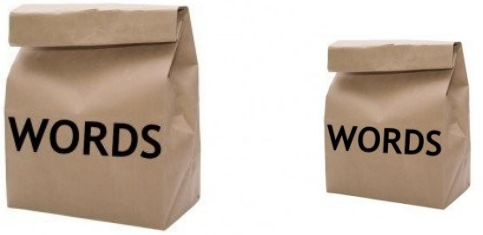
А. Островский

РАЗРАБОТКА “ИНТЕЛЛЕКТУАЛЬНОЙ“ СИСТЕМЫ, ОЦЕНИВАЮЩЕЙ РАЗВЕРНУТЫЕ ОТВЕТЫ НА ВОПРОСЫ, ПРЕДСТАВЛЕННЫЕ В ВИДЕ ТЕКСТОВ

Хакатон ПОВТАС ИИ, 2023

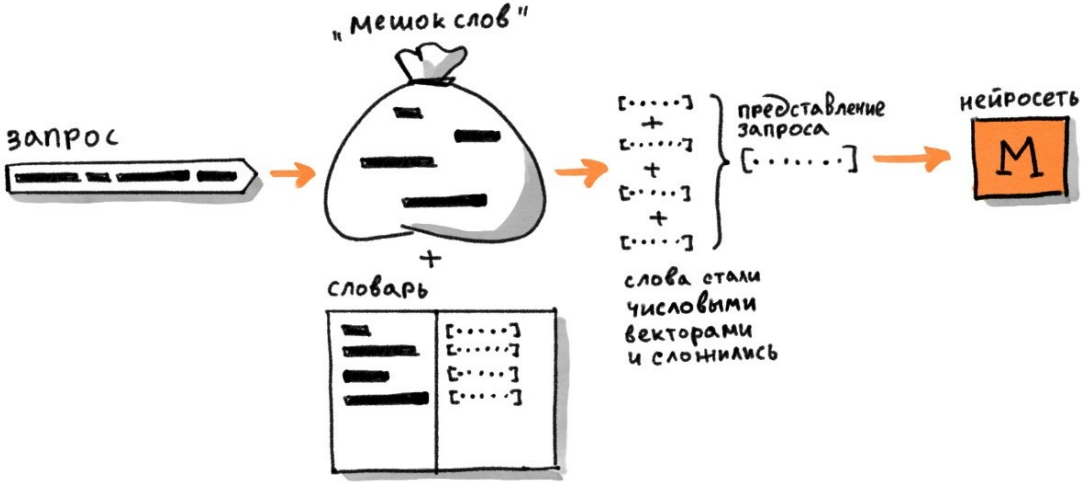
МОДЕЛЬ



В основу работы авторской модели положены принципы работы с “мешками слов”.

На ПЕРВОМ ЭТАПЕ входной текст, подлежащий оценке, и ответы, выдвинутые экспертами в качестве эталонных для обучения, обрабатываются лингвистическим препроцессором с целью нормализации ролей элементов (токенов — в данном случае слов) в текстах и помещаются в разные “мешки“.

На ВТОРОМ ЭТАПЕ производится сопоставление “мешков“ через оценки близости токенов в сообщениях, основываясь на фоновых векторных оценках корпуса текстов на русском языке и снепшота Википедии свежести 2021 года. Результат селектируется и шкалируется.



ЭТАП I

В основу работы авторской модели положены принципы работы с “мешками слов”.

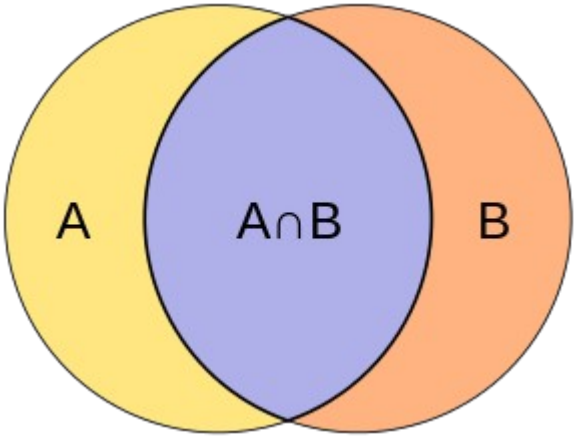
В начале ПРЕПРОЦЕССОР из входного текста, на основе морфологического анализа выделяет слова, которые приводятся к начальным формам; к полученным токенам прикрепляется определяющий тег, детерминирующий роль в структуре текста (например, tag.POS — часть речи). Так, например, “ёж”, “ежи” и “ежам” приводится к одной форме.

ЧАСТИ РЕЧИ ДЛЯ АНАЛИЗА

Часть речи

Граммема	Значение	Примеры
NOUN	имя существительное	хомяк
ADJF	имя прилагательное (полное)	хороший
ADJS	имя прилагательное (краткое)	хорош
COMP	компаратив	лучше, получше, выше
VERB	глагол (личная форма)	говорю, говорит, говорил
INFN	глагол (инфинитив)	говорить, сказать
PRTF	причастие (полное)	прочитавший, прочитанная
PRTS	причастие (краткое)	прочитана
GRND	деепричастие	прочитав, рассказывая
NUMR	числительное	три, пятьдесят
ADVB	наречие	круто
NPRO	местоимение-существительное	он
PRED	предикатив	некогда
PREP	предлог	в
CONJ	союз	и
PRCL	частица	бы, же, лишь
INTJ	междометие	ой

