Информационный поиск

Лекция 3
Семантический поиск

Дроздова Ксения
drozdova.xenia@gmail.com

Что было в прошлый раз

Формула BM25

Компоненты BM25

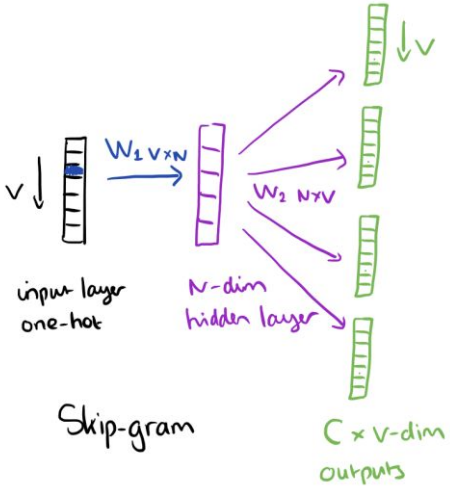
Обсуждение реализации BM25

Итоги:

Надо считать заранее все компоненты, которые не зависят от запроса, чтобы поиск работал быстро

KW search vs Semantic search

WORD2VEC

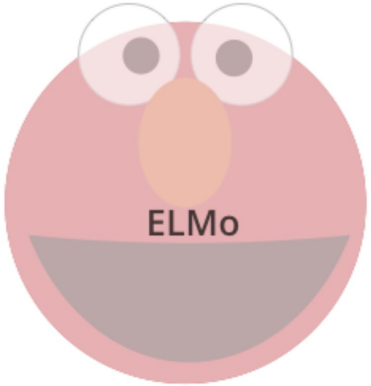


FASTTEXT

where

$n=3$

[<wh, whe, her, ere, re>]



word2vec

Для обучение и загрузки моделей W2V и FastText есть либа gensim
Она позволяет

1. Обучить модель на своем корпусе
2. Дообучить существующую модель на своих данных
3. Получить вектор слова из обученной модели, конечно
4. Обучить модель doc2vec

doc2vec

Первый способ получить вектор документа – обучить собственную модель doc2vec на корпусе
Но для хорошей итоговой модели вам потребуется действительно много текстов для обучения

Если у вас нет большого корпуса, то есть второй способ получить вектор текста - усреднить вектора входящих в него слов

Можно делать не простое среднее арифметическое, а учитывать веса для разных слов

DOCUMENTS VECTORS

Размерность вектора модели

	Hitchhiker's Guide to Galaxy	Last Chance to See	Life, Universe & Everything	Restaurant at End of Universe	So Long & Thanks for all the Fish	Starship Titanic
galaxy	**0.2204**	0	**0.2140**	**0.2125**	**0.1880**	**0.1943**
zaphod	**0.5861**	0	**0.5174**	**0.6354**	**0.2288**	0
ship	**0**	**0**	**0**	**0**	**0**	**0**
arthur	**0.6230**	0	**0.6301**	**0.5931**	**0.6160**	0
fiordland	0	**1.5171**	0	0	0	0
santorini	0	0	**1.1437**	0	0	0
wordlings	0	0	0	0	**0.7780**	0

Indexed Collection

X

Arthur has Samsung Galaxy
0.46
0
0
0.51
0
0
0

Query Vec

=

D1	0.43
D2	0
D3	0.42
D4	0.4
D5	0.41
D6	0.09

Scores

Моделируем реализацию

Видите, да, что это ничем не отличается от предыдущей идеи

Есть матрица данных проиндексированной коллекции - набор векторов для каждого документа

На входе у нас текстовый запрос, который мы индексируем тем же способом, что и коллекцию

Добавляем метрику сравнения векторов - косинусную близость, и ранжируем выдачу по убыванию метрики

Готовые модели

Где взять готовые обученные модели?

https://rusvectores.org/ru/models

На rusvectores давно можно взять обученные модели w2v и fasttext

А в конце августа Андрей Кутузов анонсировал выход Elmo, обученной на НКРЯ и Википедии

Пример работы с этой моделью можно найти в репозитории

https://github.com/ltgoslo/simple_elmo

Готовые модели

Где взять готовую модель Bert?

