кол-во неизолированных вершин	2183	7533
Кол-во компонент (без учёта изолированных вершин)	688	360
Кол-во сообществ	734	379
Модулярность	0.618	0.612

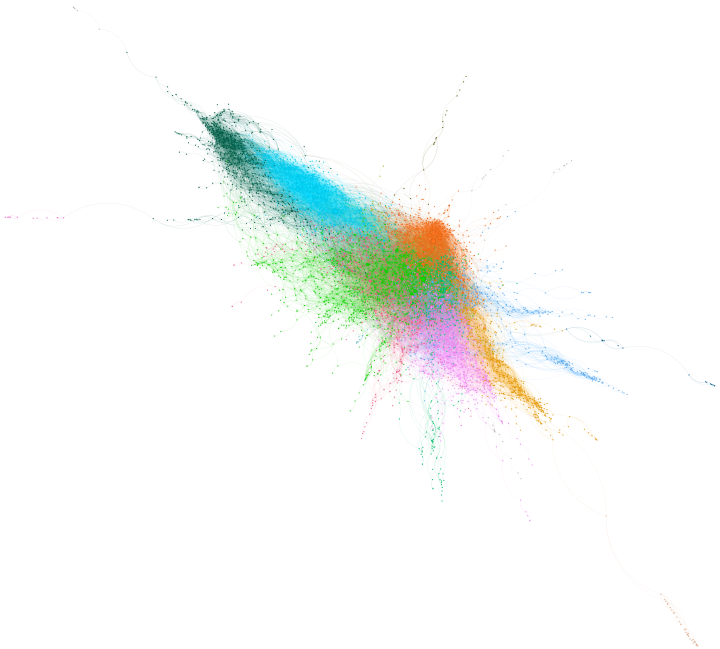

Рис.: а) Word2Vec

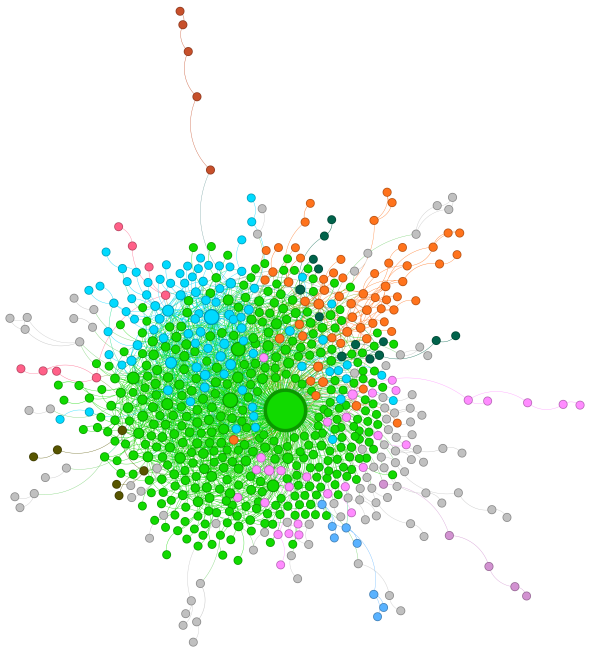

Рис.: б) GloVe

Рис.: Графы, полученные моделями Word2Vec и GloVe

Модель `ruscorpora_upos_skipgram_300_5_2018` и обученные модели Word2Vec и GloVe расширены по следующему алгоритму:

1. если эмбеддинг слова есть в обоих моделях, то используется среднее арифметическое векторов из обеих моделей;
2. если эмбеддинг слова есть только в одной из моделей, в итоговой модели используется вектор из той модели, в которой он имеется.

Эмбеддинг документа представляется как среднее эмбеддингов всех слов в тексте:

$$\bar{d} = \frac{1}{n} \sum_{i=1}^{n} \bar{w}_i$$

Алгоритм семантического поиска следующий:

1. **Предварительная обработка запроса**
2. **Получение эмбеддинга запроса**
3. **Поиск релевантных документов**
4. **Вывод результата**

В качестве запроса был взят заголовок одной из новостей: *«В.В. Володин поздравил СГУ с праздником Победы»*.

1. Предварительная обработка: ['поздравить', 'сгу', 'праздник', 'победа']
2. Эмбеддинг запроса: [[0.1731391 0.4556614 0.21552482 0.66046524, ...]]
3. ID 5 ближайших документов: [22160, 2, 29536, 29551, 24466]
4. Полученные заголовки из ID новостей:
   - В.В. Володин поздравил СГУ с праздником Победы,
   - СГУ приглашает саратовцев в парк Победы,
   - Н.В. Панков поздравляет коллектив СГУ с Новым годом,
   - В.В. Володин поздравляет коллектив СГУ с днём Великой Победы,
   - В СГУ пришли поздравления с майскими праздниками.