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

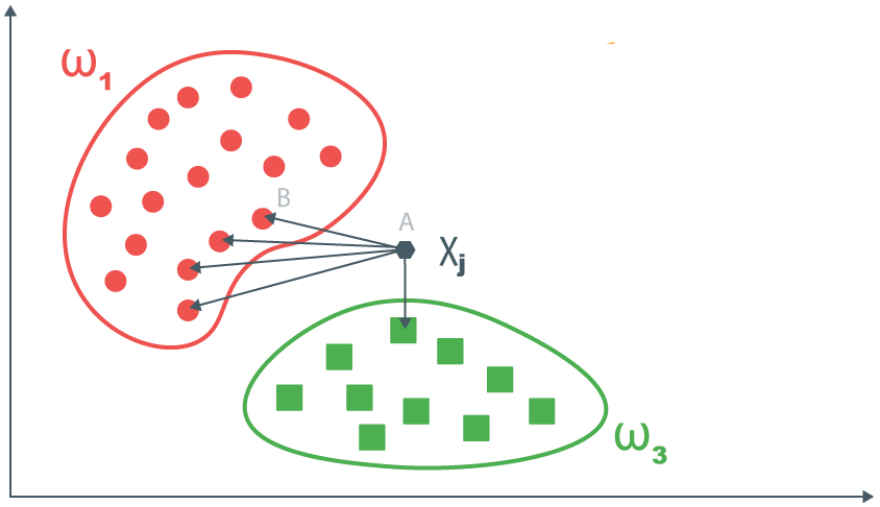


ЭТАП II

В качестве первичной метрики семантической близости “мешков“ используется мера Жаккарда с оценкой в диапазоне [0,1].
Вверху — пересечение “мешков“, внизу объединение “мешков“.

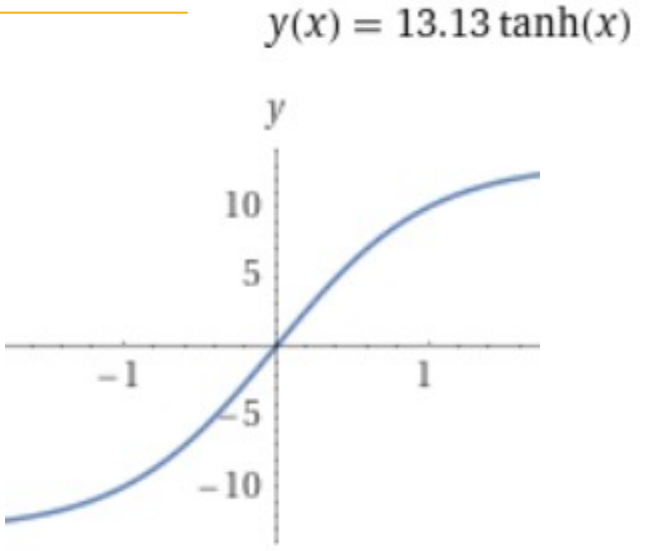
Особенности: расчет меры близости производится по векторным оценкам word2vec корпуса текстов на русском языке и снепшота Википедии.

СИНТЕЗ ОЦЕНКИ ВХОДНОГО ТЕКСТА



Синтез оценки входного текста выполняется при помощи метода ближайшего соседа. Выбирается наиболее близкий экспертный ответ к входному тексту для оценки с точки зрения выбранной меры.

НЕЛИНЕЙНАЯ НОРМАЛИЗАЦИЯ ОЦЕНКИ (ЭКСПЕРИМЕНТ)

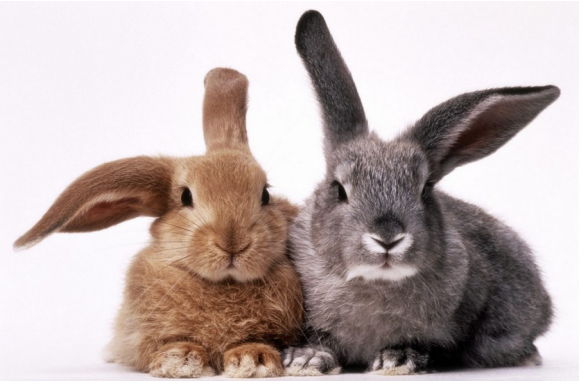


Гиперболический тангенс

Обусловлена:

1) Приведением оценки к требуемой шкале [1, 10]

2) Борьбой с негативизацией “средних оценок”



ПРОБЛЕМЫ

1) Отсутствие возможности обучения модели на ошибках (для полноценного обучения и валидации). Между тем множество парадигм ИИ уточняют модели именно на негативном опыте. Отсюда: собрать репрезентативную базу данных для обучения с “натуральными” ошибками куда важнее, чем эталонные тексты, на которых модель переобучается.

2) Отсутствие смысловой оценки. Контрпример несостоятельности мешка слов: частицы “не” или “ни”, меняющие значение высказывания, почти не меняют оценку модели.

ПРОГРАММНАЯ РЕАЛИЗАЦИЯ

Скрипт на языке Python: 242 строчки кода

Библиотеки: nltk, pymorphy2, gensim (word2vec)

Векторные представления слов на естественном языке:

https://rusvectores.org/en/models/#ruwikiruscorpora_upos_cbow_300_10_2021 (Корпус текстов на русском языке + Википедия)

MINIMUM VIABLE PRODUCT

Экспериментальная версия программы справляется с решением задачи лучше, чем ответ подброшенной монеты — 51,22%