**Домашнее задание № 4. Лексическая и дистрибутивная семантика**

На выбор варианты **A, B, C, D, Е** (*в скобках указаны баллы оценки)*

**Срок выполнения задания** – до **30 апреля** включительно

**Отчет высылается** **по почте** Олегу Дону на адрес [ordon@edu.hse.ru](file:///C:\Users\Elena\Documents\Елена\ВШЭ\Курс_КЛ\Курс_КЛ_2021\ДЗ_КЛ_21+оценки\ordon@edu.hse.ru)

В качестве темы письма указать: «**КЛ ДЗ № 4, ФИО**»

**A.**  *(5-7) баллов*

Изучить состав, функции и принципы построения системы *BabelNet* (babelnet.org) как лексического ресурса. Провести анализ ее возможностей, сравнив с принципами организации (лексические единицы и отношения) системы *WordNet* (http://wordnetweb.princeton.edu/perl/webwn) и/или *RuWordNet* (ruwordnet.ru/ru)и/или *РуТез* (<http://www.labinform.ru/pub/ruthes/index.htm>). Проиллюстрировать на 7-10 характерных примерах разного вида лексических единиц и отношений.

Отчет: характеристика функций системы и представленной в ней лексической информации (лексические единицы и отношения), иллюстрирующие примеры для слов основных частей речи, результаты сравнения системы с другими ресурсами (сходства-различия), выводы.

**B.** *(6-8 баллов)*

Сравнить представление лексической информации в 2-3 лингвистических ресурсах:

* в одном из ресурсов выбрать несколько (3-5) лексических единиц (слов и словосочетаний) и рассмотреть хранимую для них информацию, в том числе подсеть – иерархию в 2-4 уровня из 5-10 понятий;
* сопоставить эту информацию с данными в другом ресурсе для этих лексич. единиц.

Возможно сопоставление в рамках одного языка или в рамках разных языков.

Лингвистические ресурсы:

* Тезаурус РуТез: <http://www.labinform.ru/pub/ruthes/index.htm>
* Система КроссЛексика: <http://www.xl.gelbukh.com> или <http://crosslexica.ru/XLex.html>
* Идеографический словарь Баранова: <http://www.thesaurus1.narod.ru/>
* WordNet: [http://wordnetweb.princeton.edu/perl/webwn](http://wordnetweb.princeton.edu/perl/webwn%20)
* Аналоги WordNet для других языков, в частности,

-- EuroWordNet[: http://www.illc.uva.nl/EuroWordNet/](file:///C:\\Users\\Elena\\Documents\\Елена\\ВШЭ\\Курс_КЛ\\Курс_КЛ_2018\\ДЗ_КЛ_18\\:%20%20%20http:\\www.illc.uva.nl\\EuroWordNet\\)

-- WordNet для РЯ: <http://wordnet.ru>

Отчет: краткая характеристика сравниваемых лингвистических ресурсов, описание выбранных лексических единиц, сравнение их представлений в рассматриваемых ресурсах (сходства-различия), выводы.

**С.** *(6-8 баллов)*

На основе предобученной модели векторного представления слов *Word2Vec* и/или *FastText* для русского языка, взятой, например, из библиотеки *Gensim*: <https://radimrehurek.com/gensim/models/word2vec.html> или из ресурса *RusVectores:* <http://rusvectores.org/ru/models/> (можно воспользоваться его веб-интерфейсом)

провести экспериментальное исследование семантики нескольких (5-10) выбранных слов (достаточно частотных, разных частей речи): найти семантически близкие и характеризующие слова, определить близость пар слов, а также исследовать другие операции, допускаемые моделью. Рассмотреть несколько разных (2-3) обученных моделей и сравнить результаты в них для выбранных слов.

Отчет: краткая характеристика рассмотренных моделей векторного представления слов, составленная программа (если использовалась), описание проведенных экспериментов, сравнение результатов и выводы.

**D.** *(7-9 баллов)*

Для нескольких (5-10) достаточно частотных слов русского языка и семантически связанных с ними слов (синонимов, гипонимов, гиперонимов, антонимов и др.), подобранных из некоторого лексического ресурса, в частности,

* тезауруса РуТез: <http://www.labinform.ru/pub/ruthes/index.htm> (<http://wordnet.ru>) или
* системы КроссЛексика: <http://www.xl.gelbukh.com> или
* системыBabelNet ([https://babelnet.org/](https://babelnet.org/%20))

исследовать соотношение их векторных представлений на базе предобученной модели *Word2Vec* и/или *FastText*,взятой, например, из ресурса Rusvectores: <http://rusvectores.org/ru/models/> (можно воспользоваться его веб-интерфейсом) или из библиотеки Gensim: <https://radimrehurek.com/gensim/models/word2vec.html>.

Определить, есть ли какие-либо закономерности в косинусном расстоянии между лексически (тезаурусно) связанными словами? Подобны ли расстояния между парами связанных слов? Выполняются ли алгебраические зависимости?

Отчет: краткая характеристика использованной модели и выбранных для исследования слов, составленная программа (если использовалась), описание экспериментов, их результатов и выявленных закономерностей (если таковые обнаружились), выводы.

**E.** *(7-10 баллов)*

Провести построение и исследование языковой модели *Word2Vec* и/или *FastText* для русского языка, на базе известной или самостоятельно сформированной коллекции текстов (например, из Интернета), воспользовавшись для этого доступными средствами/инструментами с открытым кодом. Провести валидацию (проверку алгебраических свойств) обученной модели на основе подобранных примеров (или воспользовавшись открытыми источниками). Определить и проанализировать семантическое сходство для нескольких (5-10) выбранных слов разных частей речи (присутствующих в построенной модели), найти для них семантически близкие и характеризующие слова, по возможности выявить закономерности.

Дополнительно возможно сравнение векторных представлений выбранных слов в построенной модели с их представлениями в другой предобученной модели (например, из ресурса <http://rusvectores.org/ru/models/> ).

Отчет: указание способа построения модели и примененных для этого инструментов, характеристика обучающей коллекции, составленная программа с комментариями, описание валидации модели и ее результатов, анализ результатов проведенных экспериментов с выбранными словами, описание выявленных закономерностей (если таковые обнаружились), выводы.