Пример 1: Запрос: *Вклад ученых*

1. Ученые СГУ получили президентские гранты
2. В.С. Анищенко присвоено звание «Основатель научной школы»
3. Продолжается приём заявок на соискание премии Президента в области науки и инноваций
4. Учёный СГУ отмечен медалью ордена «За заслуги перед Отечеством» II степени
5. Фонд «БАЗИС» открыл приём заявок на конкурсы индивидуальных исследовательских грантов

Пример 2: Запрос: *Бессмертный полк*

1. В СГУ открыт студенческий штаб акции «Бессмертный полк»
2. Сегодня в ИФиЖ состоится «Праздник белых журавлей»
3. Студенты и сотрудники СГУ стали участниками мероприятий ко Дню Победы
4. Сотрудники и студенты могут увековечить память защитников блокадного Ленинграда
5. Министерство молодёжной политики предлагает сделать 70 добрых дел

Пример 3: Запрос: *Чемпионат мира по программированию*

1. СГУ участвует в проведении финала чемпионата ACM-ICPC 2013
2. Студент ИФКиС стал чемпионом России по кикбоксингу
3. Программисты СГУ отправились на финал Чемпионата мира
4. Профессор В.Н. Чинилов стал призёром Чемпионата России
5. В Саратове пройдёт Неделя информатики

Для анализа создана тестовая выборка из пар: запрос — идентификатор новости.
Запросы генерировала модель `microsoft/Phi-4-mini-instruct`.
Пример получаемого ответа:

1. Саратовский конкурс «Сияние талантов» 2015,
2. Саратовский конкурс, вокалистка, поездка в Пекин,
3. Конкурс «Сияние талантов», Саратов, награда в Пекине.

Объём выборки — 3343.

Сравнительный анализ моделей в рамках задачи семантического поиска

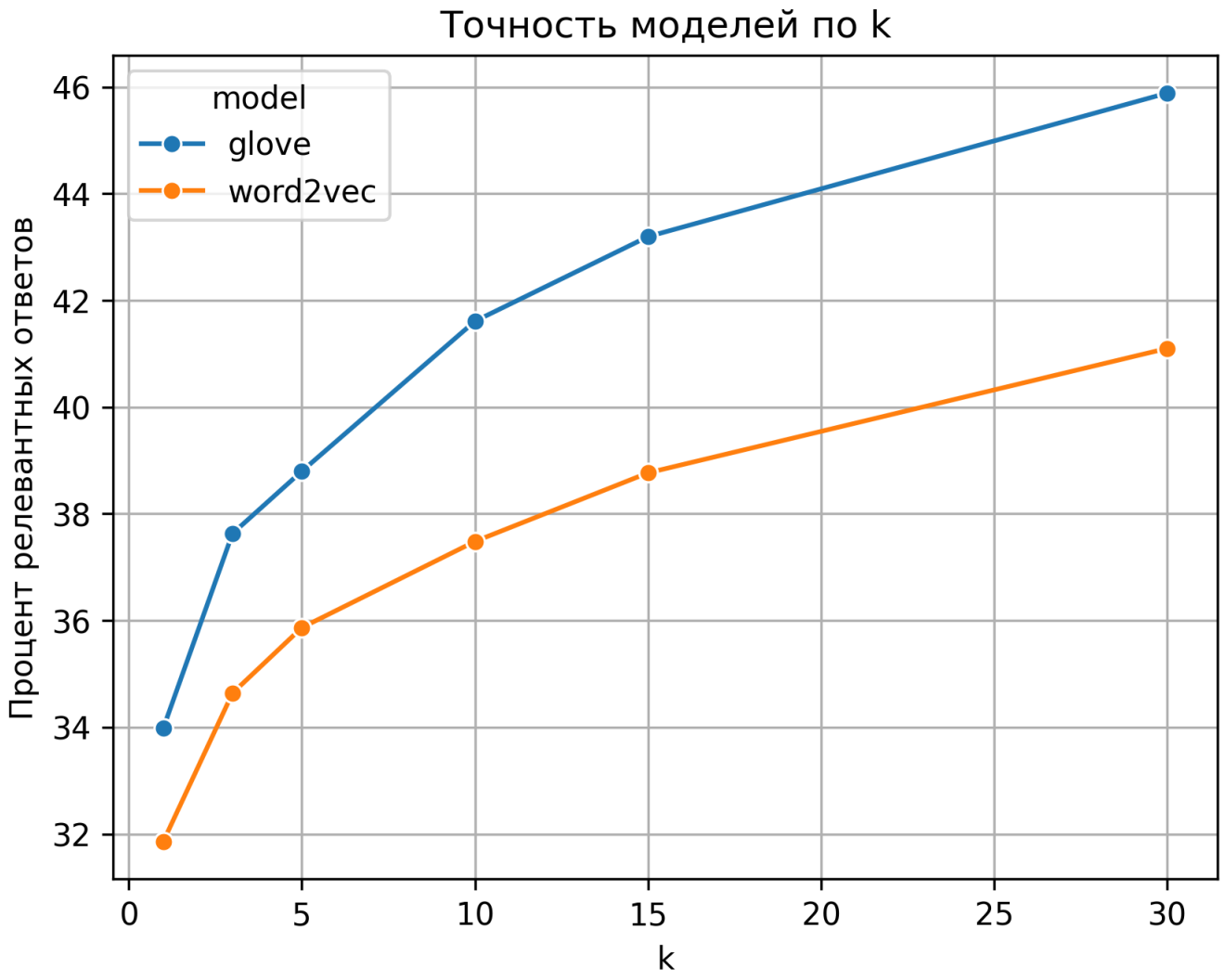

Рис.: Точность моделей Word2Vec и GloVe в задаче семантического поиска на основе корпуса новостей

- Была разработана система поиска по корпусу новостных документов на основе эмбеддингов
- GloVe показал лучшие результаты по сравнению с Word2Vec. Это делает его более подходящим для семантического поиска по новостям, особенно при работе с короткими запросами.

СПАСИБО ЗА ВНИМАНИЕ!