1) ЦЕЛЬ И ЗАДАЧИ

Цель: Разработать ресурс (в идеале веб-сервис), который будет анализировать **русскоязычные новостные тексты** и выявлять семантические сдвиги определенных абстрактных понятий в рамках коротких промежутков времени с помощью **дистрибутивной семантики**.

- Инженерные задачи:
    - собрать данные для анализа;
    - программно реализовать существующие методы (на Python);
    - выявить семантические сдвиги на имеющихся данных;
    - создание/аннотация золотого стандарта;
    - оценка.
- Исследовательские задачи:
    - изучить существующую классификацию семантических сдвигов;
    - классифицировать выявленные сдвиги;
    - проанализировать корреляцию выявленных сдвигов с лингвистическими и экстралингвистическими процессами;
    - разработка способа оценки системы;
    - разработка способа разметки золотого стандарта.

2) ДАННЫЕ И МЕТОДЫ

Данные: уже обученные модели, новостные тексты на русском языке, размеченный вручную золотой стандарт.

Методы: skip-gram negative sampling (SGNS), continuous bag of words (CBOW), Jaccard similarity coefficient, Kendall's tau coefficient, global anchors, Procrustes analysis, ...

3) ЭТАПЫ

1. Подготовительный
    a. Применение простых методов измерения к готовым датасетам с DH-школы;
    b. Прикидки относительно дальнейшего анализа: предположения о периодах, которые интересно анализировать, и об исторических событиях, которые могли повлиять.
    c. Отбор предварительного сета интересующих нас прилагательных (например, топ-10 от каждой пары лет + 10 рандомных по Прокрустову анализу/глобальным якорям).
    d. Топонимические прилагательные выделить в отдельную категорию
    e. Формулировка research question (можно ли к русскоязычному новостному материалу успешно применить существующие методы выявления диахронических семантических сдвигов в прилагательных методами дистрибутивной семантики)
    f. Решение относительно методов оценки

2. Создание evaluation test set
    a. Создание своей тестовой выборки, включающей прилагательные и пары годов с 2000 по 2014 (плюс негативные сэмплы, выбранные случайно с тем же частотным распределением; 20 прилагательных из каждой пары лет)
    b. Ручная разметка тестовых данных.
3. Эксперименты и формулировка ответа на research question, используя методы, перечисленные в (2).
    a. Подготовка корпусов
    b. Обучение дистрибутивных моделей (эмбеддингов)
    c. Имплементация и прогон алгоритмов выявления диахронических семантических сдвигов.
    d. Оценка результатов на test set и интерпретация ошибок.
4. Публикация статьи, выступление на конференции (AIST, AINL, **ACL workshop**)
5. Создание веб-сервиса (осень)

4) ТАБЛИЦА

	этап	**Вадим**	**Юля**	**Даша**
январь	(2) разметка	280 словолет	280 словолет	280 словолет тимлид
февраль	(3) эксперименты	3ac тимлид	3bc	3dc
март	(3) эксперименты	3ac	3bc тимлид	3dc
апрель	(4) 26 апреля - подача статьи на Language Change Workshop	(4)	(4)	(4) тимлид