http://docs.deeppavlov.ai/en/master/features/models/bert.html

Ребята из deeppavlov обучили две модели Bert

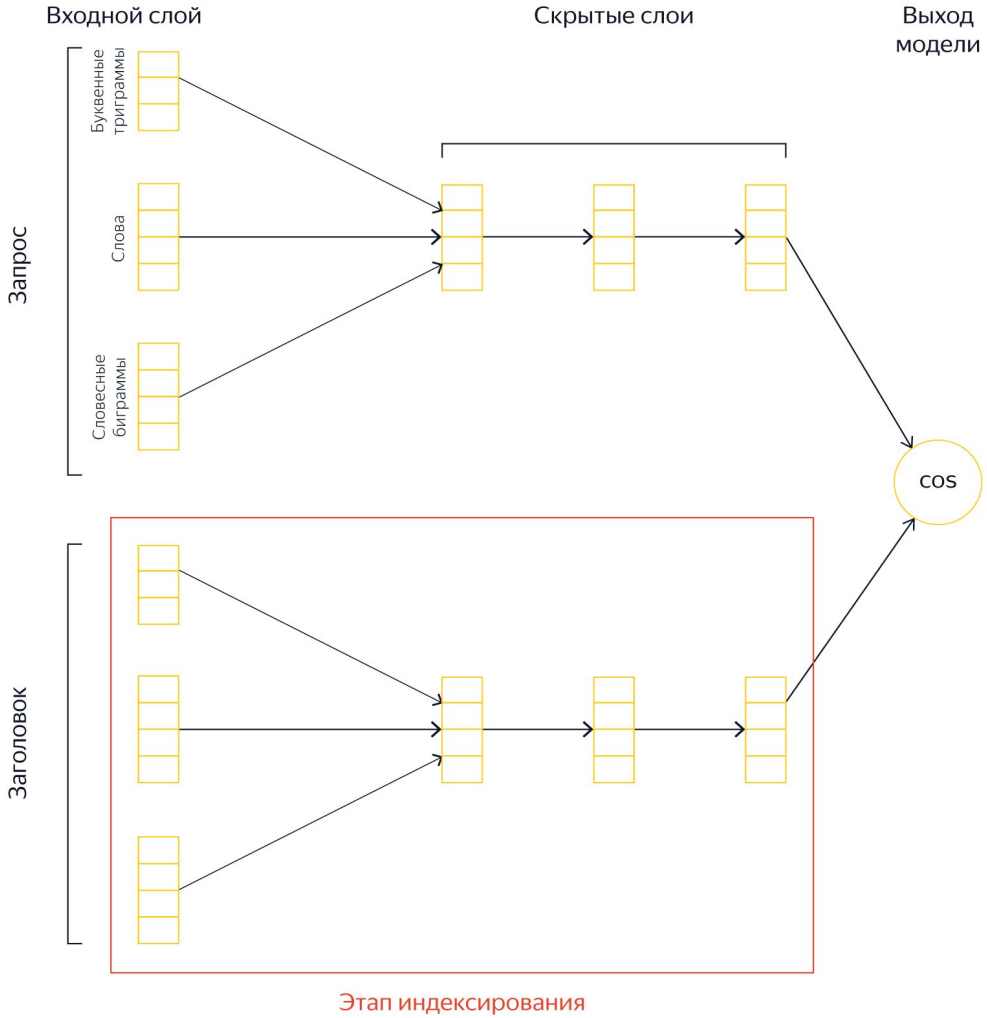
Одну классическую на Википедии и новостном корпусе

Другую более разговорную на *OpenSubtitles, Dirty, Pikabu, and Social Media segment of Taiga corpus*

Палех


Планета рецептов — картофель гратен
Календарь посадок: когда сажать картошку в разных регионах
[фильм про человека который выращивал картошку на другой планете]
Марсианин (2015) смотреть бесплатно онлайн
— поисковый запрос
— заголовки веб-страниц

Архитектура сети



Палех

Запрос [рассказ в котором раздавили бабочку]

Заголовок страницы	BM25	Нейронная модель
фильм в котором раздавили бабочку	0.79	0.82
и грянул гром википедия	0	0.43
брэдбери рэй википедия	0	0.27
машина времени роман википедия	0	0.24
домашнее малиновое варенье рецепт заготовки на зиму	0	0.06

Палех

Запрос [келлская книга]

Заголовок страницы	BM25	Нейронная модель
келлская книга википедия	0.91	0.92
ученые исследуют келлскую книгу вокруг света	0.88	0.85
book of kells wikipedia	0	0.81
ирландские иллюстрированные евангелия vii viii вв	0	0.58
икеа гипермаркеты товаров для дома и офиса ikea	0	0.09

Запрос [евангелие из келлса]

Заголовок страницы	BM25	Нейронная модель
келлская книга википедия	0	0.85
ученые исследуют келлскую книгу вокруг света	0	0.78
book of kells wikipedia	0	0.71
ирландские иллюстрированные евангелия vii viii вв	0.33	0.84
икеа гипермаркеты товаров для дома и офиса ikea	0	0.10